Behavioral Game Theory and Sociology

Herbert Gintis*

January 26, 2006

Abstract

Behavioral game theory assumes rational choice theory, but considers the content of the preference function that the agent maximizes as the subject of empirical investigation. Many behavioral experiments indicate that subjects in experimental games behave in an other-regarding manner broadly consistent with the notion that they have internalized social values that strongly affect what they choose to optimize. The paper shows that behavioral game theory vindicates a variant of rational actor sociology in modeling cooperation and punishment in social dilemmas.

1 Introduction

"Skepticism toward sociology has grown over recent years," writes Boudon (2003). "The attention granted to rational choice theory," Boudon continues, "is, to a large extent, a reaction against this situation." Yet, having given rational choice theory a try, many practitioners urge that sociology rest content with a diminished position within the behavioral sciences, and relish its capacity to illuminate the rich subtleties of social life that are perforce ignored in the grand theories treasured by economics and the other behavioral disciplines. For instance, in his Presidential Address to the American Sociological Association, Alejandro Portes (2000):1 suggests that

> Purposive social action has provided the bedrock for theoretical development and model building in several social sciences. Since its

*I would like to thank the John D. and Catherine T. MacArthur Foundation for financial support.
Affiliations: Santa Fe Institute and Central European University.

> beginnings, however, sociology has harbored a "contrarian" vocation based on examining the unrecognized, unintended, and emergent consequence of goal-oriented activity.

While such "contrarian" contributions to behavioral science are doubtless of great value, the absence of a more theoretical and systematic contribution would surely have disappointed the such great sociological theorists as Emile Durkheim, Vilfredo Pareto, Max Weber, and Talcott Parsons. It is not as though the theoretical cores of these masters of sociological theory have been assimilated into the economic, political, or psychological theories based on rational action. Surely they have not. Nor have they been surpassed or superannuated. In fact, they have simply been ignored.

We believe that "skepticism concerning sociology" has, if anything, caused more harm *outside* of sociology than it has *within*. This is because a central theoretical sociological concept, that of *socialization*, with its attendant psychological counterpart *the internalization of norms*, is simply ignored outside of the discipline. This is, of course, not because the phenomenon does not exist. Indeed, it has been thoroughly documented that an older generation instills values in a younger generation through an extended series of personal and ritualized interactions, relying on a complex interplay of affect and authority, based on a distinctive psychological impressionability of youth (Grusec and Kuczynski 1997). How could so important a principle be ignored by other behavioral disciplines?

The answer may be a predilection in liberal thought that there is a fundamental distinction between 'juvenile' and 'adult,' and the latter alone is capable of rational behavior. Socialization, in this view, is a process that, when successful, moves prerational agents into a state of rationality. Were this true, much of sociology would not be relevant to the explanation of adult behavior. The problem, however, is that socialization does more than prepare individuals for adulthood by developing the tools of mature deliberation and evaluative choice. In addition, socialization promotes -regarding values that, when acted upon, *lead individuals to eschew narrow self-regarding actions in favor of actions that help or hurt others, at an expense to the acting agent*. This notion, so central to the sociological explanation of cooperation and conflict in social life, has been spurned by those disciplines that use rational choice theory to explain these same phenomena.

In effect, rational choice theory has historically accepted as a basic principle that *rational actors are purely self-regarding*. Indeed, the term 'rational' has often been equated with 'dispassionately self-interested,' and the irrational with the emotional, the passionate, and the altruistic acts that sometimes triumph over our

better judgment. Given this intellectual framework, It is not surprising that rational choice theory has been received so reluctantly by sociologists, for whom the internalization of norms is represents, not irrationality, but rather a higher form of human ethical being.

By a *self-regarding* actor we mean an agent in a social situation game who maximizes his own payoff. A self-regarding actor thus cares about the choices and payoffs to other participants only insofar as these impact his own payoff.[1] As we explain in this paper, recent research using rational choice theory in experimental game theoretic settings *incontravertibly rejects the equation of rationality with self-regarding behavior*.

We thus resolve the tension between rational choice theory and basic sociological theory by showing that caring about others may be just as rational as caring about oneself. The position of socialization theory in an enlarged model of rational behavior is as follows: agents maximize their utility subject to the information they possess and the material constraints they face, but what humans maximize is, in part, socially determined through the process of value internalization. In effect, humans have the characteristic that their preference structures are *socially programmable*, in much the same manner at a digital computer can be programmed, and hence exhibit highly flexible behavior.[2] Internalized norms are programmable because they are accepted by their carriers not simply as *instruments toward* or *constraints upon* achieving other ends, but rather as *arguments in the preference function that the individual maximizes*.[3] People voluntarily conform to such norms and punish those who do not. The programmability of human preferences may be an elaboration upon imprinting and imitation mechanisms found in birds and other mammals (Lorenz 1981, Bolhuis and Honey 1998), but its highly developed form in humans indicates it probably had great adaptive value during the evolutionary history of our species. The extent to which social values are internalized will, in general, depend upon biological predisposition, as well as the overall effect on well-being and biological fitness of the behaviors dictated by internalized values (Gintis 2003).

[1] We prefer the term "self-regarding" to "self-interested" because an other-regarding agent—one, for instance, who prefers others to have high payoffs at personal cost—can still be considered "self-interested," since he is acting to maximize his subjective utility. We can avoid confusion (and much shallow philosophical discussion) by employing the self-regarding/other-regarding terminology.

[2] Of course, we must never assume agents are simply passive, uncritical receptors of the forces of socialization (Wrong 1961, Gintis 1975). This issue is carefully modeled in Gintis (2003).

[3] Note that this formulation *presupposes* the validity of rational choice theory.

The reader may suspect that we achieve our goals by redefining rational choice theory in some arbitrary way to suit our needs. In the Section 2 we show that this is not the case. Rational choice theory has been successful in economics, biology, and elsewhere, because it models agents as choosing best responses to the strategic choices of other agents, or equivalently, as maximizing a preference function subject to whatever material and informational constraints the agent faces. Decision theorists have long known that this assumption substantively requires nothing more than *transitive preferences* on the part of the agent.

Many who attempt to save rational choice theory by broadening the entities that agents value (Coleman 1990, Hechter and Kanazawa 1997), are criticized for being able to explain everything, and therefore explain nothing (Smelser 1992). It is, therefore, important to explain why we succeed where others have not. In Section 3, we argue that behavioral game theory gives researchers a degree of control over the collection of data concerning human behavior that has never before been possible, and therefore allows us to specify with considerable specificity exactly what features are to be added to the usual fare of rational choice theory.

The remainder of this paper consists of the description and analysis of various experiments involving social dilemmas that allow us to specify two other-regarding behaviors, which we call *strong reciprocity* and *inequality aversion*. Section 4 describes one of many experiments that show that market exchange among anonymous agents is fully explained by assuming that trading agents are self-regarding. Each succeeding section of the paper illustrates in one way or another the proposition that whenever agents engage in strategic interaction, even under conditions of anonymity and non-repetition (so reputations cannot be formed), the resulting outcome can be explained only by assuming that agents are other-regarding, caring about the behavior of others and the overall distribution of benefits, in addition to their personal material gain.

Section 5 uses the dictator game to show that most subjects have transitive preferences over choice bundles of the form (π_s, π_o), where π_o is an amount of money you keep out of a windfall gain and the amount π_o you give to another agent. Thus, in sharing a windfall gain, rational choice theory applies quite well. Section 6 shows that many subjects are *conditional altruists*, in sense that they prefer to cooperate with others, even at a cost to themselves, if they know that their partners will also cooperate.[4]

[4]Several sociologists have told us in informal conversation that they "already knew that," and the experiments are unnecessary. We reply that researchers in other behavioral disciplines have been quite shocked and disconcerted by results of this type, and we are certain that there are many

Section 7 uses the ultimatum game to show the existence of *altruistic punishment*, in which agents retaliate against those who have acted unfairly, at personal cost, in situations where there is no way in which the retaliator can expect to gain in material terms from this behavior. Section 8 uses the concept of *strong reciprocity* to explain the behavior of subjects in a simulated labor market, where the inability to make binding contracts for work effort is offset by the willingness of agents to trust their partners to behave ethically, and this trust is generally repaid by the trustworthy behavior of the recipients. Where trust is violated, the aggrieved are willing to punish their partners, at personal cost, even when they cannot in any way gain materially from doing so.

