3

Game Theory and Human Behavior

God is crafty, but He is not malicious.

Albert Einstein

My motive for doing what I am going to do is simply personal revenge. I do not expect to accomplish anything by it.

Theodore Kaczynski (the Unabomber)

Game theory is multiplayer decision theory where the choices of each player affect the payoffs to other players, and the players take this into account in their choice behavior. In this chapter we address the contribution of game theory to the design of experiments aimed at understanding the behavior of individuals engaged in strategic interaction. We call this *behavioral game theory*.

Game theory is a general lexicon that applies to all life forms. Strategic interaction neatly separates living from nonliving entities and defines life itself. Strategic interaction is the sole concept commonly used in the analysis of living systems that has no counterpart in physics or chemistry.

Game theory provides the conceptual and procedural tools for studying social interaction, including the characteristics of the players, the rules of the game, the informational structure, and the payoffs associated with particular strategic interactions. The various behavioral disciplines (economics, psychology, sociology, politics, anthropology, and biology) are currently based on distinct principles and rely on distinct types of data. However, game theory fosters a unified analytical framework available to *all* the behavioral disciplines. This facilitates cross-disciplinary information exchange that may eventually culminate in a degree of unity within the behavioral sciences now enjoyed only by the natural sciences (see chapter 12). Moreover, because behavioral game-theoretic predictions can be systematically tested, the results can be replicated by different laboratories (Plott 1979; Smith 1982; Sally 1995). This turns social science into true science.

Behavioral game theory presumes the BPC model, as developed in §1.1. Experiments subject individuals to a variety of game settings, including diverse payoffs, informational conditions, and constraints on action, and

deduce their underlying preferences from their behavior. This would be impossible if the individuals were not maximizing consistent preferences. Chapter 1 showed that the deviations that human subjects exhibit from the prescriptions of normative decision theory, while important, are compatible with preference consistency plus performance error.

3.1 Self- and Other-Regarding Preferences

This chapter deals with the interplay of self-regarding and other-regarding behavior. By a *self-regarding* actor we mean a player i in a game $\mathcal{G}$ who maximizes his own payoff π_i as defined in §2.1. A self-regarding actor thus cares about the behavior of and payoffs to the other players only insofar as these impact his own payoff π_i. The term "self-regarding" is more accurate than "self-interested" because an other-regarding individual is still acting to maximize utility and so can be described as self-interested. For instance, if I get great pleasure from your consumption, my gift to you may be self-interested, even though it is surely other-regarding. We can avoid confusion (and much pseudophilosophical discussion) by employing the self-regarding/other-regarding terminology.

One major result of behavioral game theory is that *when modeling market processes with well-specified contracts, such as double auctions (supply and demand) and oligopoly, game-theoretic predictions assuming self-regarding actors are accurate under a wide variety of social settings* (Kachelmaier and Shehata 1992; Davis and Holt 1993). In such market settings behavioral game theory sheds much new light, particularly in dealing with price dynamics and their relationship to buyer and seller expectations (Smith and Williams 1992).

The fact that self-regarding behavior explains market dynamics lends credence to the practice in neoclassical economics of assuming that individuals are self-regarding. However, it by no means justifies "Homo economicus" because many economic transactions do *not* involve anonymous exchange. This includes employer-employee, creditor-debtor, and firm-client relationships. Nor does this result apply to the welfare implications of economic outcomes (e.g., people may care about the overall degree of economic inequality and/or their positions in the income and wealth distribution), to modeling the behavior of taxpayers (e.g., they may be more or less honest than a self-regarding individual, and they may prefer to transfer resources toward or away from other individuals even at an expense to themselves)

or to important aspects of economic policy (e.g., dealing with corruption, fraud, and other breaches of fiduciary responsibility).

A second major result is that *when contracts are incomplete and individuals can engage in strategic interaction, with the power to reward and punish the behavior of other individuals, game-theoretic predictions based on the self-regarding actor model generally fail*. In such situations, the *character virtues* (including, honesty, promise keeping, trustworthiness, and decency), as well as both *altruistic cooperation* (helping others at a cost to oneself) and *altruistic punishment* (hurting others at a cost to oneself) are often observed. These behaviors are particularly common in a *social dilemma*, which is an n-player Prisoner's Dilemma—a situation in which all gain when all cooperate but each has a personal incentive to defect, gaining at the expense of the others (see, for instance, §3.9).

Other-regarding preferences were virtually ignored until recently in both economics and biology, although they are standard fare in anthropology, sociology, and social psychology. In economics, the notion that enlightened self-interest allows individuals to cooperate in large groups goes back to Bernard Mandeville's "private vices, public virtues" (1924 [1705]) and Adam Smith's "invisible hand" (2000 [1759]). The great Francis Ysidro Edgeworth considered self-interest "the first principle of pure economics" (Edgeworth 1925, p. 173). In biology, the selfishness principle has been touted as a central implication of rigorous evolutionary modeling. In *The Selfish Gene* (1976), for instance, Richard Dawkins asserts "We are survival machines—robot vehicles blindly programmed to preserve the selfish molecules known as genes. ...Let us try to teach generosity and altruism, because we are born selfish." Similarly, in *The Biology of Moral Systems* (1987, p. 3), R. D. Alexander asserts that "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." More poetically, Michael Ghiselin (1974) writes: "No hint of genuine charity ameliorates our vision of society, once sentimentalism has been laid aside. What passes for cooperation turns out to be a mixture of opportunism and exploitation. ...Scratch an altruist, and watch a hypocrite bleed."

The Darwinian struggle for existence may explain why the concept of virtue does not add to our understanding of animal behavior in general, but by all available evidence, it is a central aspect of human behavior. The reasons for this are the subject of some speculation (Gintis 2003a,2006b), but they come down to the plausible insight that human social life is so

complex, and the rewards for prosocial behavior so distant and indistinct, that adherence to general rules of propriety, including the strict control of such deadly sins as anger, avarice, gluttony, and lust, is individually fitness-enhancing (Simon 1990; Gintis 2003a).

One salient behavior in social dilemmas revealed by behavioral game theory is *strong reciprocity*. Strong reciprocators come to a social dilemma with a propensity to cooperate (*altruistic cooperation*), respond to cooperative behavior by maintaining or increasing their level of cooperation, and respond to noncooperative behavior by punishing the "offenders," even at a cost to themselves and even when they cannot reasonably expect future personal gains to flow therefrom (*altruistic punishment*). When other forms of punishment are not available, the strong reciprocator responds to defection with defection.

The strong reciprocator is thus neither the selfless altruist of utopian theory, nor the self-regarding individual of traditional economics. Rather, he is a conditional cooperator whose penchant for reciprocity can be elicited under circumstances in which self-regard would dictate otherwise. The positive aspect of strong reciprocity is commonly known as gift exchange, in which one individual behaves more kindly than required toward another with the hope and expectation that the other will treat him kindly as well (Akerlof 1982). For instance, in a laboratory-simulated work situation in which employers can pay higher than market-clearing wages in hopes that workers will reciprocate by supplying a high level of effort (§3.7), the generosity of employers was generally amply rewarded by their workers.

A second salient behavior in social dilemmas revealed by behavioral game theory is *inequality aversion*. The inequality-averse individual is willing to reduce his own payoff to increase the degree of equality in the group (whence widespread support for charity and social welfare programs). But he is especially displeased when placed on the *losing side* of an unequal relationship. The inequality-averse individual is willing to reduce his own payoff if that reduces the payoff of relatively favored individuals even more. In short, an inequality-averse individual generally exhibits a *weak* urge to reduce inequality when he is the beneficiary and a *strong* urge to reduce inequality when he is the victim (Loewenstein, Thompson, and Bazerman 1989). Inequality aversion differs from strong reciprocity in that the inequality-averse individual cares only about the distribution of final payoffs and not at all about the role of other players in bringing about this

distribution. The strong reciprocator, by contrast, does not begrudge others their payoffs but is sensitive to how fairly he is treated by others.

Self-regarding agents are in common parlance called *sociopaths*. A sociopath (e.g., a sexual predator, a recreational cannibal, or a professional killer) treats others instrumentally, caring only about what he derives from an interaction, whatever the cost to the other party. In fact, for most people, interpersonal relations are guided as much by empathy (and hostility) as by self-regard. The principle of *sympathy* is the guiding theme of Adam Smith's great book, *The Theory of Moral Sentiments*, despite the fact that his self-regarding principle of the "invisible hand" is one of the central insights of economic theory.

We conclude from behavioral game theory that one must treat individuals' objectives as a matter of *fact*, not *logic*. We can just as well build models of honesty, promise keeping, regret, strong reciprocity, vindictiveness, status seeking, shame, guilt, and addiction as of choosing a bundle of consumption goods subject to a budget constraint (§12.7), (Gintis 1972a,b, 1974, 1975; Becker and Murphy 1988; Bowles and Gintis 1993; Becker 1996; Becker and Mulligan 1997). ,

3.2 Methodological Issues in Behavioral Game Theory

Vernon Smith, who was awarded the Nobel prize in 2002, began running laboratory experiments of market exchange in 1956 at Purdue and Stanford Universities. Until the 1980's, aside from Smith, whose results supported the traditional theory of market exchange, virtually the only behavioral discipline to use laboratory experiments with humans as a basis for modeling human behavior was social psychology. Despite the many insights afforded by experimental social psychology, its experimental design was weak. For instance, the BPC model was virtually ignored and game theory was rarely used, so observed behavior could not be analytically modeled, and experiments rarely used incentive mechanisms (such as monetary rewards and penalties) designed to reveal the real, underlying preferences of subjects. As a result, social psychological findings that were at variance with the assumptions of other behavioral sciences were widely ignored.

