Incorporating Fairness into Game Theory and Economics

By MATTHEW RABIN*

People like to help those who are helping them, and to hurt those who are hurting them. Outcomes reflecting such motivations are called fairness equilibria. Outcomes are mutual-max when each person maximizes the other's material payoffs, and mutual-min when each person minimizes the other's payoffs. It is shown that every mutual-max or mutual-min Nash equilibrium is a fairness equilibrium. If payoffs are small, fairness equilibria are roughly the set of mutual-max and mutual-min outcomes; if payoffs are large, fairness equilibria are roughly the set of Nash equilibria. Several economic examples are considered, and possible welfare implications of fairness are explored. (*JEL* A12, A13, D63, C70)

Most current economic models assume that people pursue only their own material self-interest and do not care about "social" goals. One exception to self-interest which has received some attention by economists is simple altruism: people may care not only about their own well-being, but also about the well-being of others. Yet psychological evidence indicates that most altruistic behavior is more complex: people do not seek uniformly to help other people; rather, they do so according to how generous these other people are being. Indeed, *the same people who are altrustic to other altruistic people are also motivated to hurt those who hurt them*. If somebody is being nice to you, fairness dictates that you be nice to him. If somebody is being mean to you, fairness allows—and vindictiveness dictates—that you be mean to him.

Clearly, these emotions have economic implications. If an employee has been exceptionally loyal, then a manger may feel some obligation to treat that employee well, even when it is not in his self-interest to do so. Other examples of economic behavior induced by social goals are voluntary reductions of water-use during droughts, conservation of energy to help solve the energy crisis (as documented, for instance, in Kenneth E. Train et al. [1987]), donations to public television stations, and many forms of voluntary labor. (Burton A. Weisbrod [1988] estimates that, in the United States, the total value of voluntary labor is $74 billion annually.)

On the negative side, a consumer may not buy a product sold by a monopolist at an "unfair" price, even if the material value to the consumer is greater than the price. By not buying, the consumer lowers his own material well-being so as to punish the monopolist. An employee who feels she has been mistreated by a firm may engage in acts of sabotage. Members of a striking labor union may strike longer than is in their material interests because they want to punish a firm for being unfair.

*Department of Economics, 787 Evans Hall, University of California, Berkeley, CA 94720. I thank Nikki Blasberg, Jeff Ely, April Franco, and especially Gadi Barlevy for research assistance; George Akerlof, Colin Camerer, Joe Farrell, Jonathan Feinstein, Ruth Given, Danny Kahneman, Annette Montgomery, Richard Thaler, and especially Bill Dickens for useful conversations; and Gadi Barlevy, Roland Benabou, Nikki Blasberg, Gary Bolton, Jeff Ely, April Franco, Drew Fudenberg, Rachel Kranton, Michael Lii, James Montgomery, Jim Ratliff, Halsey Rogers, Lones Smith, Jeff Zweibel, two anonymous referees, and especially Eddie Dekel-Tabak, Jim Fearon, Berkeley's David Levine, and Vai-Lam Mui for helpful comments on earlier drafts of this paper. I also thank the Institute of Business and Economic Research at the University of California-Berkeley for their funding of research assistance and the National Science Foundation (Grant SES92-10323) for financial support.

By modeling such emotions formally, one can begin to understand their economic and welfare implications more rigorously and more generally. In this paper, I develop a game-theoretic framework for incorporating such emotions into a broad range of economic models.[1] My framework incorporates the following three stylized facts:

(A) People are willing to sacrifice their own material well-being to help those who are being kind.
(B) People are willing to sacrifice their own material well-being to punish those who are being unkind.
(C) Both motivations (A) and (B) have a greater effect on behavior as the material cost of sacrificing becomes smaller.

In the next section, I briefly present some of the evidence from the psychological literature regarding these stylized facts. In Section II, I develop a game-theoretic solution concept "fairness equilibrium" that incorporates these stylized facts. Fairness equilibria do not in general constitute either a subset or a superset of Nash equilibria; that is, incorporating fairness considerations can both add new predictions to economic models and eliminate conventional predictions. In Section III, I present some general results about which outcomes in economic situations are likely to be fairness equilibria. The results demonstrate the special role of "mutual-max" outcomes (in which, given the other person's behavior, each person maximizes the other's material payoffs) and "mutual-min" outcomes (in which, given the other person's behavior, each person minimizes the other's material payoffs). The following results hold:

(i) Any Nash equilibrium that is either a mutual-max outcome or mutual-min outcome is also a fairness equilibrium.
(ii) If material payoffs are small, then, roughly, an outcome is a fairness equilibrium if and only if it is a mutual-max or a mutual-min outcome.
(iii) If material payoffs are large, then, roughly, an outcome is a fairness equilibrium if and only if it is a Nash equilibrium.