Sections 9 and 10 extend the analysis of strong reciprocity and inequality aversion to n-player social dilemmas, where n is on the order of 3 to 10 persons. We show that a very high level of cooperation can be maintained because a fraction of the subjects are strong reciprocators who are willing to punish defectors in a dictator game (Section 9) or a public goods game (Section 10), at personal cost, thus leading the self-regarding defectors to cooperate, and sometimes even to shift from self-regarding to genuinely other-regarding behavior.

One of the most pressing needs in behavioral game theory is to move beyond student subjects to deal with a broader range of individuals in various life-stages, social classes, and societies. Section 11 relates our attempt to run behavioral experiments in fifteen small scale societies (hunter-gatherer, small-scale horticultural, nomadic, etc.) around the world. We found a degree of heterogeneity in the play of the ultimatum game, for instance, never before even approached in cross-cultural studies with student subjects. This study, more than any other, shows the importance of internalized values in explaining other-regarding behavior.

In many settings, it is difficult to determine whether other-regarding behaviors are motivated by strong reciprocity, where the intentions of the other players is important in one's decisions, as opposed to inequality aversion, where agents care only about the distribution of outcomes, and not how they were attained. Sections 12 and 13 attempt to tease apart these two motivations.

As we mentioned above, the simpler experimental games that form the staple of behavioral game theory should be treated as preludes to the study of more complex strategic interactions that better reflect the situation of agents in social life. Section 14 is an elaboration of the "gift exchange" model of Section 8 to

behaviors that sociologists "already know" that are contradicted in carefully executed experiments. Moreover, the simple behavioral game theory experiments we run today are the foundation for more subtle investigations by future researchers.

include the possibility of firms and employees developing long-term relationships. The result is completely supportive of many standard ideas in the sociology of the firm. When incomplete contracts are the rule, sociological notions of reciprocity, trust, and long-term commitment explain individual behaviors and institutional arrangements better than economic theory based on the self-regarding actor.

What does all this imply concerning a theory of human nature? This is, of course, a very broad subject requiring more space that available here. However, we have found that even a short discussion has clarificatory value. We present some of our own tentative understandings in Section 15. Section 16 draws some conclusions for the interpretation of the preceding material, and offers directions for future research.

2 Rational Choice Theory

Rational choice theory presupposes that agents are attempting to maximize a preference function subject to whatever informational and material constraints they face. The term 'rational' is, of course, something of a misnomer, since the term appears to imply something about the ability of the agent to give reasons for actions, to act objectively, unmoved by capricious emotionality, and even to act self-interestedly. Yet, it has long been recognized that this connotational overly is superfluous and misleading. Nothing has brought this fact home more clearly than the great success of the model in explaining animal behavior, despite the fact that no one believes that fruit flies and spiders do much in the way of cogitating (Maynard Smith 1982, Alcock 1993). Rational choice theory is the starting point for much of economic analysis, behavioral game theory, and is increasingly gaining credence with neuroscientists (Shizgal 1999, Glimcher 2003).

By a *preference ordering* $\succeq$ on a set A, we mean a binary relation, such that $x \succeq y$ may be either true or false for various pairs $x, y \in A$. We pronounce $x \succeq y$ as "x is weakly preferred to y" (Kreps 1990). We say $\succeq$ is *complete* if, for any $x, y \in A$, either $x \succeq y$ or $y \succeq x$. We say $\succeq$ is *transitive* if, for all $x, y, x \in A$, $x \succeq y$ and $y \succeq z$ imply $x \succeq z$. When these two conditions are satisfied, we say $\succeq$ is a *preference relation*. We say an agent *maximizes* $\succeq$ if, if from any subset $B \subset A$, the agent chooses one of the most preferred elements of B according to $\succeq$,

Theorem 1. *If $\succeq$ is a preference relation on set A, and if an agent maximizes $\succeq$, then there always exists a utility function $u : A \rightarrow \mathbf{R}$ ($\mathbf{R}$ are the real numbers) such that the agent behaves as if maximizing this utility function.*

3 Behavioral Game Theory

Behavioral game theory provides the tools for modeling actors, the rules of the game, the informational structure, and the payoffs associated with strategic choices. As such, behavioral game theory fosters a unified analytical framework available to *all* the behavioral disciplines. Moreover, since behavioral game-theoretic predictions can be systematically tested, the results can be replicated by different laboratories (Plott 1979, Smith 1982, Sally 1995).

A *game* consists of a set of rules specifying the order of play, the actions the players have available when it is their turn to choose, the conditions under which the game terminates, and the payoffs to the players as a function of their various choices during the course of the game. Behavioral game theory presumes rational choice theory, and attempts to reveal the underlying preference functions of agents by judiciously varying the rules of the game and the payoffs, noting the subsequent change in the behavior of the players.

Behavioral game theory thus starts by treating agents' objectives as a matter of fact, not logic. We can just as well build models of regret, altruism, vindictiveness, status-seeking, shame, guilt, and addiction as of choosing a bundle of consumption goods subject to a budget constraint (Gintis 1972a,b,1974,1975, , Bowles and Gintis 1993, Becker and Murphy 1988, , Becker 1996, Becker and Mulligan 1997, Fehr and Schmidt99)..

One salient behavior in social dilemmas revealed by behavioral game theory is *strong reciprocity*. The strong reciprocator comes to a social dilemma with a propensity to cooperate (*altruistic cooperation*), responds to cooperative behavior by maintaining or increasing his level of cooperation, and responds to noncooperative behavior by punishing the "offenders," even at a cost to himself, and even when he could not reasonably expect future personal gains to flow from such punishment (*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 agent 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 agent 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, 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 agent 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. Indeed, the inequality-averse agent is willing to reduce his own payoff if that reduces the payoff of relatively favored individuals even more. In short, an inequality-averse agent 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 agent cares only about the distribution of final payoffs, and not at all the role of other players in bringing about this distribution. The strong reciprocator, by contrast, does not begrudge others their payoffs, but it very sensitive to the degree of fairness with which he is treated by others.

Vernon Smith, who was awarded the Nobel prize in 2002, began running laboratory experiments of market exchange in 1956 at Purdue and Stanford Universities. Aside from Smith, whose results strongly supported traditional economic theory, until 1980 or so, 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, their experimental design was weak. For instance, 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 (see Section 7), 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 experiments, some of which are analyzed below, did directly challenge the widely used assumption that agents are self-regarding. Not surprisingly, the first reaction within the disciplines was to criticize the experiments rather than to change the theoretical foundations of the disciplines. Of course, this is a wholly normal and valuable reaction to new data. While behavioral game theory can easily withstand the critique, it is useful to outline, and reply to, the various objections to its findings.

The most general argument is that the behavior of subjects in simple games

under controlled and limited circumstances says nothing about their behavior in the extremely complex, rich, and temporally extended social relationships into which people enter in daily life. However, controlled experiments have served the natural sciences extremely well, and are very important in modeling animal behavior, as well as medical research and pharmacology. Of course, it should be ascertained that the behaviors exhibited in pure form in the laboratory are operative as well in daily life. This appears to be the case, as shown in studies by Andreoni, Erard and Feinstein (1998) on tax compliance. Bewley (2000) on fairness in wage setting, and Fong, Bowles and Gintis (2005) on support for income redistribution, among others.

A second argument is that games in the laboratory are bizarre and unusual, so people really 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). Indeed, post-game interviews indicate that subjects clearly comprehend the one-shot aspect of the games. Moreover, as we show in Section 9, subjects are often quite willing to punish others who do not harm them, but harm a third party by violating a social norm. It is not plausible to attribute this behavior to confusion with repeated games.

It is also simply not the case that we rarely face one-shot, anonymous interactions 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. Moreover, in these societies, greater exposure to market exchange led to stronger, not weaker, deviations from self-regarding behavior (Henrich, Boyd, Bowles, Camerer, Fehr and Gintis 2004).

Another indication that the other-regarding behavior observed in the laboratory

is not simply error 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.

Another objection often expressed is that subjects really do not believe that the 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). 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 one defined "conscience" as "the inner 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, one will nonetheless *remember* how one played a game, and one may derive great pleasure from recalling one's generosity, or one's willingness to incur the costs of punishing another player for being selfish. This is quite correct, and probably explains a good deal of non-self-regarding behavior in experimental games.[5] But, this does contradict the fact that our behavior is self-regarding! Indeed, it confirms it, although there may be some philosophi-

[5]William Shakespeare understood 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."

cal arguments (irrelevant from the behavioral standpoint) that the other-regarding behavior is nonetheless self-interested.