The results of the *Ultimatum Game* (Güth, Schmittberger, and Schwarze 1982) changed all that (§3.6), showing that in one-shot games that preserved the anonymity of subjects, people were quite willing to reject monetary rewards that they considered unfair. This, and a barrage of succeeding experi-

ments, some of which are analyzed below, did directly challenge the widely used assumption that individuals are self-regarding. Not surprisingly, the first reaction within the disciplines was to criticize the experiments rather than to question their theoretical preconceptions. This is a valuable reaction to new data, so we shall outline the various objections to these findings.

First, the behavior of subjects in simple games under controlled circumstances may bear no implications for their behavior in the complex, rich, temporally extended social relationships into which people enter in daily life. We discuss the *external validity* of laboratory experiments in §3.15.

Second, games in the laboratory are unusual, so people do not know how best to behave in these games. They therefore simply play as they would in daily life, in which interactions are repeated rather than one-shot, and take place among acquaintances rather than being anonymous. For instance, critics suggest that strong reciprocity is just a confused carryover into the laboratory of the subject's extensive experience with the value of building a reputation for honesty and willingness to punish defectors, both of which benefit the self-regarding actor. However, when opportunities for reputation building are incorporated into a game, subjects make predictable strategic adjustments compared to a series of one-shot games without reputation building, indicating that subjects are capable of distinguishing between the two settings (Fehr and Gächter 2000). Postgame interviews indicate that subjects clearly comprehend the one-shot aspect of the games.

Moreover, one-shot, anonymous interactions are not rare. We face them frequently in daily life. Members of advanced market societies are engaged in one-shot games with very high frequency—virtually every interaction we have with strangers is of this form. Major rare events in people's lives (fending off an attacker, battling hand to hand in wartime, experiencing a natural disaster or major illness) are one-shots in which people appear to exhibit strong reciprocity much as in the laboratory. While members of the small-scale societies we describe below may have fewer interactions with strangers, they are no less subject to one-shots for the other reasons mentioned. Indeed, in these societies, greater exposure to market exchange led to stronger, not weaker, deviations from self-regarding behavior (Henrich et. al al 2004).

Another indication that the other-regarding behavior observed in the laboratory is not simply confusion on the part of the subjects is that when experimenters point out that subjects could have earned more money by behaving differently, the subjects generally respond that of course they knew

that but preferred to behave in an ethically or emotionally satisfying manner rather than simply maximize their material gain. This, by the way, contrasts sharply with the experiments in behavioral decision theory described in chapter 1 where subjects generally admitted their errors.

Recent neuroscientific evidence supports the notion that subjects punish those who are unfair to them simply because this gives them pleasure. deQuervain et. al (2004) used positron emission tomography to examine the neural basis the for altruistic punishment of defectors in an economic exchange. The experimenters scanned the subjects' brains while they learned about the defector's abuse of trust and determined the punishment. Punishment activated the *dorsal striatum*, which has been implicated in the processing of rewards that accrue as a result of goal-directed actions. Moreover, subjects with stronger activations in the dorsal striatum were willing to incur greater costs in order to punish. This finding supports the hypothesis that people derive satisfaction from punishing norm violations and that the activation in the dorsal striatum reflects the anticipated satisfaction from punishing defectors.

Third, it may be that subjects really do not believe that conditions of anonymity will be respected, and they behave altruistically because they fear their selfish behavior will be revealed to others. There are several problems with this argument. First, one of the strict rules of behavioral game research is that *subjects are never told untruths or otherwise misled*, and they are generally informed of this fact by experimenters. Thus, revealing the identity of participants would be a violation of scientific integrity. Second, there are generally no penalties that could be attached to being discovered behaving in a selfish manner. Third, an exaggerated fear of being discovered cheating is *itself* a part of the strong reciprocity syndrome—it is a psychological characteristic that induces us to behave prosocially even when we are most attentive to our selfish needs. For instance, subjects might feel embarrassed and humiliated were their behavior revealed, but shame and embarrassment are themselves *other-regarding emotions* that contribute to prosocial behavior in humans (Bowles and Gintis 2004; Carpenter et. al al 2009). In short, the tendency of subjects to overestimate the probability of detection and the costs of being detected are prosocial mental processes (H. L. Mencken once defined "conscience" as "the little voice that warns us that someone may be looking"). Fourth, and perhaps most telling, in tightly controlled experiments designed to test the hypothesis that subject-experimenter anonymity is important in fostering altruistic

behavior, it is found that subjects behave similarly regardless of the experimenter's knowledge of their behavior (Bolton and Zwick 1995; Bolton, Katok, and Zwick 1998).

A final argument is that while a game may be one-shot and the players may be anonymous to one another, they nonetheless *remember* how they played a game, and they may derive great pleasure from recalling their generosity or their willingness to incur the costs of punishing others for being selfish. This is quite correct and probably explains a good deal of non-self-regarding behavior in experimental games.[1] But this does not contradict the fact that our behavior is other-regarding! Rather, it affirms that it may be in one's personal interest to engage in other-regarding acts. Only for sociopaths are the set of self-regarding acts and the set of self-interested acts the same.

In all the games described below, unless otherwise stated, subjects were college students who were anonymous to one another, were paid real money, were not deceived or misled by the experimenters, and they were instructed to the point where they fully understood the rules and the payoffs before playing for real.

3.3 An Anonymous Market Exchange

By *neoclassical economics* I mean the standard fare of microeconomics courses, including the Walrasian general equilibrium model, as developed by Kenneth Arrow, Gérard Debreu, Frank Hahn, Tjalling Koopmans, and others (Arrow 1951; Arrow and Hahn 1971; Koopmans 1957). Neoclassical economic theory holds that in a market for a product, the equilibrium price is at the intersection of the supply and demand curves for the good. It is easy to see that at any other point a self-regarding seller could gain by asking a higher price, or a self-regarding buyer could gain by offering a lower price. This situation was among the first to be simulated experimentally, the neoclassical prediction virtually always receiving strong support (Holt 1995). Here is a particularly dramatic example, provided by Holt, Langan, and Villamil (1986) (reported by Charles Holt in Kagel and Roth, 1995).

[1] William Shakespeare understands this well when he has Henry V use the following words to urge his soldiers to fight for victory against a much larger French army: "Whoever lives past today ... will rouse himself every year on this day, show his neighbor his scars, and tell embellished stories of all their great feats of battle. These stories he will teach his son and from this day until the end of the world we shall be remembered."

In the Holt-Langan-Villamil experiment there are four "buyers" and four "sellers." The good is a chip that the seller can redeem for \$5.70 (unless it is sold) but a buyer can redeem for \$6.80 at the end of the game. In analyzing the game, we assume throughout that buyers and sellers are self-regarding. In each of the first five rounds, each buyer was informed, privately, that he could redeem up to 4 chips, while 11 chips were distributed to sellers (three sellers were given 3 chips each, and the fourth was given 2 chips). Each player knew only the number of chips in his possession, the number he could redeem, and their redemption value, and did not know the value of the chips to others or how many they possessed or were permitted to redeem. Buyers should be willing to pay up to \$6.80 per chip for up to 4 chips each, and sellers should be willing to sell a chip for any amount at or above \$5.70. Total demand is thus 16 for all prices at or below \$6.80, and total supply is 11 chips at or above \$5.70. Because there is an excess demand for chips at every price between \$5.70 and \$6.80, the only point of intersection of the demand and supply curves is at the price $p = \$6.80$. The subjects in the game, however, have absolutely no knowledge of aggregate demand and supply because each knows only his own supply of or demand for chips.

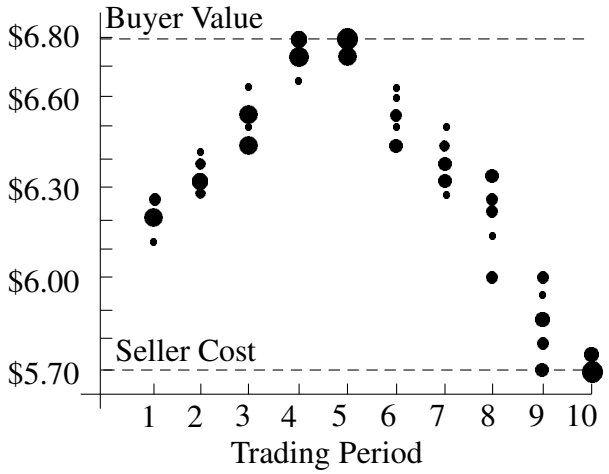

Figure 3.1. The double auction. The size of the circle is proportional to the number of trades that occurred at the stated price.

The rules of the game are that at any time a seller can call out an asking price for a chip, and a buyer can call out an offer price for a chip. This price remains "on the table" until it is accepted by another player, or a

lower asking price is called out, or a higher offer price is called out. When a deal is made, the result is recorded and that chip is removed from the game. As seen in figure 3.1, in the first period of play, actual prices were about midway between \$5.70 and \$6.80. Over the succeeding four rounds the average price increased until in period 5 prices were very close to the equilibrium price predicted by neoclassical theory.

In period 6 and each of the succeeding four periods, buyers were given the right to redeem a total of 11 chips, and each seller was given 4 chips. In this new situation, it is clear (to observers who know these facts, though not the subjects in the experiment) that there is now an *excess supply* of chips at each price between \$5.70 and \$6.80, so supply and demand intersect precisely at \$5.70. While sellers, who previously made a profit of about \$1.10 per chip in each period, must have been delighted with their additional supplies of chips, succeeding periods witnessed a steady fall in price until in the tenth period the price is close to the neoclassical prediction, and now buyers are earning about \$1.10 per chip. We see that even when agents are completely ignorant of macroeconomics conditions of supply and demand, they can move quickly to a market-clearing equilibrium under the appropriate conditions.