I hope this framework will eventually be used to study the implications of fairness in different economic situations. While I do not develop extended applications in this paper, Section IV contains examples illustrating the economic implications of my model of fairness. I develop a simple model of monopoly pricing and show that fairness implies that goods can only be sold at below the classical monopoly price. I then explore the implications of fairness in an extended labor example.

I consider some welfare implications of my model in Section V. Many researchers in welfare economics have long considered issues of fairness to be important in evaluating the desirability of different economic outcomes. Yet while such policy analysis incorporates economists' judgments of fairness and equity, it often ignores the concerns for fairness and equity of the economic actors being studied. By considering how people's attitudes toward fairness influence their behavior and well-being, my framework can help incorporate such concerns more directly into policy analysis and welfare economics.

While my model suggests that the behavioral implications of fairness are greatest when the material consequences of an economic interaction are not too large, there

[1] While many recognize the importance of social motivations in economic phenomena, these emotions have not been investigated widely within the formal apparatus of mainstream economics. Other researchers who have done so include George Akerlof (1982), Peter H. Huang and Ho-Mou Wu (1992), Vai-Lam Mui (1992), and Julio J. Rotemberg (1992); but these and other economic models have tended to be context-specific. While the current version of my model only applies to two-person complete-information games, it applies to *all* such games. If it is extended naturally, it will therefore have specific consequences in any economic or social situation that can be modeled by noncooperative game theory. (By its generality, my model may also contribute to psychological research. While some psychology researchers have tried to formulate general principles of behavior, I believe that noncooperative game theory provides a useful language for doing so more carefully. My model, for instance, helps demonstrate that some seemingly different behaviors in different contexts are explicable by common underlying principles.)

are several reasons why this does not imply that the economic implications of fairness are minor. First, while it is true that fairness influences behavior most when material stakes are small, it is not clear that it makes little difference when material stakes are large. Little empirical research on the economic implications of fairness has been conducted, and much anecdotal evidence suggests that people sacrifice substantial amounts of money to reward or punish kind or unkind behavior. Second, many major economic institutions, most notably decentralized markets, are best described as accumulations of minor economic interactions, so that the aggregate implications of departures from standard theory in these cases may be substantial. Third, the fairness component of a person's overall well-being can be influenced substantially by even small material changes.

Finally, even if material incentives in a situation are so large as to dominate behavior, fairness still matters. Welfare economics should be concerned not only with the efficient allocation of material goods, but also with designing institutions such that people are happy about the way they interact with others. For instance, if a person leaves an exchange in which he was treated unkindly, then his unhappiness at being so treated should be a consideration in evaluating the efficiency of that exchange. Armed with well-founded psychological assumptions, economists can start to address the nonmaterial benefits and costs of the free market and other institutions.[2]

I conclude the paper in Section VI with a discussion of some of the shortfalls of my model and an outline of possible extensions.

[2]Indeed, I show in Section V that there exist situations in which the unique fairness equilibrium leaves both players feeling that they have been treated unkindly. This means that negative emotions may be endogenously generated by particular economic structures. I also state and prove an unhappy theorem: *every* game contains at least one such "unkind equilibrium." That is, there does not exist any situation in which players necessarily depart with positive feelings.

I. Fairness in Games: Some Evidence

In this section, I discuss some psychological research that demonstrates the stylized facts outlined in the Introduction. Consider fact A: "People are willing to sacrifice their own material well-being to help those who are being kind." The attempt to provide public goods without coercion is an archetypical example in which departures from pure self-interest can be beneficial to society, and it has been studied by psychologists as a means of testing for the existence of altruism and cooperation. Laboratory experiments of public goods have been conducted by, among others, John M. Orbell (1978), Gerald Marwell and Ruth Ames (1981), Werner Güth (1982), Alphons J. C. van de Kragt et al. (1983), R. Mark Isaac et al. (1984, 1985), Oliver Kim and Mark Walker (1984), James Andreoni (1988a, b), and Isaac James Walker (1988a, b). These experiments typically involve subjects choosing how much to contribute toward a public good, where the self-interested contribution is small or zero. The evidence from these experiments is that people cooperate to a degree greater than would be implied by pure self-interest. Many of these experiments are surveyed in Robyn M. Dawes and Richard H. Thaler (1988), who conclude that, for most experiments of one-shot public-good decisions in which the individually optimal contribution is close to 0 percent, the contribution rate ranges between 40 percent and 60 percent of the socially optimal level.[3]