In all the games described below, unless otherwise stated, subjects are anonymous to one another, they are college students who are recruited by bulletin board and newspaper announcement. The main exception is the extensive study of hunter-gatherer and other simple, non-market societies reported in Section 11. In all cases, subjects are paid real money, they are not deceived or misled by the experimenters, and they are instructed in the to the point where they fully understand the rules and the payoffs before playing for real.

4 An Anonymous Market Exchange

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. Indeed, 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).

In the Holt, Langan, and Villamil experiment there are four "buyers" and four "sellers." The good is a chip that the seller can redeem for \$5.70 but the 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 four chips, while eleven chips were distributed to sellers (three sellers were given three chips each, and the fourth was given two chips). Clearly, buyers are willing to pay up to \$6.80 per chip for up to four chips each, and buyers are willing to sell their chip for any amount at or above \$5.70. Total demand is thus sixteen for all prices at or below \$6.80, and total supply is eleven chips at or above \$5.70. Since there is an excess demand for chips at every price between \$5.70 and \$6.80, the only point of intersection of 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, since each knew only his own supply of or demand for chips.

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 either it is accepted by another player, or a lower asking price

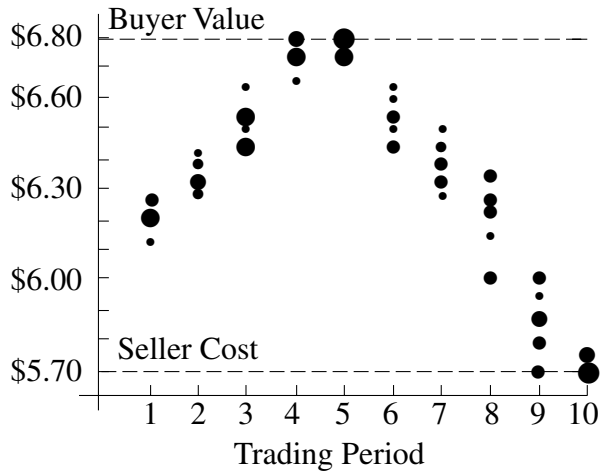

Figure 1: Simulating a Market Equilibrium: The Double Auction. The size of the circle is proportional to the number of trades that occurred at the stated price.

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 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 six and each of the succeeding four periods, buyers were given the right to redeem a total of eleven chips, and each seller was given four chips. In this new situation, it is clear (to us) that there is an excess supply of chips at each price between $5.70 and $6.80, so the only place supply and demand intersect is 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 supply, succeeding periods witnessed a steady fall in price, until in the tenth period, the price is close to the neoclassical prediction, and now the buyers are earning about $1.70 per chip. A more remarkable vindication of the neoclassical model would be difficult to imagine.

5 The Rationality of Altruistic Giving

We have stressed that there is nothing *prima facie* irrational about caring for others, so that one prefers to have less in order that others have more. But, do preferences for altruistic acts entail transitive preferences, as required by the notion of ratio-

nality in decision theory? Andreoni and Miller (2002) showed that they are.

In the dictator game, first studied by Forsythe, Horowitz, Savin and Sefton (1994), the experimenter gives a subject, called the "dictator" a certain amount of money, and instructs the subject to give any portion of this he desires to a second, anonymous, subject, called the "recipient." The dictator keeps whatever he does not choose to give to the recipient. Obviously, a self-regarding dictator will give nothing to the recipient. 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 this to the recipient is p, meaning that each point the recipient gets costs the giver p points. For instance, if $p = 4$, then it costs the dictator four points for each point that he transfers to the recipient. The dictator's choices must then satisfy the "budget constraint" $\pi_s + p\pi_o = m$, where π_s is the amount the dictator keeps and π_o is the amount the recipient 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 recipient as to care about consuming marketed commodities.

Varian (1982) showed that the following Generalized Axiom of Revealed Preference (GARP) is sufficient to ensure rationality. We say A is *directly revealed preferred* to B if B was in the choice set when A was chosen. We say A is *indirectly revealed preferred* to B if there is some sequence $A = C_1, C_2, \ldots, C_k = B$, which each C_i is directly revealed preferred to C_{i+1} for $i = 1, \ldots, k-1$. GARP then is the following condition: *If A is indirectly revealed preferred to B, the B is not strictly revealed directly preferred to A*.

The experimenters worked with 176 students in a elementary economics class, and had them play the dictator game eight 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 equalling $m = 40, 60, 75, 80$ and 100. They found that on 18 of the 176 subjects violated GARP at least once, and of these, only four were at all significant violations. By contrast, if choices were randomly generated, we would expect between 78% and 95% of subjects would violate 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, while 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 recipient).

We conclude from this study that *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 this paper.

6 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 agent 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 use the prisoner's dilemma to show that a majority of subjects (at least college students in Japan and the United States) positively value altruistic cooperation.[6] 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." It is easy to see that a self-regarding agents will exhibit $DC \succ CC \succ DD \succ CD$ (check it), while an altruistic cooperator will exhibit $CC \succ DC \succ DD \succ CD$ (where $\succ$ means "is preferred to"); i.e. the self-regarding agent prefers to defect no matter what his partner does, whereas the conditional altruistic cooperator prefers to cooperate so long as his partner cooperates. Watabe, Terai, Hayashi and Yamagishi (1996), based on 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, Ostrom, Walker and Yamagishi (1999) ran the same experiment with U.S. students with similar results. In this case, all of the subjects asked if they would cooperate if their partner was already committed to cooperating said they would.

While many agents 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 indeed if they are held at all, since 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 pay-

[6] In the prisoner's dilemma, if the two players cooperate, they both get payoff b, and if they both defect, the both get payoff $p < r$. If one cooperates and on defects, the cooperate gets $s < p$ and the defector gets $t > r$. A self-interested player will defect no matter what his partner does.

offs using 149 Japanese university students. The experimenters ran three distinct treatments, with about equal numbers of the subjects in each treatment. The first treatment was the 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 defect would be made known to the second player before the latter made his own choice. The experimenters found that 38% of 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 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.

7 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 agent wants to create a more equal distribution of outcomes, even at the cost of lower outcomes for all. The simplest game exhibiting altruistic punishment is the *ultimatum game* (Güth et al. 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. 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 agent: accept anything you are offered. Knowing this, a self-regarding proposer who believes he faces a self-regarding responder, will offer the minimum possible amount, $1, and this will be 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 respondents very substantial amounts (50% of the total generally being the modal offer), and respondents 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. (1991) conducted ultimatum games 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 respondents 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 difference in behavior are found, reflecting different standards of what it means to be "fair" in different types of society.

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 under 40% as a form of altruistic punishment of the norm-violating proposer. Proposer offer 50% because they are altruistic cooperators, or 40% because they fear rejection. To support this interpretation, we note that if the offer in an ultimatum game are generated by a computer rather than the proposer, and if respondents know this, low offers 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 getting 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 (Bolton and Zwick 1995). As a final indication that strong reciprocity motives are operative in this game, when, 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. In all of the above experiments a significant fraction of subjects (about a quarter, typically) conform to self-regarding preferences.

A more difficult issue is the extent to which behavior reflects strong reciprocity as opposed to inequality aversion. The difference between the two motivations is straightforward: strong reciprocity is sensitive to the intentions of other players,

whereas inequality aversion is not. It follows that in any game in which a subject acts to reduce inequality in outcomes, yet the inequality is not due to the elective behavior (actual or expected) of other players, inequality aversion must be at work. The problem is that behavioral games, unless explicitly structured to deal with the role of intentions versus outcomes in altruistic behavior, ineluctably confound the two issues. Special versions of the trust game and the moonlight game confront this problem.

8 Strong Reciprocity in the Labor Market

Akerlof (1982) suggested that many puzzling facts about labor markets could be better understood if it were recognized that in many situations, 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) performed an experiment to validate this *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 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*.