3.4 The Rationality of Altruistic Giving

There is nothing irrational about caring for others. But do preferences for altruistic acts entail transitive preferences as required by the notion of rationality in decision theory? Andreoni and Miller (2002) showed that in the case of the Dictator Game, they do. Moreover, there are no known counterexamples.

In the Dictator Game, first studied by Forsythe et. al al (1994), the experimenter gives a subject, called the Dictator, a certain amount of money and instructs him to give any portion of it he desires to a second, anonymous, subject, called the Receiver. The Dictator keeps whatever he does not choose to give to the Receiver. Obviously, a self-regarding Dictator will give nothing to the Receiver. Suppose the experimenter gives the Dictator m points (exchangeable at the end of the session for real money) and tells him that the price of giving some of these points to the Receiver is p, meaning that each point the Receiver gets costs the giver p points. For instance, if $p = 4$, then it costs the Dictator 4 points for each point that he transfers to the Receiver. The Dictator's choices must then satisfy the bud-

get constraint $\pi_s + p\pi_o = m$, where π_s is the amount the Dictator keeps and π_o is the amount the Receiver gets. The question, then, is simply, is there a preference function $u(\pi_s, \pi_o)$ that the Dictator maximizes subject to the budget constraint $\pi_s + p\pi_o = m$? If so, then it is just as rational, from a behavioral standpoint, to care about giving to the Receiver as to care about consuming marketed commodities.

Varian (1982) showed that the following generalized axiom of revealed preference (GARP) is sufficient to ensure not only rationality but that individuals have nonsatiated, continuous, monotone, and concave utility functions—the sort expected in traditional consumer demand theory. To define GARP, suppose the individual purchases bundle $x(p)$ when prices are p. We say consumption bundle $x(p_s)$ is *directly revealed to be preferred* to bundle $x(p_t)$ if $p_s x(p_t) \leq p_s x(p_s)$; i.e., $x(p_t)$ could have been purchased when $x(p_s)$ was purchased. We say $x(p_s)$ is *indirectly revealed to be preferred* to $x(p_t)$ if there is a sequence $x(p_s) = x(p_1), x(p_2), \ldots, x(p_k) = x(p_t)$, where each $x(p_i)$ is directly revealed preferred to $x(p_{i+1})$ for $i = 1, \ldots, k-1$. GARP then is the following condition: if $x(p_s)$ is indirectly revealed to be preferred to $x(p_t)$, then $p_t x(p_t) \leq p_t x(p_s)$; i.e., $x(p_s)$ does not cost less than $x(p_t)$ when $x(p_s)$ is purchased.

Andreoni and Miller (2002) worked with 176 students in an elementary economics class and had them play the Dictator Game multiple times each, with the price p taking on the values $p = 0.25, 0.33, 0.5, 1, 2, 3,$ and 4, with amounts of tokens equaling $m = 40, 60, 75, 80,$ and 100. They found that only 18 of the 176 subjects violated GARP at least once and that of these violations, only four were at all significant. By contrast, if choices were randomly generated, we would expect that between 78% and 95% of subjects would have violated GARP.

As to the degree of altruistic giving in this experiment, Andreoni and Miller found that 22.7% of subjects were perfectly selfish, 14.2% were perfectly egalitarian at all prices, and 6.2% always allocated all the money so as to maximize the total amount won (i.e., when $p > 1$, they kept all the money, and when $p < 1$, they gave all the money to the Receiver).

We conclude from this study that, at least in some cases, and perhaps in all, we can treat altruistic preferences in a manner perfectly parallel to the way we treat money and private goods in individual preference functions. We use this approach in the rest of the problems in this chapter.

3.5 Conditional Altruistic Cooperation

Both strong reciprocity and inequality aversion imply *conditional altruistic cooperation* in the form of a predisposition to cooperate in a social dilemma as long as the other players also cooperate, although they have different reasons: the strong reciprocator believes in returning good for good, whatever the distributional implications, whereas the inequality-averse individual simply does not want to create unequal outcomes by making some parties bear a disproportionate share of the costs of cooperation.

Social psychologist Toshio Yamagishi and his coworkers used the Prisoner's Dilemma (§2.10) to show that a majority of subjects (college students in Japan and the United States) positively value altruistic cooperation. In this game, let CC stand for "both players cooperate," let DD stand for "both players defect," let CD stand for "I cooperate but my partner defects," and let DC stand for "I defect and my partner cooperates." A self-regarding individual will exhibit $DC \succ CC \succ DD \succ CD$ (check it), while an altruistic cooperator will exhibit $CC \succ DC \succ DD \succ CD$ (for notation, see §1.1); i.e. the self-regarding individual prefers to defect no matter what his partner does, whereas the conditional altruistic cooperator prefers to cooperate so long as his partner cooperates. Watabe et. al al (1996), using 148 Japanese subjects, found that the average desirability of the four outcomes conformed to the altruistic cooperator preferences ordering. The experimenters also asked 23 of the subjects if they would cooperate if they already knew that their partner was going to cooperate, and 87% (20) said they would. Hayashi et. al al (1999) ran the same experiment with U.S. students with similar results. In this case, all the subjects said they would cooperate if their partners were already committed to cooperating.

While many individuals appear to value conditional altruistic cooperation, the above studies did not use real monetary payoffs, so it is unclear how strongly these values are held, or if they are held at all, because subjects might simply be paying lip service to altruistic values that they in fact do not hold. To address this issue, Kiyonari, Tanida, and Yamagishi (2000) ran an experiment with real monetary payoffs using 149 Japanese university students. The experimenters ran three distinct treatments, with about equal numbers of subjects in each treatment. The first treatment was a standard "simultaneous" Prisoner's Dilemma, the second was a "second-player" situation in which the subject was told that the first player in the Prisoner's Dilemma had already chosen to cooperate, and the third was a "first-player" treatment in which the subject was told that his decision to cooperate or de-

fect would be made known to the second player before the latter made his own choice. The experimenters found that 38% of the subjects cooperated in the simultaneous treatment, 62% cooperated in the second player treatment, and 59% cooperated in the first-player treatment. The decision to cooperate in each treatment cost the subject about $5 (600 yen). This shows unambiguously that a majority of subjects were conditional altruistic cooperators (62%). Almost as many were not only cooperators, but were also willing to bet that their partners would be (59%), provided the latter were assured of not being defected upon, although under standard conditions, without this assurance, only 38% would in fact cooperate.

3.6 Altruistic Punishment

Both strong reciprocity and inequality aversion imply *altruistic punishment* in the form of a predisposition to punish those who fail to cooperate in a social dilemma. The source of this behavior is different in the two cases: the strong reciprocator believes in returning harm for harm, whatever the distributional implications, whereas the inequality-averse individual wants to create a more equal distribution of outcomes even at the cost of lower outcomes for himself and others. The simplest game exhibiting altruistic punishment is the *Ultimatum Game* (Güth, Schmittberger, and Schwarze 1982). Under conditions of anonymity, two player 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 and the Responder 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. The two players do not face each other again.

There is only *one* Responder strategy that is a best response for a self-regarding individual: accept anything you are offered. Knowing this, a self-regarding Proposer who believes he faces a self-regarding Responder, offers the minimum possible amount, $1, and this is accepted.

However, when actually played, the self-regarding outcome is almost never attained or even approximated. In fact, as many replications of this experiment have documented, under varying conditions and with varying amounts of money, Proposers routinely offer Responders very substantial amounts (50% of the total generally being the modal offer) and Respon-

ders frequently reject offers below 30% (Güth and Tietz 1990; Camerer and Thaler 1995). Are these results culturally dependent? Do they have a strong genetic component or do all successful cultures transmit similar values of reciprocity to individuals? Roth et. al al (1991) conducted the Ultimatum Game in four different countries (United States, Yugoslavia, Japan, and Israel) and found that while the level of offers differed a small but significant amount in different countries, the probability of an offer being rejected did not. This indicates that both Proposers and Responders share the same notion of what is considered fair in that society and that Proposers adjust their offers to reflect this common notion. The differences in level of offers across countries, by the way, were relatively small. When a much greater degree of cultural diversity is studied, however, large differences in behavior are found, reflecting different standards of what it means to be fair in different types of societies (Henrich et. al 2004).

Behavior in the Ultimatum Game thus conforms to the strong reciprocity model: fair behavior in the Ultimatum Game for college students is a 50–50 split. Responders reject offers under 40% as a form of altruistic punishment of the norm-violating Proposer. Proposers offer 50% because they are altruistic cooperators, or 40% because they fear rejection. To support this interpretation, we note that if the offers in an Ultimatum Game are generated by a computer rather than by the Proposer, and if Responders know this, low offers are 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 getting nothing but allows the Proposer to keep the share he suggested for himself, Responders never reject offers, and proposers make considerably smaller (but still positive) offers (Bolton and Zwick 1995). As a final indication that strong reciprocity motives are operative in this game, after the game is over, when asked why they offered more than the lowest possible amount, Proposers commonly said that they were afraid that Responders will consider low offers unfair and reject them. When Responders rejected offers, they usually claimed they want to punish unfair behavior. In all of the above experiments a significant fraction of subjects (about a quarter, typically) conformed to self-regarding preferences.

3.7 Strong Reciprocity in the Labor Market

Gintis (1976) and Akerlof (1982) suggested that, in general, employers pay their employees higher wages than necessary in the expectation that workers will respond by providing higher effort than necessary. Fehr, Gächter, and Kirchsteiger (1997) (see also Fehr and Gächter 1998) performed an experiment to validate this *legitimation* or *gift exchange* model of the labor market.