These experiments indicate that contributions toward public goods are not, however, the result of "pure altruism," where people seek unconditionally to help others. Rather, the willingness to help seems highly contingent on the behavior of others. If people do not think that others are doing their fair share, then their enthusiasm for sacrificing for others is greatly diminished.

[3]Further examples of stylized fact A can be found in Richard E. Goranson and Leonard Berkowitz (1966), Martin Greenberg and David Frisch (1972), Elizabeth Hoffman and Matthew Spitzer (1982), and Daniel Kahneman et al. (1986a, b).

Indeed, stylized fact B says that people will in some situations not only refuse to help others, but will sacrifice to hurt others who are being unfair. This idea has been most widely explored in the "ultimatum game," discussed at length in Thaler (1988). The ultimatum game consists of two people splitting some fixed amount of money X according to the following rules: a "proposer" offers some division of X to a "decider." If the decider says yes, they split the money according to the proposal. If the decider says no, they both get no money. The result of pure self-interest is clear: proposers will never offer more than a penny, and deciders should accept any offer of at least a penny. Yet experiments clearly reject such behavior. Data show that, even in one-shot settings, deciders are willing to punish unfair offers by rejecting them, and proposers tend to make fair offers.[4] Some papers illustrating stylized fact B are Goranson and Berkowitz (1966), Jerald Greenberg (1978), Güth et al. (1982), Kahneman et al. (1986a, b), and Alvin E. Roth et al. (1991).

Stylized fact C says that people will not be as willing to sacrifice a great amount of money to maintain fairness as they would be with small amounts of money. It is tested and partially confirmed in Gerald Leventhal and David Anderson (1970), but its validity is intuitive to most people. If the ultimatum game were conducted with \$1, then most deciders would reject a proposed split of (\$0.90, \$0.10). If the ultimatum game were conducted with \$10 million, the vast majority of deciders would accept a proposed split of (\$9 million, \$1 million).[5] Consider also the following example from Dawes and Thaler (1988 p. 145):

> In the rural areas around Ithaca it is common for farmers to put some fresh produce on a table by the road. There is a cash box on the table, and customers are expected to put money in the box in return for the vegetables they take. The box has just a small slit, so money can only be put in, not taken out. Also, the box is attached to the table, so no one can (easily) make off with the money. We think that the farmers who use this system have just about the right model of human nature. They feel that enough people will volunteer to pay for the fresh corn to make it worthwhile to put it out there. The farmers also know that if it were easy enough to take the money, someone would do so.

This example is in the spirit of stylized fact C: people succumb to the temptation to pursue their interests at the expense of others in proportion to the profitability of doing so.

From an economist's point of view, it matters not only whether stylized facts A–C are true, but whether they have important economic implications. Kahneman et al. (1986a, b) present strong arguments that these general issues are indeed important. For anyone unconvinced of the importance of social goals empirically or intuitively, one purpose of this paper is to help test the proposition theoretically: will adding fairness to economic models substantially alter conclusions? If so, in what situations will conclusions be altered, and in what way?

II. A Model

To formalize fairness, I adopt the framework developed by John Geanakoplos, David Pearce, and Ennio Stacchetti (1989) (hereafter, GPS). They modify conventional game theory by allowing payoffs to depend on players' *beliefs* as well as on their actions (see also Itzhak Gilboa and David Schmeidler, 1988).[6] While explicitly incorporating

[4]The decision by proposers to make fair offers can come from at least two motivations: self-interested proposers might be fair because they know unfair offers may be rejected, and proposers themselves have a preference for being fair.

[5]Clearly, however, a higher percentage of deciders would turn down an offer of (\$9,999,999.90, \$0.10) than turn down (\$0.90, \$0.10). In his footnote 6, Thaler (1988) concurs with these intuitions, while pointing out the obvious difficulty in financing experiments of the scale needed to test them fully.