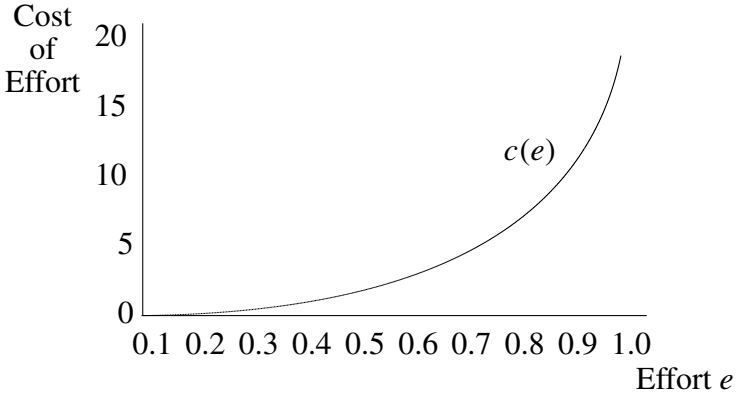

Figure 2: The Cost of Effort Schedule in Fehr, Gächter, and Kirchsteiger (1997).

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, which *need not equal the contracted effort*, e^*. In effect, there is no penalty if the employee does not keep his promise, so the employee can choose any effort level, $e \in [0.1, 1]$, with impunity. Although subjects may play this game several times with different partners, each employer-employee interaction is a one-shot (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 (assuming only integral wage offers are permitted).[7] 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 $\pi = 0.1 \times 100 - 1 = 9$.

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 receive higher effort, as a means to increase both their own and the employee's payoff, as depicted in Figure 3. Similar results have been observed in Fehr, Kirchsteiger and Riedl 1993, 1998.

Figure 3 also shows that, though most employees are strong reciprocators, at any wage rate there still is a significant gap between the amount of 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 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, since they would not (knowingly) interact with the same worker a second time. However, 68% of the time, employers punished employees that did not fulfill

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

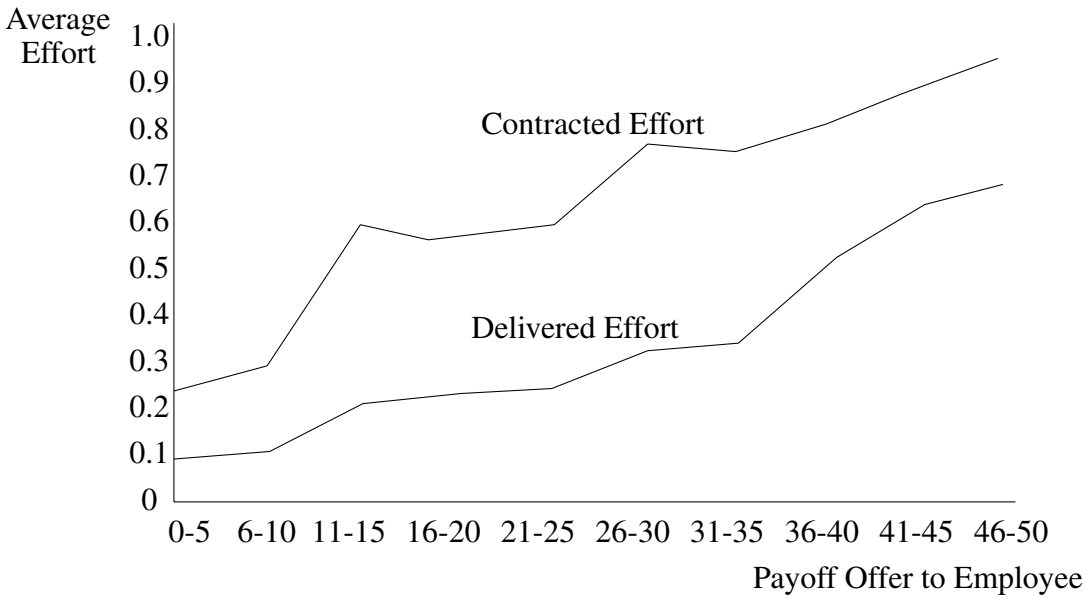

Figure 3: Relation of Contracted and Delivered Effort to Worker Payoff (141 subjects). From Fehr, Gächter, and Kirchsteiger (1997).

their contracts, and 70% of the time, employers rewarded employees who overfulfilled their contracts. Indeed, 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 83% 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 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 reward good 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 only 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 so *for its*

own sake.

9 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* the 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, will help the individual and his family in times of need, will preferentially trade favors with the individual, and otherwise will 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 will refuse aid even when it is not costly to do so, will shun the offender and approve of the offender's ostracism from beneficial group activities, again at low cost to the third party, but highly costly to the offender.

It is hard to conceive of human societies operating at a high level of cooperative efficiency in the absence of such third party altruistic behavioral. Yet, self-regarding actors will never engage in such behavior if it is at all costly. An experiment conducted by Fehr and Fischbacher (2004) addresses this question by conducting a series of Third Party Punishment games using prisoner's dilemma and dictator games. The experimenters implemented four experimental treatments, in each of which subjects were groups into threes. In each group, in stage one, subject A played a prisoner's dilemma or dictator game with subject B as the recipient, 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 three points deducted from A's score cost C one 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 ten points, and each could either keep the ten points, or transfer the ten points to the other subject, in which case it was tripled by the experimenter. Thus, if both cooperated, each earned 30 points, and if both defected, each earned ten points. If one cooperated

and one defected, however, the cooperator earned zero and the defector 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 As 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 only punish his *own* dictator, whereas in the T condition, a B player could only punish 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 player A' 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 in the S condition, an AB pair could only punish each other, whereas in the T condition, each agent could punish only a randomly assigned subject from another pair.[8]

In the first two treatments, since 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 (C's) punish dictators (A's) who give less than 50% of the endowment to recipients (B's). 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 AB pair defected and the other cooperated, the defector would have on average 10.05 points deducted by C's, but if both defected, the punished player lost only an average of 1.75 points. This shows that

[8]It is worth repeating that the experimenters never use value-laden terms such as "punish," but rather neutral terms, such as "deduct points."

third parties (C's) care not only about the intentions of defectors, but how much harm they caused and/or how unfair they turned out to be. Overall, 45.8% of third parties punished defectors whose partners cooperated, whereas only 20.8% of third parties punished defectors whose partners defected.

Turning to the third treatment (T&SP-DG), second party sanctions of selfish dictators are found to be considerably stronger than third party sanctions, although both were highly significant. On average, in the first condition, where recipients 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 points per point kept on third party dictators. In the final treatment (T&SP-PD), defectors are severely punished by both second and third parties, but second party punishment is again found to be much more severe than third. 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 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.

10 Altruism in Social Dilemmas

The *public goods game* reflects several real-life social dilemmas, such as the voluntary contribution to team and community goals. Researchers (Ledyard 1995, Yamagishi 1986, Ostrom, Walker and Gardner 1992, Gächter and Fehr 1999) uniformly find 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 ten. 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 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 '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 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 the whole endowment into the common account, a player loses twelve points but the

other three group members gain in total 24 ($= 8 \times 3$) points. The players keep whatever is in their private account at the end of the round.

A self-regarding player will contribute nothing to the common account. However, only a fraction of subjects in fact conform to the self-regarding model. Subjects begin by contributing on average about half of their endowment 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. 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 et al. (1992) subjects in a public goods game, by paying a "fee," could impose costs on others by "fining" them. Since 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 rational choice theory, 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, since costly punishment of defectors could increase cooperation in future periods, yielding a positive net return for the punisher. It is true that backward induction rules out such a strategy, but we know that people do not backward induct very far anyway (Camerer 2003).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 six and ten round public goods

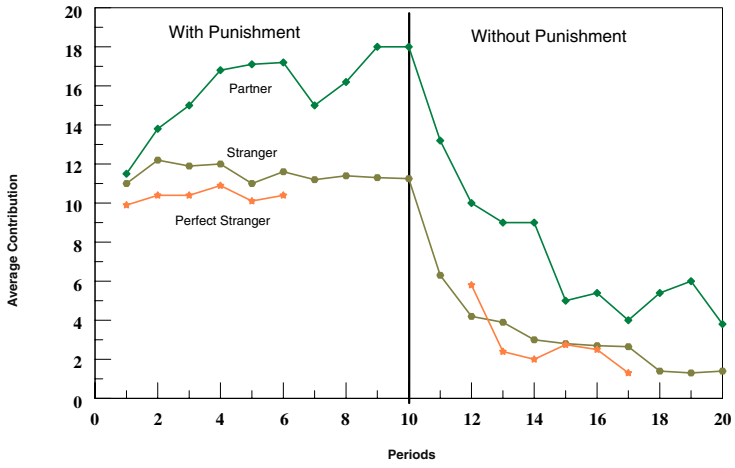

Figure 4: Average Contributions over Time in the Partner, Stranger, and Perfect Stranger Treatments when the Punishment Condition is Played First (Fehr and Gächter, 2000).

games with groups of size four, 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 ten periods. Under the *Stranger* treatment, the subjects were randomly reassigned after each round. Finally, under the *Perfect Stranger* treatment the subjects were randomly reassigned, but assured that they would never meet the same subject more than once.