The experimenters divided a group of 141 subjects (college students who had agreed to participate in order to earn money) into "employers" and "employees." The rules of the game are as follows. If an employer hires an employee who provides effort e and receives a wage w, his profit is $\pi = 100e - w$. The wage must be between 1 and 100, and the effort is 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 shown in figure 3.2. All payoffs involve real money that the subjects are paid at the end of the experimental session. We call this the *Experimental Labor Market Game*.

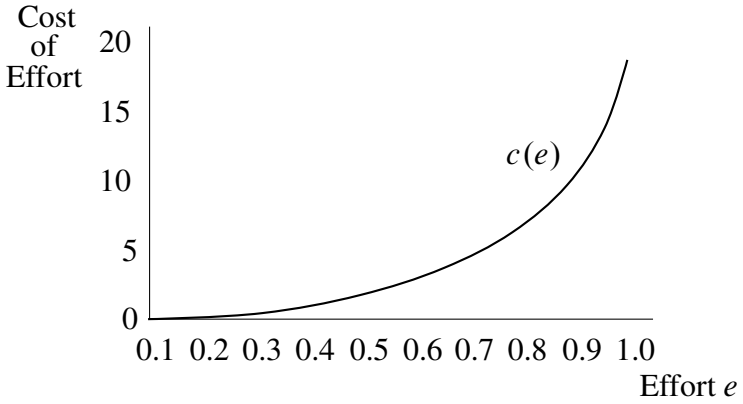

Figure 3.2. The Cost-of-effort schedule in Fehr, Gächter, and Kirchsteiger (1997).

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 that *need not equal the contracted effort* e^*. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level, $e \in [0.1, 1]$, with impunity. Although subjects may play this game

several times with different partners, each employer-employee interaction is a one-shot (nonrepeated) 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 (assuming only integer wage offers are permitted).[2] The employee will accept this offer and will set $e = 0.1$. Because $c(0.1) = 0$, the employee's payoff is $u = 1$. The employer's payoff is $\pi = 0.1 \times 100 - 1 = 9$.

In fact, however, this self-regarding outcome rarely occurred in this experiment. 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 provided. In effect, employers presumed the strong reciprocity predispositions of the employees, making quite generous wage offers and receiving higher effort, as a means to increase both their own and the employee's payoff, as depicted in figure 3.3. Similar results have been observed in Fehr, Kirchsteiger, and Riedl (1993, 1998).

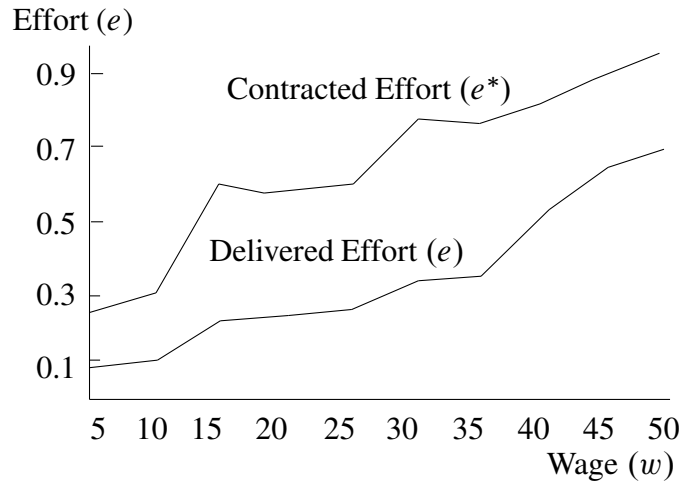

Figure 3.3. Relation of contracted and delivered effort to worker wage (141 subjects). From Fehr, Gächter, and Kirchsteiger (1997).

Figure 3.3 also shows that, though most employees are strong reciprocators, at any wage rate there still is a significant gap between the amount of

[2]This is because the experimenters created more employees than employers, thus ensuring an excess supply of employees.

effort agreed upon and the amount actually delivered. This is not because there are a few "bad apples" among the set of employees but because only 26% of the employees delivered the level of effort they promised! We conclude that strong reciprocators are inclined to compromise their morality to some extent.

To see if employers are also strong reciprocators, the authors extended the game 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 because they would not (knowingly) interact with the same worker a second time. However, 68% of the time, employers punished employees who did not fulfill their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Employers rewarded 41% 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 71% to 26% of the exchanges, and overfulfilled contracts rose from 3% to 38% of the total. Finally, allowing employers to reward and punish led to a 40% increase in 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 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 reward good behavior and punish bad behavior when they are allowed, and employees expect this behavior and adjust their own effort levels accordingly. In general, then, subjects follow an internalized norm not because it is prudent or useful to do so, or because they will suffer some material loss if they do not, but rather because they desire to do this *for its own sake*.

3.8 Altruistic Third-Party Punishment

Prosocial behavior in human society occurs not only because those directly helped and harmed by an individual's actions are likely to reciprocate in

kind but also because there are general *social norms* that foster prosocial behavior and many people are willing to bestow favors on someone who conforms to social norms, and to punish someone who does not, even if they are not personally helped or hurt by the individual's actions. In everyday life, third parties who are not the beneficiaries of an individual's prosocial act, help the individual and his family in times of need, preferentially trade favors with the individual, and otherwise reward the individual in ways that are not costly but are nonetheless of great benefit to the cooperator. Similarly, third parties who have not been personally harmed by the selfish behavior of an individual refuse aid even when it is not costly to do so, shun the offender, and approve of the offender's ostracism from beneficial group activities, again at low cost to the third party but at high cost to the offender.

It is hard to conceive of human societies operating at a high level of efficiency in the absence of such third-party reward and punishment. Yet, self-regarding actors will never engage in such behavior if it is at all costly. Fehr and Fischbacher (2004) addressed this question by conducting a series of third-party punishment experiments using the Prisoner's Dilemma (§2.10) and the dictator game (§3.4). The experimenters implemented four experimental treatments in each of which subjects were grouped into threes. In each group, in stage 1, subject A played a Prisoner's Dilemma or the Dictator Game with subject B as the Receiver, and subject C was an outsider whose payoff was not affected by A's decision. Then, in stage two, subject C was endowed with 50 points and allowed to deduct points from subject A such that every 3 points deducted from A's score cost C 1 point. In the first treatment, TP-DG, the game was the Dictator Game, in which A was endowed with 100 points, and could give 0, 10, 20, 30, 40, or 50 points to B, who had no endowment.

The second treatment (TP-PD) was the same, except that the game was the Prisoner's Dilemma. Subjects A and B were each endowed with 10 points, and each could either keep the 10 points or transfer them to the other subject, in which case the points were tripled by the experimenter. Thus, if both cooperated, each earned 30 points, and if both defected, each earned 10 points. If one cooperated and one defected, however, the cooperator earned 0 points and the defector earned 40 points. In the second stage, C was given an endowment of 40 points, and was allowed to deduct points from A and/or B, just as in the TP-DG treatment.

To compare the relative strengths of second- and third-party punishment in the Dictator Game, the experimenters implemented a third treatment,

S&P-DG. In this treatment, subjects were randomly assigned to player A and player B, and A-B pairs were randomly formed. In the first stage of this treatment, each A was endowed with 100 points and each B with none, and the A's played the Dictator Game as before. In the second stage of each treatment, each player was given an additional 50 points, and the B players were permitted to deduct points from A players on the same terms as in the first two treatments. S&P-DG also had two conditions. In the S condition, a B player could punish only his *own* Dictator, whereas in the T condition, a B player could punish only an A player *from another pair*, to which he was randomly assigned by the experimenters. In the T condition, each B player was informed of the behavior of the A player to which he was assigned.

To compare the relative strengths of second and third-party punishment in the Prisoner's Dilemma, the experimenters implemented a fourth treatment, S&P-PG. This was similar to the S&P-DG treatment, except that now they played the Prisoner's Dilemma.[3]

In the first two treatments, because subjects were randomly assigned to positions A, B, and C, the obvious fairness norm is that all should have equal payoffs (an *equality norm*). For instance, if A gave 50 points to B and C deducted no points from A, each subject would end up with 50 points. In the Dictator Game treatment, TP-DG, 60% of third parties (Cs) punished Dictators (As) who give less than 50% of the endowment to Receivers (Bs). Statistical analysis (ordinary least squares regression) showed that for every point an A kept for himself above the 50-50 split, he was punished an average 0.28 points by C's, leading to a total punishment of $3 \times 0.28 = 0.84$ points. Thus, a Dictator who kept the whole 100 points would have $0.84 \times 50 = 42$ points deducted by C's, leaving a meager gain of 8 points over equal sharing.

The results for the Prisoner's Dilemma treatment, TP-PD, was similar, with an interesting twist. If one partner in the A-B pair defected and the other cooperated, the defector would have on average 10.05 points deducted by Cs, but if both defected, the punished player lost only an average of 1.75 points. This shows that third parties (Cs) cared not only about the intentions of defectors but also about how much harm they caused and/or how unfair they turned out to be. Overall, 45.8% of third parties punished defectors

[3]The experimenters never used value-laden terms such as "punish" but rather used neutral terms, such as "deduct points."

whose partners cooperated, whereas only 20.8% of third parties punished defectors whose partners defected.

Turning to the third treatment (S&P-DG), second-party sanctions of selfish Dictators were found to be considerably stronger than third-party sanctions, although both were highly significant. On average, in the first condition, where Receivers could punish their own Dictators, they imposed a deduction of 1.36 points for each point the Dictator kept above the 50-50 split, whereas they imposed a deduction of only 0.62 point per point kept on third-party Dictators. In the final treatment, S&P-PD, defectors were severely punished by both second and third parties, but second-party punishment was again found to be much more severe than third-party punishment.. Thus, cooperating subjects deducted on average 8.4 points from a defecting partner, but only 3.09 points from a defecting third party.

