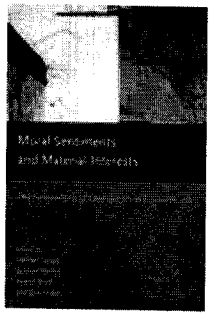
PsycCRITIQUES
Contemporary Psychology: APA Review of Books
American Psychological Asso

Game Theory Revolving

A review of



Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life
by Herbert Gintis, Samuel Bowles, Robert Boyd, and Ernst Fehr (Eds.)

Cambridge, MA: MIT Press, 2005. 404 pp. ISBN 0-262-07252-1. $50.00

Reviewed by
Joachim I. Krueger

Humans are intensely social animals. They are smarter than other species by most accounts, able to look out for themselves and to outwit others. At the same time, they are more dependent on group life and cooperation. Humans require a long period of socialization, and even in adulthood many of their economic needs demand effective coordination and cooperation with others. How do humans solve the dilemma of doing well individually without destroying the groups that sustain them?

Traditional answers fall into two classes. According to one view, which is dominant in the fields of economics and evolutionary biology, self-interest is the principal motive. Concern about others depends on whether others reciprocated aid in the past. Self-interest also contributes to group formation and survival because the strong can coerce the weak. According to another view, which is dominant in the fields of anthropology and sociology, regard for the group is the principal

motive. In the limiting case, self-interest is set aside and replaced by a kind of group mentality. Identification with the group and internalization of its norms are the effective forces that determine behavior. In egalitarian groups, mutual coercion is mutually agreed upon; in stratified groups, a minority of selfish strongmen manage to impose their norms on others.

Homo reciprocans

Gintis, Bowles, Boyd, and Fehr argue that both views are incomplete. To overcome theoretical imperfections and disciplinary divisions, they introduce a third paradigm. In 13 chapters, the editors and their colleagues review evidence for their theory of "strong reciprocity." Strong reciprocity is a "predisposition to cooperate with others, and to punish (at personal cost, if necessary) those who violate the norms of cooperation, even when it is implausible to expect that these costs will be recovered at a later date" (p. 8). Strong reciprocity means that many humans pursue their self-interest while also internalizing social norms. Yet, strong reciprocity is not a mere blend of the two traditional paradigms. Unlike traditional theories, strong reciprocity can explain moral sentiments and morally motivated behaviors among interacting group members. Far from being irrational, the emotions of guilt, envy, pride, and outrage serve critical purposes for both the individual and the group. Likewise, behaviors such as investments, reciprocations, punishments, and rewards are understood as strategic in both a self-interested and collectively reasonable way. Many, if not most, people care about the welfare of others, the fairness of resource distributions, and their own and others' adherence to social norms. The theory of strong reciprocity thus goes beyond individualism and collectivism. It also goes beyond the old nature versus nurture debate. Individuals, groups, and even cultures

are seen as coevolving. This assumption is critical as part of the explanation for why strong reciprocators are not crowded out by pure egoists.

These ideas are not new. In a classic article on trust and suspicion, Morton Deutsch (1958) recognized that for a "cooperative interchange to be a stable ongoing system, each person must have *a way of reacting to violation of his expectation* which is known to the other and which can serve as an inhibitor of violation" (p. 273). Now there is a wealth of evidence supporting Deutsch's insight. Consider positive reciprocity (or "conditional reciprocity"). In a trust game, a trustor decides what fraction of an endowment to turn over to a trustee. Any investment is multiplied at a fixed rate, and the trustee then decides what fraction of the product to return to the trustor. Most trustors and trustees exchange funds, which is irrational from the perspective of pure self-interest. The trustee's generosity constitutes positive reciprocity. Now consider negative reciprocity (or "altruistic punishment"). In the ultimatum game, a proposer offers to share an endowment with a responder. Most proposers offer 50%, and most responders reject offers below 40%. The latter is a punishment that hurts the responder, but it hurts the proposer even more.[1] Again, self-interested rationality forbids such punishments, but strong reciprocity is not irrational. Why?

The public goods game provides an illustration. Each player may contribute a portion of a personal endowment to a common pool. Contributions multiply in value and are redistributed to all players regardless of their individual contributions. A self-interested player contributes nothing. The standard finding, however, is that about half of the players contribute in early rounds of the game, and their contributions gradually decay over time. Positive reciprocity means that players contribute to the extent that others contribute.

This explains why contributions decay only slowly.

A radical change occurs when negative reciprocity is permitted. Fehr and Gächter (2000, reviewed in Chapters 1 and 5) announced all contributions after each round of the game. With everyone's contributions in plain view, concerns about reputation mattered. More important, players were allowed to subtract points from others at a personal cost that was smaller than the punishment. Those who contributed the most also punished the most. Punishment worked; it brought free riders back into the fold, and the group did collectively well. In order to totally unconfound strong reciprocity and self-interest, Fehr and Gächter ran an experiment in which players knew they would never play again with the same people. Nevertheless, they punished the free riders, and contributions did not decay (although they did not increase either). In other words, free riding triggered moral outrage and retaliation among the strong reciprocators, and ultimately the common good was served.