Fehr and Gächter (2000) performed their experiment for ten rounds with punishment and ten rounds without. Their results are illustrated in Figure 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 on 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 in this treatment the punished subjects are certain that, once they have been punished in previous rounds, 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.[9]

11 Experimental Games in the Field

To expand the diversity of cultural and economic circumstances of experimental subjects, Herbert Gintis and Robert Boyd, with funds provided by the John D. and Catherine T. MacArthur Foundation, undertook a large cross-cultural study of behavior in various games including the ultimatum game and the public goods game (Henrich, Boyd, Bowles, Camerer, Fehr, Gintis and McElreath 2001, Henrich et al. 2004). These societies, exhibiting a wide variety of economic and cultural conditions, 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 and 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 Zimbabwe farmers in Africa). As in games described previously, these games were played anonymously, without misleading or deceiving subjects, and for real stakes—the local equivalent of one or more day's wages. The results can be summarized in four points.

- *There was no society in which experimental behavior was consistent with the self-regarding preferences*. In the ultimatum game, in some societies there were virtually no rejections, even of very low offers, but proposers still made high offers to respondents. In others, responders were quick to reject low offers, and proposer correspondingly made relatively high offers. In still other societies, neither group of players conformed to the self-regarding actor model.

[9] 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 the start of the game. It was not.

- *There was markedly more variation among groups than had been previously reported*. While mean ultimatum game offers in experiments with student subjects are typically between 43% and 48%, the mean offers from proposers in our sample range from 26% to 58%. While modal ultimatum game offers are consistently 50% among university students, sample modes with this data range from 15% to 50%. In some groups rejections were extremely rare, 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 50%).

  Typical distributions of contributions in a one-round public goods game in previous studies have a U-shape with the mode at full defection and a secondary mode at full cooperation, with mean contribution between 40% and 60%. By contrast, the Machiguenga have a mode at full defection exhibit zero full cooperation, and have mean 22%. The Aché and Tsimané distributions resemble American distributions, but with very low rejection rates. The Orma and Huinca (non-Mapuche Chileans living among the Mapuche) have modal offers near the center of the distribution, but show secondary peaks at full cooperation.

- *Differences among societies in* "market integration" *and* "cooperation in production" *explain a substantial portion of the behavioral variation between groups.* 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, market integration and cooperation in production, were significant, and they together accounted for 66% of the variation among societies in mean ultimatum game offers.

- *Behavior in experimental games tended to mirror patterns of interaction found in everyday life in that society.*

In a number of cases the parallels between experimental game play and the structure of daily life are quite striking. Nor was this relationship lost on the subjects themselves. Here are some examples.

1. 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 58% with 25% full contributors).

2. Among the Au and Gnau, many proposers offered more than half the pie, 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.

3. Among the whale hunting Lamalera, 63% of the proposers in the ultimatum game divided the pie equally, and most of those who did not, offered more than 50% (the mean offer was 57%). In real life, a large catch, always the product of cooperation among many individual whalers, is meticulously divided into pre-designated parts and carefully distributed among the members of the community.

4. Among the Aché, 79% of proposers offered either 40% or 50%, and 16% offered more than 50%, with no rejected offers. In daily life, the Aché regularly share meat, which is being distributed equally among all other households, irrespective of which hunter made the catch.

5. The Hadza made low offers and had high rejection rates in the ultimatum game, the opposite of the Aché. This reflect 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.

6. 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 show little fear of social sanctions and care little about "public opinion."

7. The Mapuche's social relations are characterized by mutual suspicion, envy, and fear of being envied. This pattern is consistent with the Mapuche's post-game interviews in 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 rare spiteful responders, who would be willing to reject even 50/50 offers.

12 Intentions Matter I: 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, player A, is told he can transfer any number of dollars, from zero to ten, to his (anonymous) partner, player B, and keep the remainder. The amount transferred will be tripled by the experimenter and given to player B, who can then give any number of dollars back to player A (this amount is not tripled). A player A who transfers a lot is called *trusting*, and a player B who returns a lot to player A is called *trustworthy*.

Clearly, if all agents have self-regarding preferences, and if each player A believes his partner has self-regarding preferences, then player A will give nothing to player B. On the other hand, if player A believes player B is inequality averse, he will transfer all $10 to player B, who will then have $40. To avoid inequality, player B will give $20 back to player A. A similar result would obtain if player A believes player B is a strong reciprocator. On the other hand, if player A is altruistic, he may transfer something to player B, on the grounds that the money is worth more to player B (since it is tripled) than it is to himself, even if player A does not expect anything back. It follows that several distinct motivations can lead to a positive transfer of money from A to B and then back to A.

Berg et al. (1995) found that on average, $5.16 was transferred from A to B, and on average, $4.66 was transferred back from B's to A's. Furthermore, when the experimenters revealed this result to the subjects and had them play the game a second time, $5.36 was transferred from A's to B's, and $6.46 was transferred back from B's to A's. In both sets of games there was a great deal of variability, some player A's transferring everything, some nothing, and some player B's more than fully repaying their partner, and some giving back nothing.

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 exactly like treatment A, except that now player B cannot return anything to player A. Treatment C differs from treatment A in that each player A is matched one-to-one with a player A in treatment A, and each player B is matched one-to-one with a player B in treatment A. Each player in treatment C is 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 player A group and the player B group have exactly what they had under treatment A, except that player A now has nothing to do with player B's endowment, so nothing transferred from B to A can 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, the subjects in treatment C were not told the reasoning behind the size 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 where anonymous to one another and to the experimenter).

In treatment B, the player A dictator game counterpart to the trust game, player A 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. Since 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.

13 Intentions Matter II: The Moonlighting Game

The moonlighting game provides another way to isolate reciprocity motives, which involve the intentionality of players, from pure fairness motives, which depend only on the distribution of payoffs. The moonlighting game is a version of the trust game in which both positive and negative reciprocal behavior can be observed. First, player A chooses an action $a \in \{-6, -5, \ldots, 5, 6\}$. If $a \geq 0$, the experimenter triples a and gives it to player B. If $a < 0$, player A takes $|a|$ points from player B. Player B then chooses action $b \in \{-6, -5, \ldots, 17, 18\}$. If $b \geq 0$, one point is transferred from B to A, and if $b < 0$, B loses $|b|$ and A loses $3|b|$. Falk, Fehr and Fischbacher (2002) implemented the moonlighting game with 206 subjects who were students at the University of Zürich. They set up two treatments. Treatment A was the game as described above, and treatment B was the same game, except that a randomizing device, rather than player A, was used to determine a.

The probabilities according to which a was chosen in treatment B were set to be equal to the frequencies with which various values of a were chosen in an

implementation of the game by Abbink, Irlenbusch and Renner (2000). Moreover, B-players in both treatments were shown this frequency distribution (which varied from 2% for $a = -5$,-2, or +5, to 24% for $a = 6$ and 13% for $a = 0$) and asked to indicate how much they would return for each value of a before they actually saw the choice of player A in treatment A, or the random choice in treatment B (this is called the *strategy method* of eliciting behavioral responses).[10]

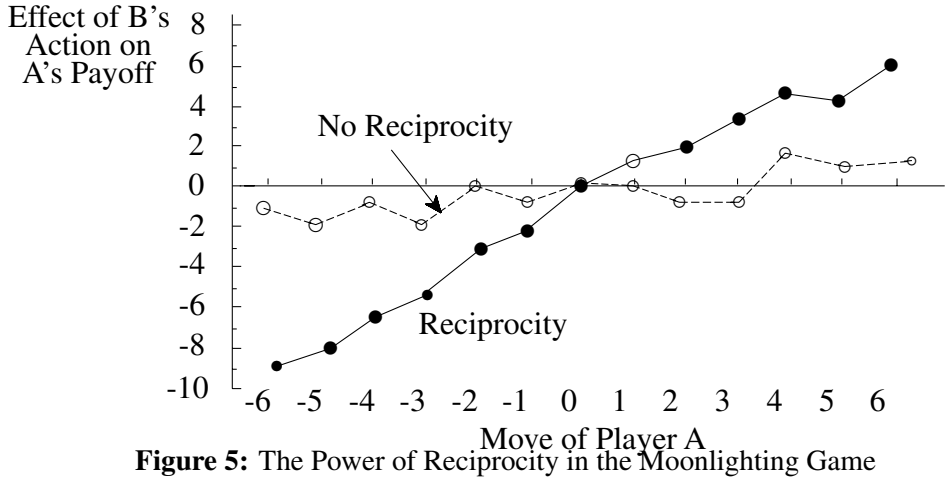

Figure 5: The Power of Reciprocity in the Moonlighting Game

In either treatment, a self-regarding player B will neither reward nor punish his partner. As can be seen in figure 5, this was nearly the case for treatment B, in which player A intentions could not be determined. But in treatment A, where player A's intentions were known to player B, and hence reciprocity was possible, there was a strong relationship between player A's generosity and the extent to which player B punished or rewarded player A. Indeed, in treatment A, 30% of player B's behaved perfectly selfishly, while in treatment B, none did.