This study confirms the general principle that punishing norm violators is very common but not universal, and that individuals are prone to be more harsh in punishing those who hurt them personally, as opposed to violating a social norm that hurts others than themselves.

3.9 Altruism and Cooperation in Groups

A *Public Goods Game* is an n-person game in which, by cooperating, each individual A adds more to the payoff of the other members than A's cost of cooperating, but A's share of the total gains he creates is less than his cost of cooperating. By not contributing, the individual incurs no personal cost and produces no benefit for the group. The Public Goods Game captures many social dilemmas, such as voluntary contribution to team and community goals. Researchers (Ledyard 1995; Yamagishi 1986; Ostrom, Walker, and Gardner 1992; Gächter and Fehr 1999) uniformly found that groups exhibit a much higher rate of cooperation than can be expected assuming the standard model of the self-regarding actor.

A typical Public Goods Game consists of a number of rounds, say 10. In each round, each subject is grouped with several other subjects—say 3 others. Each subject is then given a certain number of points, say 20, 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 "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, say 40%, of the

total amount in the common account. So if a subject contributes his whole 20 points to the common account, each of the 4 group members will receive 8 points at the end of the round. In effect, by putting the whole endowment into the common account, a player loses 12 points but the other 3 group members gain in total 24 (8 times 3) points. The players keep whatever is in their private accounts at the end of the round.

A self-regarding player contributes nothing to the common account. However, only a fraction of the subjects in fact conform to the self-regarding 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 10 rounds until in the final rounds most players are behaving in a self-regarding manner. This is, of course, exactly what is predicted by the strong reciprocity model. Because they are altruistic contributors, strong reciprocators start out by contributing to the common pool, but in response to the norm violation of the self-regarding types, they begin to refrain from contributing themselves.

How do we know that the decay of cooperation in the Public Goods Game is due to cooperators punishing free riders by refusing to contribute themselves? Subjects often report this behavior retrospectively. More compelling, however, is the fact that when subjects are given a more constructive way of punishing defectors, they use it in a way that helps sustain cooperation (Orbell, Dawes, and Van de Kragt 1986, Sato 1987, and Yamagishi 1988a, 1988b, 1992).

For instance, in Ostrom, Walker, and Gardner (1992) subjects in a Public Goods Game, by paying a "fee," could impose costs on others by "fining" them. Because fining costs the individual who uses it but the benefits of increased compliance accrue to the group as a whole, the only subgame perfect Nash equilibrium in this game is for no player to pay the fee, so no player is ever punished for defecting, and all players defect by contributing nothing to the public account. However, the authors found a significant level of punishing behavior. The experiment was then repeated with subjects being allowed to communicate without being able to make binding agreements. In the framework of the self-regarding actor model, such communication is called *cheap talk* and cannot lead to a distinct subgame perfect equilibrium. But in fact such communication led to almost perfect cooperation (93%) with very little sanctioning (4%).

The design of the Ostrom-Walker-Gardner study allowed individuals to engage in strategic behavior because costly punishment of defectors could

increase cooperation in future periods, yielding a positive net return for the punisher. What happens if we remove any possibility of punishment being strategic? This is exactly what Fehr and Gächter (2000) studied.

Fehr and Gächter (2000) set up an experimental situation in which the possibility of strategic punishment was removed. They used 6- and 10-round Public Goods Games with groups of size 4, and with costly punishment allowed at the end of each round, employing three different methods of assigning members to groups. There were sufficient subjects to run between 10 and 18 groups simultaneously. Under the Partner treatment, the four subjects remained in the same group for all 10 periods. Under the Stranger treatment, the subjects were randomly reassigned after each round. Finally, under the Perfect Stranger treatment, the subjects were randomly reassigned but assured that they would never meet the same subject more than once.

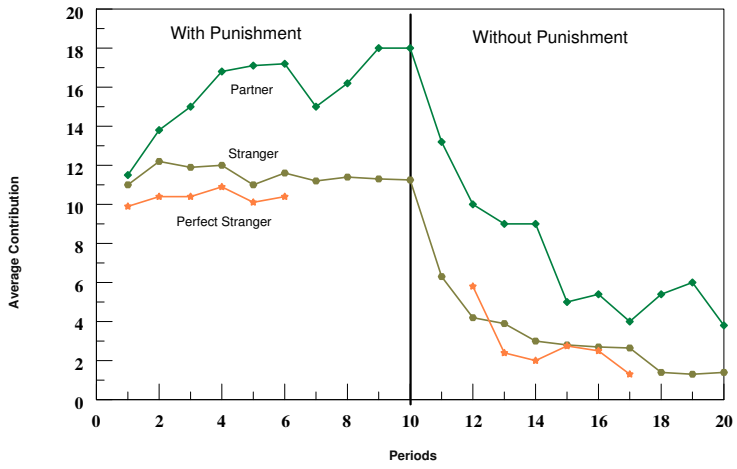

Figure 3.4. Average contributions over time in the Partner, Stranger, and Perfect Stranger Treatments when the punishment condition is played first (Fehr and Gächter 2000).

Fehr and Gächter (2000) performed their experiment for 10 rounds with punishment and 10 rounds without. Their results are illustrated in figure 3.4. We see that when costly punishment is permitted, cooperation does not deteriorate, and in the Partner game, despite strict anonymity, cooperation increases almost to full cooperation even in the final round. When punishment is not permitted, however, the same subjects experienced the deterioration of cooperation found in previous Public Goods Games. The contrast in cooperation rates between the Partner treatment 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 in this treatment the punished subjects are certain that, once they have been punished in previous rounds, the punishing subjects are in their group. The prosociality impact of strong reciprocity on cooperation is thus more strongly manifested, the more coherent and permanent the group in question.[4]

Many behavioral game theorists have found that, while altruistic punishment increases participation, it often leads to such a high level of punishment that overall average payoffs, net of punishment, are low (Carpenter and Matthews 2005; Page, Putterman, and Unel 2005; Casari and Luini 2007; Anderson and Putterman 2006; Nikiforakis 2008). Some have interpreted this as showing that strong reciprocity "could not have evolved," or "is not an adaptation." It is more likely, however, that the problem is with the experiments themselves. These experiments attempt to refute the standard "homo economicus" model of the self-regarding actor and do not attempt to produce realistic punishment scenarios in the laboratory. In fact, the motive for punishing norm violators is sufficiently strong as to lower overall payoffs when not subject to some social regulation. In real societies, there tends to be collective control over the meting out of punishment, and the excessive zeal of individual punishers is frowned upon and socially punished. Indeed, in one of the rare studies that allowed groups to regulate punishment, Ertan, Page, and Putterman (2005) found that groups that voted to permit only punishment of below-average or of average and below-average contributors achieved significantly higher earnings than groups not using punishment.

[4]In Fehr and Gächter (2002), the experimenters reverse the order of the rounds with and without punishment to be sure that the decay in the "without punishment" phase was not due to its occurring at the end rather than at the start of the game. It was not.

3.10 Inequality Aversion

The inequality-averse individual exhibits a *weak* urge to reduce inequality when on top and a *strong* urge to reduce inequality when on the bottom (Loewenstein, Thompson, and Bazerman 1989). Since the advent of hierarchical societies based on settled agriculture, societies have attempted to inculcate in their less fortunate members precisely the opposite values—subservience to and acceptance of the status quo. The widely observed distaste for relative deprivation is thus probably a genetically based behavioral characteristic of humans. Because small children spontaneously share (even the most sophisticated of nonhuman primates, such as chimpanzees, fail to do this), the urge of the fortunate to redistribute may also be part of human nature, though doubtless a weaker impulse in most of us.

Support for inequality aversion comes from the anthropological literature. *H. sapiens* evolved in small hunter-gatherer groups. Contemporary groups of this type, although widely dispersed throughout the world, display many common characteristics. This commonality probably reflects their common material conditions. From this and other considerations we may tentatively infer the social organization of early human society from that of these contemporary foraging societies (Woodburn 1982; Boehm 1982, 2000).

Such societies have no centralized structure of governance (state, judicial system, church, Big Man), so the enforcement of norms depends on the voluntary participation of peers. There are many unrelated individuals, so cooperation cannot be explained by kinship ties. Status differences are very circumscribed, monogamy is widely enforced,[5] members who attempt to acquire personal power are banished or killed, and there is widespread sharing of large game and other food sources that are subject to substantial stochasticity, independent of the skill and/or luck of the hunters. Such conditions are, of course, conducive to the emergence of inequality aversion.

We model inequality aversion following Fehr and Schmidt (1999). Suppose the monetary payoffs to n players are given by $\pi = (\pi_1, \ldots, \pi_n)$. We take the utility function of player i to be

$$u_i(\pi) = \pi_i - \frac{\alpha_i}{n-1} \sum_{\pi_j > \pi_i} (\pi_j - \pi_i) - \frac{\beta_i}{n-1} \sum_{\pi_j < \pi_i} (\pi_i - \pi_j). \qquad (3.1)$$

[5]Monogamy in considered to be an extremely egalitarian institution for men because it ensures that virtually all adult males will have a wife.

A reasonable range of values for β_i is $0 \leq \beta_i < 1$. Note that when $n = 2$ and $\pi_i > \pi_j$, if $\beta_i = 0.5$, then i is willing to transfer income to j dollar for dollar until $\pi_i = \pi_j$, and if $\beta_i = 1$ and i has the highest payoff, then i is willing to throw away money (or give it to the other players) at least until $\pi_i = \pi_j$ for some player j. We also assume $\beta_i < \alpha_i$, reflecting the fact that people are more sensitive to inequality when on the bottom than when on the top.