[6]Outside the context of noncooperative game theory, Akerlof and William T. Dickens (1982) presented an earlier model incorporating beliefs directly into people's utility functions.

		Player 2: Opera	Player 2: Boxing
Player 1	Opera	2X, X	0, 0
Player 1	Boxing	0, 0	X, 2X

FIGURE 1. EXAMPLE 1: BATTLE OF THE SEXES

beliefs substantially complicates analysis, I argue that the approach is necessary to capture aspects of fairness. Fortunately, GPS show that many standard techniques and results have useful analogues in these "psychological games."

In developing my model of fairness, I extend the GPS approach with an additional step which I believe will prove essential for incorporating psychology into economic research: I derive psychological games from basic "material games." Whereas GPS provide a technique for analyzing games that already incorporate emotions, I use assumptions about fairness to derive psychological games from the more traditional material description of a situation. Doing so, I develop a model that can be applied generally and can be compared directly to standard economic analysis.

To motivate both the general framework and my specific model, consider Example 1 (see Fig. 1), where X is a positive number. (Throughout the paper, I shall represent games with the positive "scale variable" X. This allows me to consider the effects of increasing or decreasing a game's stakes without changing its fundamental strategic structure.) This is a standard battle-of-the-sexes game: two people prefer to go to the same event together, but each prefers a different event. Formally, both players prefer to play either (opera, opera) or (boxing, boxing) rather than not coordinating; but player 1 prefers (opera, opera), and player 2 prefers (boxing, boxing).

The payoffs are a function only of the moves made by the players. Suppose, however, that player 1 (say) cares not only about his own payoff, but depending on player 2's motives, he cares also about player 2's payoff. In particular, if player 2 seems to be intentionally helping player 1, then player 1 will be motivated to help player 2; if player 2 seems to be intentionally hurting player 1, then player 1 will wish to hurt player 2.

Suppose player 1 believes (a) that player 2 is playing boxing, and (b) that player 2 believes player 1 is playing boxing. Then player 1 concludes that player 2 is choosing an action that helps both players (playing opera would hurt both players). Because player 2 is not being either generous or mean, neither stylized fact A nor B applies. Thus, player 1 will be neutral about his effect on player 2 and will pursue his material self-interest by playing boxing. If this argument is repeated for player 2, one can show that, in the natural sense, (boxing, boxing) is an equilibrium: if it is common knowledge that this will be the outcome, then each player is maximizing his utility by playing his strategy.

Of course, (boxing, boxing) is a conventional Nash equilibrium in this game. To see the importance of fairness, suppose player 1 believes (a) that player 2 will play boxing, and (b) that player 2 believes that player 1 is playing opera. Now player 1 concludes that player 2 is lowering her own payoff in order to hurt him. Player 1 will therefore feel hostility toward player 2 and will wish to harm her. If this hostility is strong enough, player 1 may be willing to sacrifice his own material well-being, and play opera rather than boxing. Indeed, if both players have a strong enough emotional reaction to each other's behavior, then (opera, boxing) is an equilibrium. If it is common knowledge that they are playing this outcome, then, in the induced atmosphere of hostility, both players will wish to stick with it.

Notice the central role of expectations: player 1's payoffs do not depend simply on the actions taken, but also on his beliefs about player 2's motives. Could these emotions be directly modeled by transforming the payoffs, so that one could analyze this transformed game in the conventional way? This turns out to be impossible. In the natural sense, both of the equilibria discussed above are strict: each player strictly prefers to play his strategy given the equilibrium. In the equilibrium (boxing, boxing), player 1

strictly prefers playing boxing to opera. In the equilibrium (opera, boxing) player 1 strictly prefers opera to boxing. No matter what payoffs are chosen, these statements would be contradictory if payoffs depended solely on the actions taken. To formalize these preferences, therefore, it is necessary to develop a model that explicitly incorporates beliefs.[7] I now construct such a model, applicable to all two-person, finite-strategy games.