Many ingenious studies are reviewed throughout the book. Falk and Fischbacher (Chapter 6), for example, show that strong reciprocity can be distinguished from simple inequity aversion. Only strong reciprocity involves a sensitivity to others' intentions, moral outrage if those intentions are judged to be bad, and retaliation to send a message. Other chapters explore the implications of strong reciprocity for family relations (Chapter 3) and social policy (Chapters 9-13). The common theme is that they all refute the idea that self-interested rationality can explain collective action. The doctrine of self-interest implies that when collectively satisfactory outcomes fail to emerge spontaneously they can still be attained with explicit regimens of rewards and punishments. Such schemes turn out to be counterproductive because they

"crowd out" the spontaneously grown social preferences for strong reciprocity. Groups that comprise enough strong reciprocators form self-policing communities. Social psychological research on intrinsic motivation supports the idea that explicit norms and centralized enforcement erode cooperation. External rewards and punishments undercut the motivation to act on personal preferences (Deci & Moeller, 2005).

The theory of strong reciprocity poses a more serious challenge to classic notions of self-interested rationality than cognitive theories of judgment and decision-making do. Most cognitive theories accept the doctrine of self-interest but seek to demonstrate that people lack the ability to reason coherently (Kahneman & Tversky, 1984). In contrast, the theory of strong reciprocity does not impugn people's ability to make sound judgments. Instead, it brings to light an adaptive web of preferences, emotions, judgments, and behaviors. Despite the great appeal of this paradigm, three cautionary notes are in order, one theoretical, one empirical, and one pragmatic.

Why Cooperate in a One-Shot Game?

In repeated games, players can learn about the preferences of others, build reputations for themselves, and try to shape the behavior of others. Before free riders can be altruistically punished, their defections have to be on record. Fehr and Fischbacher (Chapter 5) are most explicit in their claim that strong reciprocity applies to one-shot games. The evidence, however, comes from the ultimatum game. Although the ultimatum game can be seen as a one-shot affair, the proposer and the responder have to act in sequence. The responder cannot punish the proposer by refusing the deal without knowing what the deal is. In contrast, players in public goods dilemmas or prisoner's dilemma must act

simultaneously. In a one-shot game, or in the first round of a repeated game, they know nothing of the choices of others. Why do about half of them cooperate? Likewise, why do so many still cooperate in the last round when they know that defection will go unpunished?

Any game theory that ignores the problem of one-shot cooperation is incomplete. One remedial hypothesis is that people generalize to one-shot dilemmas what they have learned from repeated interactions elsewhere. Fehr and Fischbacher (Chapter 5) have little regard for this idea, noting that "the vast majority of the subjects understand the strategic differences between one-shot and repeated interactions quite well" (pp. 156-157). To translate: People know that defection is the dominating choice in one-shot games. If the overgeneralization hypothesis is to be rejected, it must be rejected in all its forms. One cannot assume that people cooperate because they have been punished for defection in the past. They cannot be punished in one-shot games, and they know it. Negative reciprocity does not apply. What about positive reciprocity? The theory says that strong reciprocators will cooperate "if they are sure that the other people who are involved in the cooperation problem will also cooperate" (p. 164). The point of one-shot games is precisely that they cannot be sure. Therefore, positive reciprocity does not apply either.

There is an alternative. The theory of evidential reasoning suggests that people can choose to cooperate in one-shot dilemmas—or in the first and last rounds of repeated dilemmas—if they predict what others will do from what they themselves do. When people have no information about others, they can mentally simulate the implications of their available response options. Should they cooperate, they expect that many others will also cooperate; should they defect, they expect that many others will defect. Given two

different conditional forecasts, people can estimate the expected values of cooperation and defection and select whichever is larger. This decision rule follows from Bayes's theorem, and it does not imply that people believe they can magically cause others to cooperate by cooperating themselves. It is enough to realize that one's own behavior—by definition—is more likely matched than mismatched by others (Krueger & Acevedo, 2005). The theory of evidential reasoning does not diminish the theory of strong reciprocity. The latter explains how cooperation can be increased over repeated interactions, whereas the former explains how cooperation gets started before people know each other. There is a qualitative difference, however. Cooperation as a result of evidential reasoning can be understood in completely self-interested terms. If people predict that others do as they themselves do, they can choose to cooperate because the payoff for mutual cooperation is higher than the payoff for mutual defection. They need not worry about social utilities (i.e., value the benefit for others or abhor unfairness; see also Chapter 4).

The theory of strong reciprocity relies on social categorization as an important moderator variable. Nine of the 13 chapters emphasize that strong reciprocity flourishes within groups to the extent that the groups are homogeneous. If preferences for strong reciprocity evolved in hunter-gatherer kinship groups, they are now extended to larger groups comprising strangers. Group formation is assortative (Chapter 8), and social institutions emphasize existing similarities (Chapter 4). Likewise the theory of evidential reasoning holds that people project their own choices only to those within their own groups. They understand shared group membership as a diagnostic signal for a whole set of self-other similarities (Acevedo & Krueger, 2004). The theory can explain why people reward in-group members, but not out-group members, before

having been rewarded themselves (Gaertner & Insko, 2000).