It is worth noting that in this experiment, as in virtually every behavioral experiment, there is a great deal of subject heterogeneity, masked by the presentation of average results. For instance, in treatment A in this experiment, about 35% of subjects full strong reciprocity, while another 21% exhibited only positively reciprocal responses, and 15% exhibited only negatively reciprocal responses.

[10]It is possible that B-players would behave differently using the strategy method than if they chose their response facing an actual behavior of the A-player, but Cason and Mui (1998), among others, have shown that there is no observable effect.

14 The Economy as Social System

"An economic transaction," says Abba Lerner (1972), "is a solved political problem. Economics has gained the title of queen of the social sciences by choosing solved political problems as its domain." Lerner's observation is correct, however, only insofar as economic transactions are indeed *solved* political problems. The assumption in neoclassical economic theory that gives this result is that *all economic transactions involved contractual agreements that are enforced by third parties (e.g., the judiciary) at no cost to the exchanging parties.* However, some of the most important economic transactions are characterized by the *absence of third-party enforcement.*

Consider, for instance, the relationship between a employer and an employee, analyzed in Section 10. The employer promises to pay the worker, and the worker agrees to work hard on behalf of the firm. The worker's promise, however, is typically not suitably specific to be enforceable in a court of law. Rather than suing an employee for not working sufficiently hard, the employer generally simply dismisses the worker. For this threat to be effective, the employer must pay a wage sufficiently high that the worker can expect incur very high unemployment and search costs to secure an equally good alternative position. Hence, the exchange between employer and employee is not a "solved political problem," and both the gift exchange issue analyzed in Section 8 and the disciplining of labor by virtue of the authority relationship between employer and employee are involved in the determination of wages, labor productivity, and indeed the overall organization of the production process.

An experiment conducted by Brown, Falk and Fehr (2004) shows clearly that if third party enforcement is ruled out, employers prefer to establish long-term relationships with employees, offering a high wages, and using the threat of ending the relationship to induce high effort. Rather than market clearing determining the wage, as in the neoclassical labor market, the result in this experiment is a labor market dominated by long-term relationships, with a positive level of unemployment in equilibrium, and employed workers enjoying a payoff advantage over unemployed workers. Labor market competition has little effect on the wage rate in this case, because employers will not rupture long-term relationships by hiring the unemployed at a lower wage.

Brown, Falk, and Fehr (BFF) used 15 trading periods with 238 subjects and three treatments. The first treatment was the standard complete contract condition (C condition) in which labor effort is contractually specified and guaranteed. The second treatment was an incomplete contract condition (ICF condition) with

exactly the same characteristics, including costs and payoffs to employer and employee, as in Section 8. In addition, however, workers were given a payment of 5 points in each period that they were unemployed. In both conditions, subjects had identification numbers that allow long-term relationships to develop. The third treatment, which we call ICR, was identical to ICF, except that long-term relationships were ruled out (subjects received shuffled identification numbers in each experimental period). This treatment is thus identical to the gift exchange model in Section 8, except for the 5 point "unemployment compensation."

All contracts formally lasted only one period, so even long-term relationships had to be explicitly renewed in each period. If agents are self-regarding, it is easy to see that in the ICR treatment, all employees will supply the lowest possible effort $e = 1$, and employers will offer wage $w = 5$. Each firm then has a profit of $10e - 5 = 5$, and each worker has payoff $w - c(e) = 5 - c(0) = 5$. This outcome will also occur in the last period of the ICF treatment, and hence by backward induction, will hold in all periods. In the C treatment with self-regarding agents, it is easy to show that the employer will set $w = 23$ and require $e = 10$, so workers get $w - c(e) = 23 - c(10) = 5$ and employers get $10e - w = 100 - 23 = 77$ in each period. Workers are, in effect indifferent between being employed an unemployed in all cases.

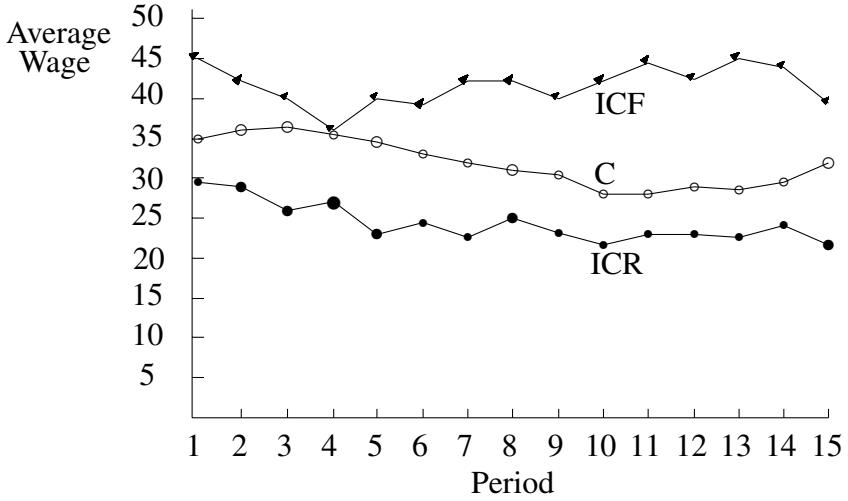

Figure 6: Wages over Fifteen Periods (Brown et al. 2004). The C treatment is complete contracting, the ICF treatment is incomplete contracting with long-term relationships permitted, and the ICR treatment is incomplete contracting with no long-term relationships permitted.

The actual results were, not surprisingly, quite at variance with the self-regarding preferences assumption. Figure 6 shows the path of wages over the fifteen periods under the three treatments. The ICR condition reproduces the result of Section 8, wages being consistently well above the self-regarding level of $w = 5$. If the C condition were a two-sided double auction, we would expect wages to converge to $w = 23$, as in Section 4. In fact, the ICR conditions gives wages closer to the prediction for complete contracting than the C condition. The ICF condition gives the highest wages after the fourth period, validating the claim that under conditions of incomplete contracting, long-term relationships will prevail, and the distribution of gains will be more equal between buyers and sellers.

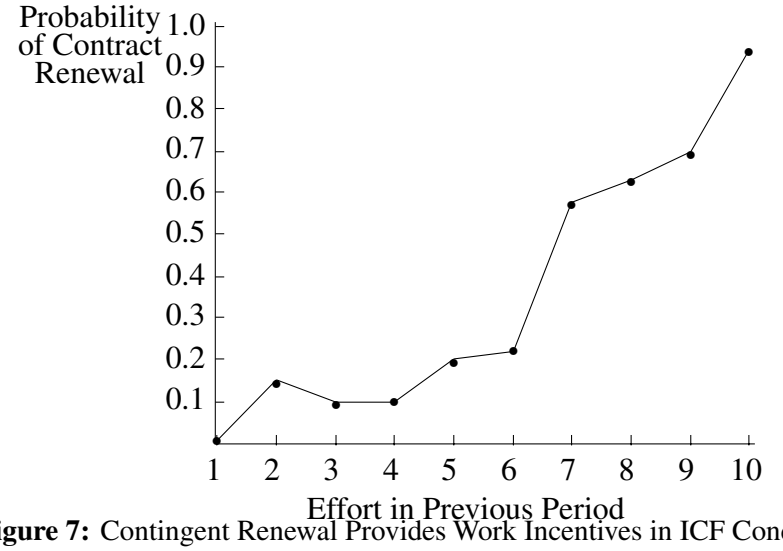

Figure 7: Contingent Renewal Provides Work Incentives in ICF Condition

By paying high wages in the ICF condition, employers were capable of effectively threatening their employees with dismissal (non-renewal of contract) if they were dissatisfied with worker performance. Figure 7 shows that this threat was in fact often exercised. Workers with effort close to $e = 10$ were non-renewed only about 5% of the time, whereas workers with effort below $e = 7$ were rarely renewed.