Consider a two-player, normal form game with (mixed) strategy sets S_1 and S_2 for players 1 and 2, derived from finite pure-strategy sets A_1 and A_2. Let $\pi_i : S_1 \times S_2 \to \mathbb{R}$ be player i's *material payoffs*.[8]

From this "material game," I now construct a "psychological game" as defined in GPS. I assume that each player's subjective expected utility when he chooses his strategy will depend on three factors: (i) his strategy, (ii) his beliefs about the other player's strategy choice, and (iii) his beliefs about the other player's beliefs about his strategy. Throughout, I shall use the following notation: $a_1 \in S_1$ and $a_2 \in S_2$ represent the strategies chosen by the two players; $b_1 \in S_1$ and $b_2 \in S_2$ represent, respectively, player 2's beliefs about what strategy player 1 is choosing, and player 1's beliefs about what strategy player 2 is choosing; $c_1 \in S_1$ and $c_2 \in S_2$ represent player 1's beliefs about what player 2 believes player 1's strategy is, and player 2's beliefs about what player 1 believes player 2's strategy is.

[7]My point here is that the results I get could not be gotten simply by respecifying the payoffs over the physical actions in the game. Van Kolpin (1993) argues that one can apply conventional game theory to these games by including the choice of beliefs as additional parts of players' strategies.

[8]I shall emphasize pure strategies in this paper, though formal definitions allow for mixed strategies, and all stated results apply to them. One reason I de-emphasize mixed strategies is that the characterization of preferences over mixed strategies is not straightforward. In psychological games, there can be a difference between interpreting mixed strategies literally as purposeful mixing by a player versus interpreting them as uncertainty by other players. Such issues of interpretation are less important in conventional game theory, and consequently incorporating mixed strategies is more straightforward.

The first step to incorporating fairness into the analysis is to define a "kindness function," $f_i(a_i, b_j)$, which measures how kind player i is being to player j. (I assume in this paper that players have a shared notion of kindness and fairness and that they apply these standards symmetrically. In Rabin (1992), I show that most of the results of this paper hold if multiple kindness functions are allowed.)

If player i believes that player j is choosing strategy b_j, how kind is player i being by choosing a_i? Player i is choosing the payoff pair $(\pi_i(a_i, b_j),\ \pi_j(b_j, a_i))$ from among the set of all payoffs feasible if player j is choosing strategy b_j [i.e., from among the set $\Pi(b_j) \equiv \{(\pi_i(a, b_j), \pi_j(b_j, a)) | a \in S_i\}$]. The players might have a variety of notions of how kind player i is being by choosing any given point in $\Pi(b_j)$. While I shall now proceed with a specific (and purposely simplistic) measure of kindness, I show in the Appendix that the results of this paper are valid for any kindness function that specifies the equitable payoffs as some rule for sharing along the Pareto frontier.

Let $\pi_j^h(b_j)$ be player j's highest payoff in $\Pi(b_j)$, and let $\pi_j^\ell(b_j)$ be player j's lowest payoff *among points that are Pareto-efficient* in $\Pi(b_j)$. Let the "equitable payoff" be $\pi_j^e(b_j) = [\pi_j^h(b_j) + \pi_j^\ell(b_j)]/2$. When the Pareto frontier is linear, this payoff literally corresponds to the payoff player j would get if player i "splits the difference" with her among Pareto-efficient points. More generally, it provides a crude reference point against which to measure how generous player i is being to player j. Finally, let $\pi_j^{\min}(b_j)$ be the worst possible payoff for player j in the set $\Pi(b_j)$.

From these payoffs, I define the kindness function. This function captures how much more than or less than player j's equitable payoff player i believes he is giving to player j.

Definition 1: Player i's kindness to player j is given by

$$f_i(a_i, b_j) \equiv \frac{\pi_j(b_j, a_i) - \pi_j^e(b_j)}{\pi_j^h(b_j) - \pi_j^{\min}(b_j)}.$$

If $\pi_j^h(b_j) - \pi_j^{\min}(b_j) = 0$, then $f_i(a_i, b_j) = 0$.

Note that $f_i = 0$ if and only if player i is trying to give player j her equitable payoff.[9] If $f_i < 0$, player i is giving player j less than her equitable payoff. Recalling the definition of the equitable payoff, there are two general ways for f_i to be negative: either player i is grabbing more than his share on the Pareto frontier of $\Pi(b_j)$ or he is choosing an inefficient point in $\Pi(b_j)$. Finally, $f_i > 0$ if player i is giving player j more than her equitable payoff. Recall that this can happen only if the Pareto frontier of $\Pi(b_j)$ is a nonsingleton; otherwise $\pi_j^e = \pi_j^h$.