We shall show that with these preferences we can reproduce some of the salient behaviors in the Ultimatum and Public Goods games, where fairness appears to matter, as well as in market games, where it does not.

Consider first the Ultimatum Game. Let y be the share the Proposer offers the Responder, so the Proposer gets $x = 1 - y$. Because $n = 2$, we can write the two utility functions as

$$u(x) = \begin{cases} x - \alpha_1(1 - 2x) & x \leq 0.5 \\ x - \beta_1(2x - 1) & x > 0.5 \end{cases} \tag{3.2}$$

$$v(y) = \begin{cases} y - \alpha_2(1 - 2y) & y \leq 0.5 \\ y - \beta_2(2y - 1) & y > 0.5 \end{cases} \tag{3.3}$$

We have the following theorem.

THEOREM 3.1 *Suppose the payoffs in the Ultimatum Game are given by (3.2) and (3.3) and α_2 is uniformly distributed on the interval $[0, \alpha^*]$. Writing $y^* = \alpha^*/(1 + 2\alpha^*)$, we have the following:*

a. If $\beta_1 > 0.5$, the Proposer offers $y = 0.5$.
b. If $\beta_1 = 0.5$, the Proposer offers $y \in [y^*, 0.5]$.
c. If $\beta_1 < 0.5$, the Proposer offers y^*.

In all cases the Responder accepts. We leave the proof, which is straightforward, to the reader.

Now suppose we have a Public Goods Game $\mathcal{G}$ with $n \geq 2$ players. Each player i is given an amount 1 and decides independently what share x_i to contribute to the public account, after which the public account is multiplied by a number a, with $1 > a > 1/n$, and shared equally among the players. Because $1 > a$, contributions are costly to the contributor, and because $na > 1$, the group benefits of contributing exceed the costs, so contributing is a public good. The monetary payoff for each player then becomes $\pi_i = 1 - x_i + a \sum_{j=1}^{n} x_j$, and the utility payoffs are given by (3.1). We then have the following theorem.

THEOREM 3.2 *In the n-player Public Goods Game $\mathcal{G}$,*

a. *If $\beta_i < 1 - a$ for player i, then contributing nothing to the public account is a dominant strategy for i (a strategy is* dominant *for player i if it is a best response to any strategy profile of the other players).*
b. *If there are $k > a(n-1)/2$ players with $\beta_i < 1 - a$, then the only Nash equilibrium is for all players to contribute nothing to the public account.*
c. *If there are $k < a(n-1)/2$ players with $\beta_i < 1 - a$ and if all players i with $\beta_i > 1 - a$ satisfy $k/(n-1) < (a + \beta_i - 1)/(\alpha_i + \beta_i)$, then there is a Nash equilibrium in which the latter players contribute all their money to the public account.*

Note that if a player has a high β and hence could possibly contribute, but also has a high α so the player strongly dislikes being below the mean, then condition $k/(n-1) < (a+\beta_i-1)/(\alpha_i+\beta_i)$ in part (c) of the theorem will fail. In other words, cooperation with defectors requires that contributors not be excessively sensitive to relative deprivation.

The proof of this theorem is a bit tedious but straightforward and will be left to the reader (or consult Fehr and Schmidt 1999). We prove only part (c). We know from part (a) that players i with $\beta_i < 1-a$ will not contribute. Suppose $\beta_i > 1 - a$ and assume all other players satisfying this inequality contribute all their money to the public account. By reducing his contribution by $\delta > 0$, player i saves $(1 - a)\delta$ directly and receives $k\alpha_i\delta/(n-1)$ in utility from the higher returns compared to the noncontributors, minus $(n-k-1)\delta\beta_i/(n-1)$ in utility from the lower returns compared with the contributors. The sum must be nonpositive in a Nash equilibrium, which reduces to the inequality in part (c).

Despite the fact that players have egalitarian preferences given by (3.1) if the game played has sufficiently marketlike qualities, the unique Nash equilibrium may settle on the competitive equilibrium however unfair this appears to be to the participants. Consider the following theorem.

THEOREM 3.3 *Suppose preferences are given by (3.1) and that \$1 is to be shared between player 1 and one of the players $i = 2, \ldots, n$ who submit simultaneous bids y_i for the share they are willing to give to player 1. The highest bid wins, and among equal highest bids, the winner is drawn at random. Then, for any set of (α_i, β_i), in every subgame perfect Nash equilibrium player 1 receives the whole \$1.*

The proof is left to the reader. Show that at least two bidders will set their y_i's to 1, and the seller will accept this offer.

3.11 The Trust Game

In the Trust Game, first studied by Berg, Dickhaut, and McCabe (1995), subjects are each given a certain endowment, say \$10. Subjects are then randomly paired, and one subject in each pair, Alice, is told she can transfer any number of dollars, from 0 to 10, to her (anonymous) partner, Bob, and keep the remainder. The amount transferred will be tripled by the experimenter and given to Bob, who can then give any number of dollars back to Alice (this amount is not tripled). If Alice transfers a lot, she is called "trusting," and if Bob returns a lot to Alice, he is called "trustworthy." In the terminology of this chapter, a trustworthy player is a strong reciprocator, and a trusting player is an individual who expects his partner to be a strong reciprocator.

If all individuals have self-regarding preferences, and if Alice believes Bob has self-regarding preferences, she will give nothing to Bob. On the other hand, if Alice believes Bob can be trusted, she will transfer all \$10 to Bob, who will then have \$40. To avoid inequality, Bob will give \$20 back to Alice. A similar result will obtain if Alice believes Bob is a strong reciprocator. On the other hand, if Alice is altruistic, she may transfer some money to Bob, on the grounds that it is worth more to Bob (because it is tripled) than it is to her, even if she does not expect anything back. It follows that several distinct motivations can lead to a positive transfer of money from Alice to Bob and then back to Alice.

Berg, Dickhaut, and McCabe (1995) found that, on average, \$5.16 was transferred from Alices to Bobs and on average, \$4.66 was transferred back from Bobs to Alices. Furthermore, when the experimenters revealed this result to the subjects and had them play the game a second time, \$5.36 was transferred from Alices to Bobs, and \$6.46 was transferred back from Bobs to Alices. In both sets of games there was a great deal of variability: some Alices transferring everything and some transferring nothing, and some Bobs more than fully repaying their partners, and some giving back nothing.

Note that the term "trustworthy" applied to Bob is inaccurate because Bob never, either explicitly or implicitly, promised to behave in any particular manner, so there is nothing concrete that Alice might trust him to do. The

Trust Game is really a strong reciprocity game in which Alice believes with some probability that Bob is a sufficiently motivated strong reciprocator and Bob either does or does not fulfill this expectation. To turn this into a real Trust Game, the second player should be able to promise to return a certain fraction of the money passed to him. We investigate this case in §3.12.

To tease apart the motivations in the Trust Game, Cox (2004) implemented three treatments, the first of which, treatment A, was the Trust Game as described above. Treatment B was a Dictator Game (§3.8) exactly like treatment A, except that now Bob could not return anything to Alice. Treatment C differs from treatment A in that each Alice was matched one-to-one with an Alice in treatment A, and each Bob was matched one-to-one with a Bob in treatment A. Each player in treatment C was then given an endowment equal to the amount his corresponding player had after the A-to-B transfer, but before the B-to-A transfer in treatment A. In other words, in treatment C, the Alice group and the Bob group have exactly what they had under treatment A, except that Alice now had nothing to do with Bob's endowment, so nothing transferred from Bob to Alice could be accounted for by strong reciprocity.

In all treatments, the rules of the game and the payoffs were accurately revealed to the subjects. However, in order to rule out third-party altruism (§3.8), the subjects in treatment C were not told the reasoning behind the sizes of their endowments. There were about 30 pairs in each treatment, each treatment was played two times, and no subject participated in more than one treatment. The experiment was run double-blind (subjects were anonymous to one another and to the experimenter).

In treatment B, the Dictator Game counterpart to the Trust Game, Alice transferred on average \$3.63 to player B, as opposed to \$5.97 in treatment A. This shows that \$2.34 of the \$5.97 transferred to B in treatment A can be attributed to trust, and the remaining \$3.63 to some other motive. Because players A and B both have endowments of \$10 in treatment B this other motive cannot be inequality aversion. This transfer may well reflect a reciprocity motive of the form, "If someone can benefit his partner at a cost that is low compared to the benefit, he should do so, even if he is on the losing end of the proposition." But we cannot tell from the experiment exactly what the \$3.63 represents.

In treatment C, the player B Dictator Game counterpart to the Trust Game, player B returned an average of \$2.06, as compared with \$4.94 in

treatment A. In other words, \$2.06 of the original \$4.94 can be interpreted as a reflection of inequality aversion, and the remaining \$2.88 is a reflection of strong reciprocity.

Several other experiments confirm that other-regarding preferences depend on the actions of individuals and not simply on the distribution of payoffs, as is the case with inequality aversion. Charness and Haruvy (2002), for instance, developed a version of the gift exchange labor market described in §3.7 capable of testing self-regarding preferences, pure altruism, inequality aversion, and strong reciprocity simultaneously. Strong reciprocity had by far the greatest explanatory value.