Figure 8 shows that the effect of different contracting availabilities strongly affects the level of productivity of the system, as measured by average effort levels. Under complete contracting, effort levels quickly attain near-efficiency ($e = 10$), and remain there. Contingent renewal of long-term relationships achieves between 80% and 90% efficiency, with a significant end-game effect, as the threat of non-renewal is not very effective on the last few rounds. The gift exchange treatment

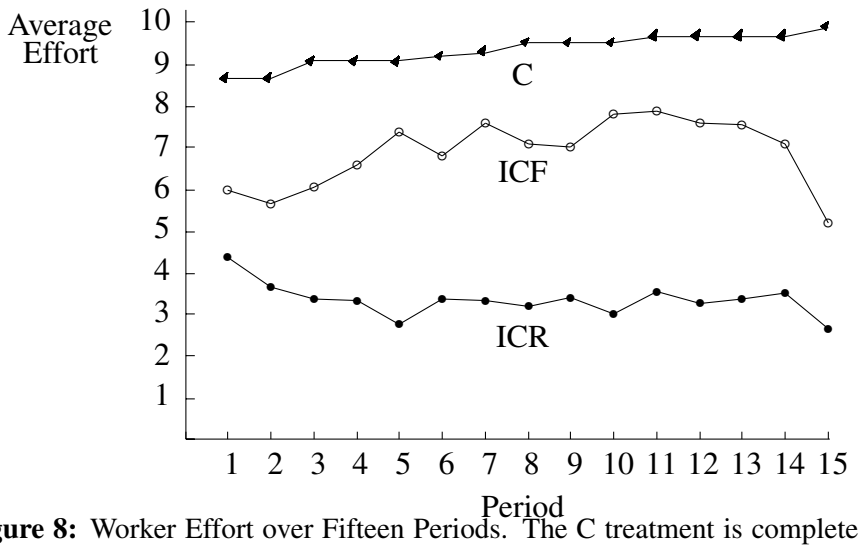

Figure 8: Worker Effort over Fifteen Periods. The C treatment is complete contracting, the ICF treatment is incomplete contracting with long-term relationships permitted, and the ICR treatment is incomplete contracting with no long-term relationships permitted.

(ICR), while supporting effort levels considerably above the self-regarding level, is considerably less efficient that either of the others, although it predictably suffers a smaller end-game effect that the ICF condition.

One extremely interesting pattern emerging from this study is the interaction of gift exchange and threat in the employer-employee relationship. One might thing that they would be mutually exclusive, on the grounds that one cannot both feel charitable towards one's employer while at the same time being threatened by him. Yet, many of us will recall from personal experience this ambiguous co-presence of good will and fear. In this study, the importance of gift exchange in the long-term relationship is exhibited by the fact that even in the last two periods, where the threat of dismissal is weak or absent, effort levels are considerably above those of the pure gift exchange condition. Thus, gift exchange appears to be stronger when accompanied by the capacity of the employer to harm, as though the fact that the employer has not exercised this capacity increases the worker's gratitude and willingness to supply effort.

15 Altruism and Human Nature

One naturally asks whether behaviors such as strong reciprocity and inequality aversion are products of culture or part of our genetic constitution as a species. Before answering this question, it is important to be address two points. First, the behavior of any biological organism is the product of an interaction between genes and environment. Genes predispose an organism to be affected differential by different environments and not others (Schlichting and Pigliucci 1998), and genes predispose an organism to seek out certain environments and to transform environments in particular ways to meet their needs (Odling-Smee, Laland and Feldman 2003). It follows that for humans, it is *almost never reasonable to ask whether a certain behavior is genetic or cultural, or even which is "more important."*

Second, many human behaviors appear to be *structurally universal*, but *functionally culturally specific*. For instance, all humans use highly complex linguistic structures with a good deal of underlying commonality, but languages are extremely diverse and mutually unintelligible (Chomsky 1957). Also, the behavioral patterns surrounding *shame* are universal (downcast eyes, blushing, shallow breathing, etc.), and shame is always the results of being discovered in violation of social norms. However, the *content* of the social norms that trigger the shame response are highly culturally specific. Indeed, acting a certain way may trigger shame in one society, and not acting that way may trigger shame in a different society.

Given these background facts, we can say that human being have a universal, probably genetic, predisposition to display such behaviors as strong reciprocity, but the situations that trigger altruistic cooperation and/or punishment differ widely in different societies. Such behaviors are thus, like the language facility or the shame syndrome, structurally universal with culturally specific expressions. This is, of course, most obviously the case for the internalization of norms, which is a potent capacity, the content of which is highly culturally variable. We should also add that there is a great deal of individual heterogeneity in the expression of prosocial traits, from the exclusively positive altruism associated with some especially saintly humans, to the exclusively negative altruism associated with certain puritanical personalities, and embracing the purely self-regarding behavior of the sociopath.

16 Modeling Human Behavior: Conclusion and Review

Experimental games are artificial and quite removed from everyday life. How, then do they relate to everyday life? The laboratory allows us to control the social environment so that experiments can be replicated and the results from different experiments can be compared. In physics and chemistry, the experimental method has the additional goal of *eliminating all influences on the behavior of the object of study except those controlled by the experimenter*. This goal can be achieved because elementary particles, and even chemical compounds, are completely interchangeable, given a few easily measurable characteristics (atomic number, energy, spin, chemical composition, and the like). Experiments in human social interaction, however, *cannot* achieve this goal, even in principle, because experimental subjects bring their personal history with them into the laboratory. Their behavior is therefore *ineluctably* an interaction between the subject's personal history and the experimenter's controlled laboratory conditions.

This observation is intimately related to the basic structure of evolutionary game theory (and human psychology, as stressed by Loewenstein 1999). In strategic interaction nature abhors low probability events, and for an experimental subject, *the experiment is precisely a low probability event!* Neither personal history nor general cultural/genetic evolutionary history has prepared subjects for the Ultimatum, Dictator, Common Pool Resource, and other games that they are asked to confront. An agent treats a low probability event as a high probability event by assigning a novel situation to one of a small number of pre-given *situational contexts*, and then deploying the behavioral repertoire—payoffs, probabilities, and actions—appropriate to that context. We may call this *choosing a frame* for interpreting the experimental situation. This is how subjects bring their history to an experiment.[11]

The results of the ultimatum game in Section 7, for instance, suggest that in a two-person bargaining situation, in the absence of other cues, the situational context applied by most subjects dictates some form of "sharing." Suppose we change the rules such that both proposer and respondent are members of different

[11]For a similar view, see Hoffman, McCabe and Smith (1996). A caveat: It is incorrect to think that the subjects are "irrational" or "confused" because they drag their history into an experimental situation. In fact, they are acting normally on the basis of the preferences they exhibit in daily life. Of course, if this low probability event (being a subject in an experiment) turns into a high probability event (e.g., by being repeatedly asked to be a subject), agents may change their framing or even create a wholly new situational context for the purpose at hand. The process is not well understood.

teams and each is told that their respective winning will be paid to the team rather than the individual. A distinct situational context, involving "winning," is now often deemed appropriate, dictating acting on behalf of one's team and suppressing behaviors that would be otherwise individually satisfying—such as "sharing." In this case, proposers offer much less, and respondents very rarely reject positive offers (Shogren 1989). Similarly, if the experimenters introduce notions of property rights into the strategic situation (e.g., that the proposer in an ultimatum game has "earned" or "won" the right to this position), then motivations concerning "fairness" are considerably attenuated in the experimental results (Hoffman, McCabe, Shachat and Smith 1994, Hoffman et al. 1996).

Unless there is some substantive relationship between the behavior of agents in the laboratory and these agents in daily life, the relevance of experimental data will be severely circumscribed. The next step in behavioral game theory is to take games to such natural social settings as schools, workplaces, churches, community centers, prisons, and hospitals. Does the structure of everyday life explain how people play experimental games? Do individual behavioral characteristics isolated in the laboratory reflect the real-life behavior of individuals? Although we have some suggestive evidence (see Section 11), by and large, we simply do not know.