The Empirical Frontier

- In the reported studies, each player can punish any other player (p. 169). A lone defector can be thoroughly traumatized by a group of strong reciprocators, whereas a lone reciprocator will go bankrupt trying to punish a horde of free riders. Are cumulative punishments more effective, or is there an optimal level? If a single punishment is enough, the reciprocators face a secondary dilemma. Indeed, evidence is reported that reciprocators punish even cooperators who fail to punish free riders. This opens the door to ad absurdum progression. Perhaps those who fail to punish the nonpunishers should also be punished, and so on. Although this is a logical possibility, it seems intuitively unlikely. Perhaps the theory needs a weighting function that characterizes how people discount higher order dilemmas.

- How is it that free riders do not retaliate in turn? Why do they return to cooperation after being punished? Deutsch (1958) recognized this problem and postulated that there must be a "method of absolution" (p. 273). Participants in his experiment communicated by exchanging notes describing their intentions. Whereas the threat of retaliation increased cooperation, the promise of desisting from further punishments after the reestablishment of mutual cooperation was most effective.

- Many real-life dilemmas do not have punishment options. Self-restraining fishermen, donors to public broadcasting, and voters in presidential elections have no means to sour the defectors' payoffs. To be most convincing, the theory of strong reciprocity needs a clear statement of its boundary conditions. The theory must also survive

competition with other theories. Although the contributors report much progress in this regard, a remaining question of interest is how strong reciprocity would fare relative to the simpler tit-for-tat strategy. Both strategies could be programmed and let loose on untutored human players. To win, strong reciprocity would have to yield higher cooperation rates than tit-for-tat. If the rates were the same, tit-for-tat would win because it does not involve punishments or the costs of punishing.

The Moral of the Story

- Groups with a high ratio of strong reciprocators over egoists outcompete other groups. That groups of reciprocators will replace groups of egoists sounds innocuous, even desirable. In social reality, however, "replacement" often means war. In war, defectors are deserters, and strong reciprocators are informers and executioners. The moral sense balks at the idea that Nazi judge Roland Freisler was a strong reciprocator par excellence. What of East Germans spying on their fellow citizens and reporting such infractions as listening to "Westradio"? The ability to punish others can certainly be abused. Someone who holds a grudge can viciously damage another under the pretense of upholding community norms. The theory of strong reciprocity, powerful as it is, will have to find a way to separate the moral sense that drives people to punish others from broader moral issues. At the present stage, this theory, like other game theories, sends an implicit message that cooperation is good. Punishment must be good if it increases cooperation.

- The implications of strong reciprocity for intergroup relations deserve further exploration. The theory locates moral sentiments and the decisions that flow from them within individuals. Intergroup conflict is seen as a by-product of within-group dynamics

(Chapter 4). Few social scientists today seriously endorse the group mind hypothesis (Krueger, Acevedo, & Robbins, 2006). Yet, the actions of groups sometimes uncannily resemble the actions of individuals. When a nation goes to war on behalf of another (England for Poland, the United States for Kuwait), it is seldom clear whether the decision was motivated by self-interest (e.g., access to oil) or negative reciprocity on behalf of supranational norms. Because mobilization is usually announced, and the prospective enemy has the option to pull back, such wars may simply be games of chicken gone bad.

References

Acevedo, M., & Krueger, J. I. (2004). Two egocentric sources of the decision to vote: The voter's illusion and the belief in personal relevance. *Political Psychology, 25*, 115-134. PsycINFO Article

Deci, E. L., & Moeller, A. C. (2005). The concept of competence: A starting point for understanding intrinsic motivation and self-determined extrinsic motivation. In A. J. Elliott & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 579-597). New York: Guilford Press.

Deutsch, M. (1958). Trust and suspicion. *Journal of Conflict Resolution, 2*, 265-279.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment. *American Economic Review, 90*, 980-994.

Fehr, E., & Rockenbach, B. (2003, March 13). Detrimental effects of sanctions on human altruism. *Nature, 422*, 127-140.

Gaertner, L., & Insko, C. A. (2000). Intergroup discrimination in the minimal group paradigm: Categorization, reciprocation, or fear? *Journal of Personality and Social Psychology, 79*, 77-94. PsycINFO Article

Kahneman, D., & Tversky, A. (1984). Choices,

values, and frames. *American Psychologist, 39*, 341-350. PsycINFO Article

Krueger, J. I., & Acevedo, M. (2005). Social projection and the psychology of choice. In M. D. Alicke, D. Dunning, & J. I. Krueger (Eds.), *The self in social perception* (pp. 17-41). New York: Psychology Press.

Krueger, J. I., Acevedo, M., & Robbins, J. M. (2006). Self as sample. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 353-377). New York: Cambridge University Press.

[1]It is interesting to note that negative reciprocity backfires in the trust game. Trustees pay back less if trustors go on record that they will fine trustees who return less than some expected amount (Fehr & Rockenbach, 2003).

PsycCRITIQUES April 12, 2006 Vol. 51 (15), Article 14

1554-0138