3.12 Character Virtues

Character virtues are ethically desirable behavioral regularities that individuals value for their own sake, while having the property of facilitating cooperation and enhancing social efficiency. Character virtues include *honesty*, *loyalty*, *trustworthiness*, *promise keeping*, and *fairness*. Unlike such other-regarding preferences as strong reciprocity and empathy, these character virtues operate without concern for the individuals with whom one interacts. An individual is honest in his transactions because this is a desired state of being, not because he has any particular regard for those with whom he transacts. Of course, the sociopath "Homo economicus" is honest only when it serves his material interests to be so, whereas the rest of us are at times honest even when it is costly to be so and even when no one but us could possibly detect a breach.

Common sense, as well as the experiments described below, indicate that honesty, fairness, and promise keeping are not absolutes. If the cost of virtue is sufficiently high, and the probability of detection of a breach of virtue is sufficiently small, many individuals will behave dishonestly. When one is aware that others are unvirtuous in a particular region of their lives (e.g., marriage, tax paying, obeying traffic rules, accepting bribes), one is more likely to allow one's own virtue to lapse. Finally, the more easily one can delude oneself into inaccurately classifying an unvirtuous act as virtuous, the more likely one is to allow oneself to carry out such an act.

One might be tempted to model honesty and other character virtues as *self-constituted constraints* on one's set of available actions in a game, but a more fruitful approach is to include the state of being virtuous in a certain way as an argument in one's preference function, to be traded off against

other valuable objects of desire and personal goals. In this respect, character virtues are in the same category as ethical and religious preferences and are often considered subcategories of the latter.

Numerous experiments indicate that most subjects are willing to sacrifice material rewards to maintain a virtuous character even under conditions of anonymity. Sally (1995) undertook a meta-analysis of 137 experimental treatments, finding that face-to-face communication, in which subjects are capable of making verbal agreements and promises, was the strongest predictor of cooperation. Of course, face-to-face interaction violates anonymity and has other effects besides the ability to make promises. However, both Bochet, Page, and Putterman (2006) and Brosig, Ockenfels, and Weimann (2003) report that only the ability to exchange verbal information accounts for the increased cooperation.

A particularly clear example of such behavior is reported by Gneezy (2005), who studied 450 undergraduate participants paired off to play three games of the following form, all payoffs to which were of the form (b, a), where player 1, Bob, receives b and player 2, Alice, receives a. In all games, Bob was shown two pairs of payoffs, A:(x, y) and B:(z, w) where x, y, z, and w are amounts of money with $x < z$ and $y > w$, so in all cases B is better for Bob and A is better for Alice. Bob could then say to Alice, who could not see the amounts of money, either "Option A will earn you more money than option B," or "Option B will earn you more money than option A." The first game was A:(5,6) vs. B:(6,5) so Bob could gain 1 by lying and being believed while imposing a cost of 1 on Alice. The second game was A:(5,15) vs. B:(6,5), so Bob could gain 1 by lying and being believed, while still imposing a cost of 10 on Alice. The third game was A:(5,15) vs. B:(15,5), so Bob could gain 10 by lying and being believed, while imposing a cost of 10 on Alice.

Before starting play, Gneezy asked the various Bobs whether they expected their advice to be followed. He induced honest responses by promising to reward subjects whose guesses were correct. He found that 82% of Bobs expected their advice to be followed (the actual number was 78%). It follows from the Bobs' expectations that if they were self-regarding, they would always lie and recommend B to Alice.

The experimenters found that, in game 2, where lying was very costly to Alice and the gain from lying was small for Bob, only 17% of Bobs lied. In game 1, where the cost of lying to Alice was only 1 but the gain to Bob was the same as in game 2, 36% of Bobs lied. In other words, Bobs

were loathe to lie but considerably more so when it was costly to Alices. In game 3, where the gain from lying was large for Bob and equal to the loss to Alice, fully 52% of Bobs lied. This shows that many subjects are willing to sacrifice material gain to avoid lying in a one-shot anonymous interaction, their willingness to lie increasing with an increased cost to them of truth telling, and decreasing with an increased cost to their partners of being deceived. Similar results were found by Boles, Croson, and Murnighan (2000) and Charness and Dufwenberg (2004). Gunnthorsdottir, McCabe, and Smith (2002) and Burks, Carpenter, and Verhoogen (2003) have shown that a socio-psychological measure of "Machiavellianism" predicts which subjects are likely to be trustworthy and trusting.

3.13 The Situational Character of Preferences

This chapter has deepened the rational actor model, allowing it to apply to situations of strategic interaction. We have found that preferences are other-regarding as well as self-regarding. Humans have social preferences that facilitate cooperation and exchange, as well as moral preferences for such personal character virtues as honesty and loyalty. These extended preferences doubtless contribute to longrun individual well-being (Konow and Earley 2008). However, social and moral preferences are certainly not merely instrumental, because individuals exercise these preferences even when no longrun benefits can accrue.

Despite this deepening of rational choice, we have conserved the notion that the individual has an immutable underlying preferences ordering that entails situationally specific behaviors, depending on the particular strategic interaction involved. Our analysis in §7.8, however, is predicated upon the denial of this immutability. Rather, we suggest that generally a social situation, which we call a *frame*, is imbued with a set of customary social norms that individuals often desire to follow simply because these norms are socially appropriate in the given frame. To the extent that this occurs, preferences themselves, and not just their behavioral implications, are situationally specific. The desire to conform to the moral and conventional standards that people associate with particular social frames thus represents a *meta-preference* that regulates revealed preferences in specific social situations.

We present two studies by Dana, Cain, and Dawes (2006) that illustrate the situational nature of preferences and the desire to conform to so-

cial norms (which we term *normative predisposition* in chapter 7). The first study used 80 Carnegie-Mellon University undergraduate subjects who were divided into 40 pairs to play the Dictator Game (§3.4), one member of each pair being randomly assigned to be the Dictator, the other to be the Receiver. Dictators were given $10, and asked to indicate how many dollars each wanted to give the Receiver, but the Receivers were not informed they were playing a Dictator Game. After making their choices, but before informing the Receivers about the game, the Dictators were presented with the option of accepting $9 rather than playing the game. They were told that if a Dictator took this option, the Receiver would never find out that the game was a possibility and would go home with their show-up fee alone.

Eleven of the 40 Dictators took this exit option, including 2 who had chosen to keep all of the $10 in the Dictator Game. Indeed, 46% of the Dictators who had chosen to give a positive amount to their Receivers took the exit option in which the Receiver got nothing. This behavior is not compatible with the concept of immutable preferences for a division of the $10 between the Dictator and the Receiver because individuals who would have given their Receiver a positive amount in the Dictator Game instead gave them nothing by avoiding playing the game, and individuals who would have kept the whole $10 in the Dictator Game were willing to take a $1 loss not to have to play the game.

To rule out other possible explanations of this behavior, the authors executed a second study in which the Dictator was told that the Receiver would never find out that a Dictator Game had been played. Thus, if the Dictator gave $5 to the Receivers, the latter would be given the $5 but would be given no reason why. In this new study, only 1 of 24 Dictators chose to take the $9 exit option. Note that in this new situation, the same social situation between Dictator and Receiver obtains both in the Dictator Game and in the exit option. Hence, there is no difference in the norms applying to the two options, and it does not make sense to forfeit $1 simply to have the game not called a Dictator Game.

The most plausible interpretation of these results is that many subjects felt obliged to behave according to certain norms when playing the Dictator Game, or violated these norms in an uncomfortable way, and were willing to pay simply not to be in a situation subject to these norms.

3.14 The Dark Side of Altruistic Cooperation

The human capacity to cooperate in large groups by virtue of prosocial preferences extends not only to exploiting nature but also to conquering other human groups as well. Indeed, even a slight hint that there may be a basis for inter-group competition induces individuals to exhibit insider loyalty and outsider hostility (Dawes, de Kragt, and Orbell 1988; Tajfel 1970; Tajfel et. al 1971; Turner 1984). Group members then show more generous treatment to in-group members than to out-group members even when the basis for group formation is arbitrary and trivial (Yamagishi, Jin, and Kiyonari 1999; Rabbie, Schot, and Visser 1989).

An experiment conducted by Abbink et. al al (2007), using undergraduate students recruited at the University of Nottingham, is an especially dramatic example of the tendency for individuals willingly to escalate a conflict well beyond the point of serving their interests in terms of payoffs alone. Experimenters first had pairs of students $i = 1, 2$ play the following game. Each individual was given 1000 points and could spend any portion of it, x_i, on "armaments." The probability of player i winning was then set to $p_i = x_i/(x_1 + x_2)$.

We can find the Nash equilibrium of this game as follows. If player 1 spends x_1, then the expenditure of player 2 that maximizes the expected payoff is given by

$$x_2^* = \sqrt{1000x_1} - x_1.$$

The symmetric Nash equilibrium sets $x_1^* = x_2^*$, which gives $x_1^* = x_2^* = 250$. Indeed, if one player spends more than 250 points, the other player's best response is to spend less than 250 points.

Fourteen pairs of subjects played this game in pairs for 20 rounds, each with the same partner. The average per capita armament expenditure started at 250% of the Nash equilibrium in round 1 and showed some tendency to decline, reaching 160% of the Nash level after 20 rounds.

The experimenters also played the same game with 4 players on each team, where each player on the winning team received 1000 points. It is easy to show that now the Nash equilibrium has each team spending 250 points on armaments. To see this, we write player 1's expected payoff as

$$\frac{1000 \sum_{i=1}^{4} x_i}{\sum_{i=1}^{8}}.$$

Differentiating this expression, setting the result to zero, and solving for x_1 gives

$$x_1 = \sqrt{1000(x_5 + x_6 + x_7 + x_8)} - \sum_{i=2}^{8} x_i .$$

Now, equating all the x_i's to find the symmetric equilibrium, we find $x_i^* = 62.5 = 250/4$. In this case, however, the teams spent about 600% of the optimum in the first few periods, and this declined fairly steadily to 250% of the optimum in the final few periods.