REFERENCES

Abbink, Klaus, B. Irlenbusch, and E. Renner, "The Moonlighting Game—An Experimental Study on Reciprocity and Retribution," *Journal of Economic Behavior and Organization* 42 (2000):265–277.

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).

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.

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

Becker, Gary S., *Accounting for Tastes* (Cambridge, MA: Harvard University Press, 1996).

____ and Casey B. Mulligan, "The Endogenous Determination of Time Preference," *Quarterly Journal of Economics* 112,3 (August 1997):729–759.

____ and Kevin M. Murphy, "A Theory of Rational Addiction," *Journal of Political Economy* 96,4 (August 1988):675–700.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior* 10 (1995):122–142.

Bewley, Truman F., *Why Wages Don't Fall During a Recession* (Cambridge: Cambridge University Press, 2000).

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.

Bolhuis, J. J. and R. C. Honey, "Imprinting, Learning and Development: From Behaviour to Brain and Back," *Trends in Neuroscience* 21 (1998):306–311.

Bolton, Gary E. and Rami Zwick, "Anonymity versus Punishment in Ultimatum Games," *Games and Economic Behavior* 10 (1995):95–121.

____ , Elena Katok, and Rami Zwick, "Dictator Game Giving: Rules of Fairness versus Acts of Kindness," *International Journal of Game Theory* 27,2 (July 1998):269–299.

Boudon, Raymond, "Beyond Rational Choice Theory," *Annual Review of Sociology* 29 (2003):1–21.

Bowles, Samuel and Herbert Gintis, "The Revenge of Homo Economicus: Contested Exchange and the Revival of Political Economy," *Journal of Economic Perspectives* 7,1 (Winter 1993):83–102.

____ and ____ , "The Origins of Human Cooperation," in Peter Hammerstein (ed.) *Genetic and Cultural Origins of Cooperation* (Cambridge, MA: The MIT Press, 2004).

Brown, Martin, Armin Falk, and Ernst Fehr, "Relational Contracts and the Nature of Market Interactions," *Econometrica* 72,3 (May 2004):747–780.

Camerer, Colin, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton, NJ: Princeton University Press, 2003).

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

Cason, T. and V-L. Mui, "Social Influence in the Sequential Dictator Game," *Journal of Mathematical Psychology* 42 (1998):248–265.

Chomsky, Noam, *Syntactic Structures* (The Hague: Mouton & Co., 1957).

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

Cox, James C., "How to Identify Trust and Reciprocity," *Games and Economic Behavior* 46 (2004):260–281.

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

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

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

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

____ and Urs Fischbacher, "Third Party Punishment and Social Norms," *Evolution & Human Behavior* 25 (2004):63–87.

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

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

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

Fong, Christina M., Samuel Bowles, and Herbert Gintis, "Reciprocity and the Welfare State," in Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr (eds.) *Moral Sentiments and Material Interests: On the Foundations of Cooperation in Economic Life* (Cambridge: The MIT Press, 2005).

Forsythe, Robert, Joel Horowitz, N. E. Savin, and Martin Sefton, "Replicability, Fairness and Pay in Experiments with Simple Bargaining Games," *Games and Economic Behavior* 6,3 (May 1994):347–369.

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, "Consumer Behavior and the Concept of Sovereignty," *American Economic Review* 62,2 (May 1972):267–278.

____ , "A Radical Analysis of Welfare Economics and Individual Development," *Quarterly Journal of Economics* 86,4 (November 1972):572–599.

____ , "Welfare Criteria with Endogenous Preferences: The Economics of Education," *International Economic Review* 15,2 (June 1974):415–429.

____ , "Welfare Economics and Individual Development: A Reply to Talcott Parsons," *Quarterly Journal of Economics* 89,2 (February 1975):291–302.

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

Glimcher, Paul W., *Decisions, Uncertainty, and the Brain: The Science of Neuroeconomics* (Cambridge, MA: MIT Press, 2003).

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.

Grusec, Joan E. and Leon Kuczynski, *Parenting and Children's Internalization of Values: A Handbook of Contemporary Theory* (New York: John Wily & Sons, 1997).

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

Güth, Werner, R. Schmittberger, and B. Schwarze, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3 (May 1982):367–388.

Hayashi, N., E. Ostrom, J. Walker, and T. Yamagishi, "Reciprocity, Trust, and the Sense of Control: a Cross-societal Study," *Rationality and Society* 11 (1999):27–46.

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

Henrich, Joe, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, and Herbert Gintis, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-scale Societies* (Oxford: Oxford University Press, 2004).

Henrich, Joseph, 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.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith, "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review* 86,3 (June 1996):653–660.

____ , ____ , Keith Shachat, and Vernon L. Smith, "Preferences, Property Rights, and Anonymity in Bargaining Games," *Games and Economic Behavior* 7 (1994):346–380.

Holt, Charles A., *Industrial Organization: A Survey of Laboratory Research* (Princeton, NJ: Princeton University Press, 1995).

____ , Loren Langan, and Anne Villamil, "Market Power in an Oral Double Auction," *Economic Inquiry* 24 (1986):107–123.

Kagel, J. H. and A. E. Roth, *Handbook of Experimental Economics* (Princeton, NJ: Princeton University Press, 1995).

Kiyonari, Toko, Shigehito Tanida, and Toshio Yamagishi, "Social Exchange and Reciprocity: Confusion or a Heuristic?," *Evolution and Human Behavior* 21 (2000):411–427.

Kreps, David M., *A Course in Microeconomic Theory* (Princeton, NJ: Princeton University Press, 1990).

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) pp. 111–194.

Lerner, Abba, "The Economics and Politics of Consumer Sovereignty," *American Economic Review* 62,2 (May 1972):258–266.

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

Loewenstein, George F., Leigh Thompson, and Max H. Bazerman, "Social Utility and Decision Making in Interpersonal Contexts," *Journal of Personality and Social Psychology* 57,3 (1989):426–441.

Lorenz, Konrad, *Foundations of Ethology* (New York: Springer-Verlag, 1981).

Maynard Smith, John, *Evolution and the Theory of Games* (Cambridge, UK: Cambridge University Press, 1982).

Odling-Smee, F. John, Keven N. Laland, and Marcus W. Feldman, *Niche Construction: The Neglected Process in Evolution* (Princeton: Princeton University Press, 2003).

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.

Plott, Charles R., "The Application of Laboratory Experimental Methods to Public Choice," in Clifford S. Russell (ed.) *Collective Decision Making: Applications from Public Choice Theory* (Baltimore, MD: Johns Hopkins University Press, 1979) pp. 137–160.

Portes, Alejandro, "The Hidden Abode: Sociology as Analysis of the Unexpected," *American Sociological Review* 65,1 (February 2000):1–18.

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.

Sally, David, "Conversation and Cooperation in Social Dilemmas," *Rationality and Society* 7,1 (January 1995):58–92.

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

Schlichting, Carl D. and Massimo Pigliucci, *Phenotypic Evoltuion: A Reaction Norm Perspective* (Sunderland, MA: Sinauer Associated, 1998).

Shizgal, Peter, "On the Neural Computation of Utility: Implications from Studies of Brain Stimulation Reward," in Daniel Kahneman, Edward Diener, and Norbert Schwarz (eds.) *Well-Being: The Foundations of Hedonic Psychology* (New York: Russell Sage, 1999) pp. 502–526.

Shogren, Jason F., "Fairness in Bargaining Requires a Context: An Experimental Examination of Loyalty," *Economic Letters* 31 (1989):319–323.

Smelser, Neil J., "The Rational Choice Perspective: a Theoretical Assessment," *Rational Sociology* 4,4 (1992):381–410.

Smith, Vernon, "Microeconomic Systems as an Experimental Science," *American Economic Review* 72 (December 1982):923–955.

Varian, Hal R., "The Nonparametric Approach to Demand Analysis," *Econometrica* 50 (1982):945–972.

Watabe, M., S. Terai, N. Hayashi, and T. Yamagishi, "Cooperation in the One-Shot Prisoner's Dilemma based on Expectations of Reciprocity," *Japanese Journal of Experimental Social Psychology* 36 (1996):183–196.

Wrong, Dennis H., "The Oversocialized Conception of Man in Modern Sociology," *American Sociological Review* 26 (April 1961):183–193.

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 (1988):265–271.

____ , "Seriousness of Social Dilemmas and the Provision of a Sanctioning System," *Social Psychology Quarterly* 51,1 (1988):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) pp. 267–287.