This experiment showcases the tendency of subjects to overspend vastly for competitive purposes, although familiarity with the game strongly dampens this tendency, and had the participants played another 20 periods, we might have seen an approach to best response behavior.

However, the experimenters followed up the above treatments with another in which, after each round, players were allowed to punish other players based on the level of their contributions in the previous period. The punishment was costly, three tokens taken from the punishee costing the punisher one token. This, of course, mirrors the Public Goods Game with costly punishment (§3.9), and indeed this game does have a public goods aspect since the more one team member contributes, the less the best response contribution of the others, because the optimal total contribution of team members is 250, no matter how it is divided up among the members.

In this new situation, competition with punishment, spending started at 640% of the best response level, rose to a high of 1000% of this level, and settled at 900% of the best response level in period 7, showing no tendency to increase or decrease in the remaining 13 periods. This striking behavior shows that the internal dynamics of altruistic punishment are capable of sustaining extremely high levels of combat expenditure far in excess of the material payoff-maximizing level. While much more work in this area remains to be done, it appears that the same prosocial preferences that allow humans to cooperate in large groups of unrelated individuals are also turned into the goal of mutual self-destruction with great ease.

3.15 Norms of Cooperation: Cross-Cultural Variation

Experimental results in the laboratory would not be very interesting if they did not aid us in understanding and modeling real-life behavior. There are strong and consistent indications that the external validity of experimental results is high. For instance, Binswanger (1980) and Binswanger and

Sillers (1983) used survey questions concerning attitudes towards risk and experimental lotteries with real financial rewards to successfully predict the investment decisions of farmers. Glaeser et. al al (2000) explored whether experimental subjects who trusted others in the Trust Game (§3.11) also behaved in a trusting manner with their own personal belongings. The authors found that experimental behavior was a quite good predictor of behavior outside the laboratory, while the usual measures of trust, based on survey questions, provided virtually no information. Genesove and Mayer (2001) showed that loss aversion determined seller behavior in the 1990s Boston housing market. Condominium owners subject to nominal losses set selling prices equal to the market rate plus 25% to 35% of the difference between their purchase price and the market price and sold at prices 3% to 18% of this difference. These findings show that loss aversion is not confined to the laboratory but affects behavior in a market in which very high financial gains and losses can occur.

Similarly, Karlan (2005) used the Trust Game and the Public Goods Game to predict the probability that loans made by a Peruvian microfinance lender would be repaid. He found that individuals who were trustworthy in the Trust Game were less likely to default. Also, Ashraf, Karlan, and Yin (2006) studied Phillipino women, identifying through a baseline survey those women exhibited a lower discount rate for future relative to current tradeoffs. These women were indeed significantly more likely to open a savings account, and after 12 months, average savings balances increased by 81 percentage points for those clients assigned to a treatment group based on their laboratory performance, relative to those assigned to the control group. In a similar vein, Fehr and Goette (2007) found that in a group of bicycle messengers in Zürich, those and only those who exhibited loss aversion in a laboratory survey also exhibited loss aversion when faced with real-life wage rate changes. For additional external validity studies, see Andreoni, Erard, and Feinstein (1998) on tax compliance (§3.4), Bewley (2000) on fairness in wage setting, and Fong, Bowles, and Gintis (2005) on support for income redistribution.

In one very important study, Herrmann, Thöni, and Gächter (2008) had subjects play the Public Goods Game with punishment (§3.9) with 16 subject pools in 15 different countries with highly varying social characteristics (one country, Switzerland, was represent by two subject pools, one in Zurich and one in St. Gallen). To minimize the social diversity among sub-

ject pools, they used university students in each country. The phenomenon they aimed to study was *anti-social punishment*.

The phenomenon itself was first noted by Cinyabuguma, Page, and Putterman (2004), who found that some free riders, when punished, responded not by increasing their contributions, but rather by punishing the high contributors! The ostensible explanation of this perverse behavior is that some free riders believe it is their personal right to free-ride if they so desire, and they respond to the "bullies" who punish them in a strongly reciprocal manner—they retaliate against their persecutors. The result, of course, is a sharp decline in the level of cooperation for the whole group.

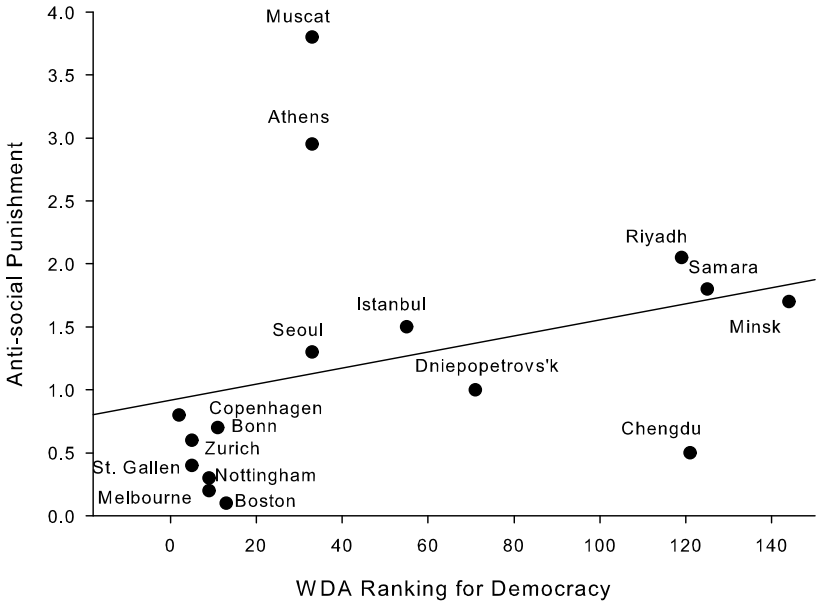

Figure 3.5. Countries judged highly democratic (political rights, civil liberties, press freedom, low corruption) by the World Democracy Audit engage in very little anti-social punishment, and conversely. (Statistics from Herrmann, Thöni, and Gächter, 2008.)

This behavior was later reported by Denant-Boemont, Masclet, and Noussair (2007) and Nikiforakis (2008), but because of its breadth, the Herrmann, Thöni, and Gächter study is distinctive for its implications for social theory. They found that in some countries, antisocial punishment was very rare, while in others it was quite common. As can be seen in fig-

ure 3.5, there is a strong negative correlation between the amount of anti-punishment exhibited and the World Development Audit's assessment of the level of democratic development of the society involved.

Figure 3.6 shows that a high level of antisocial punishment in a group translates into a low level of overall cooperation. The researchers first ran 10 rounds of the Public Goods Game without punishment (the *N* condition), and then another 10 rounds with punishment (the *P* condition). The figures show clearly that the more democratic countries enjoy a higher average payoff from payoffs in the Public Goods Game.

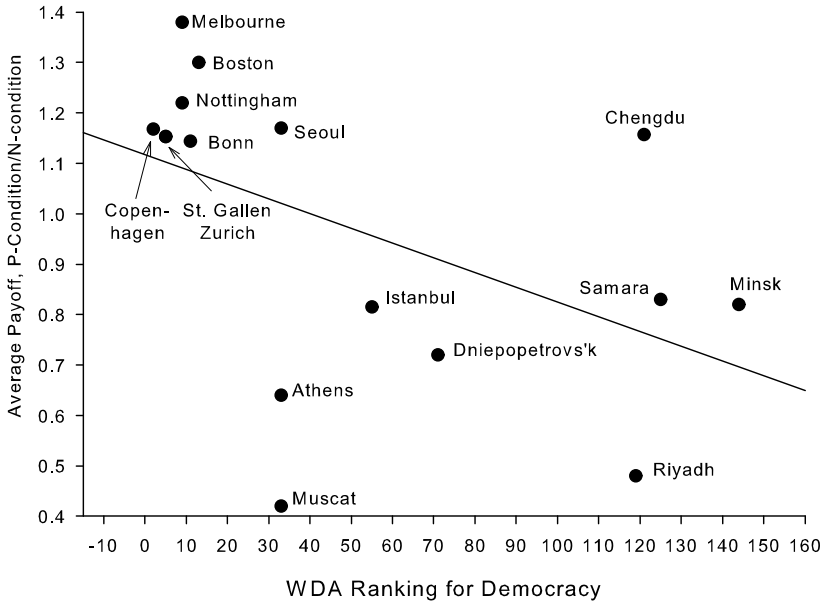

Figure 3.6. Antisocial punishment leads to low payoffs (Statistics from Herrmann, Thöni, and Gächter, Online Supplementary Material, 2008).

How might we explain this highly contrasting social behavior in university students in democratic societies with advanced market economies on the one hand, and more traditional societies based on authoritarian and parochial social institutions on the other? The success of democratic market societies may depend critically upon moral virtues as well as material interests, so the depiction of economic actors as "homo economicus" is as incorrect in real life as it is in the laboratory. These results indicate that individuals in modern democratic capitalist societies have a deep reservoir

of public sentiment that can be exhibited even in the most impersonal interactions with unrelated others. This reservoir of moral predispositions is based upon an innate prosociality that is a product of our evolution as a species, as well as the uniquely human capacity to internalize norms of social behavior. Both forces predispose individuals to behave morally, even when this conflicts with their material interests, and to react to public disapprobation for free-riding with shame and penitence rather than anti-social self-aggrandizement.

More pertinent to the purposes of behavioral game theory, this experiment shows that laboratory games can be deployed to shed light on real-life social regularities that cannot be explained by participant observation or cross-country statistical analysis alone.