I shall let the function $\tilde{f}_j(b_j, c_i)$ represent player i's beliefs about how kindly player j is treating him. While I shall keep the two notationally distinct, this function is formally equivalent to the function $f_j(a_j, b_i)$.

Definition 2: Player i's belief about how kind player j is being to him is given by

$$\tilde{f}_j(b_j, c_i) \equiv \frac{\pi_i(c_i, b_j) - \pi_i^e(c_j)}{\pi_i^h(c_i) - \pi_i^{\min}(c_i)}.$$

If $\pi_i^h(c_i) - \pi_i^{\min}(c_i) = 0$, then $\tilde{f}_j(b_j, c_j) = 0$.

Because the kindness functions are normalized, the values of $f_i(\cdot)$ and $\tilde{f}_j(\cdot)$ must lie in the interval $[-1, \frac{1}{2}]$. Further, the kindness functions are insensitive to positive affine transformations of the material payoffs (overall utility, as defined shortly, will be sensitive to such transformations).

These kindness functions can now be used to specify fully the players' preferences. Each player i chooses a_i to maximize his expected utility $U_i(a_i, b_j, c_i)$, which incorporates both his material utility and the players' shared notion of fairness:

$$U_i(a_i, b_j, c_i) \equiv \pi_i(a_i, b_j) + \tilde{f}_j(b_j, c_i) \cdot \left[1 + f_i(a_i, b_j)\right].$$

The central behavioral feature of these preferences reflects the original discussion. If player i believes that player j is treating him badly—$\tilde{f}_j(\cdot) < 0$—then player i wishes to treat player j badly, by choosing an action a_i such that $f_i(\cdot)$ is low or negative. If player j is treating player i kindly, then $\tilde{f}_j(\cdot)$ will be positive, and player i will wish to treat player j kindly. Of course, the specified utility function is such that players will trade off their preference for fairness against their material well-being, and material pursuits may override concerns for fairness.

Because the kindness functions are bounded above and below, this utility function reflects stylized fact C: the bigger the material payoffs, the less the players' behavior reflects their concern for fairness. Thus, the behavior in these games is sensitive to the scale of material payoffs. Obviously, I have not precisely determined the relative power of fairness versus material interest or even given units for the material payoffs; my results in specific examples are, therefore, only qualitative.

Notice that the preferences $V_i(a_i, b_j, c_i) \equiv \pi_i(a_i, b_j) + \tilde{f}_j(b_j, c_i) \cdot f_i(a_i, b_j)$ would yield precisely the same behavior as the utility function $U_i(a_i, b_j, c_i)$. I have made the preferences slightly more complicated so as to capture one bit of realism: whenever player j is treating player i unkindly, player i's overall utility will be lower than his material payoffs. That is, $\tilde{f}_j(\cdot) < 0$ implies $U_i(\cdot) \leq \pi_i(\cdot)$. If a person is treated badly, he leaves the situation bitter, and his ability to take revenge only partly makes up for the loss in welfare.[10]

Because these preferences form a psychological game, I can use the concept of *psychological Nash equilibrium* defined by GPS; this is simply the analog of Nash equilibrium for psychological games, imposing the additional condition that all higher-order

[9] When $\pi^h = \pi^{\min}$, all of player i's responses to b_j yield player j the same payoff. Therefore, there is no issue of kindness, and $f_i = 0$.

[10] As Lones Smith has pointed out to me, however, this specification has one unrealistic implication: if player 1 is being "mean" to player 2 ($f_1 < 0$), then the nicer player 2 is to player 1, the happier is player 1, even if one ignores the implication for material payoffs. While this is perhaps correct if people enjoy making suckers of others, it is more likely that a player will feel guilty if he is mean to somebody who is nice to him.

beliefs match actual behavior. I shall call the solution concept thus defined "fairness equilibrium."[11]

Definition 3: The pair of strategies $(a_1, a_2) \in (S_1, S_2)$ is a *fairness equilibrium* if, for $i = 1, 2$, $j \neq i$,

$$(1) \qquad a_i \in \arg\max_{a \in S_i} U_i(a, b_j, c_i)$$

$$(2) \qquad c_i = b_i = a_i.$$

Is this solution concept consistent with the earlier discussion of Example 1? In particular, is the "hostile" outcome (opera, boxing) a fairness equilibrium? If $c_1 = b_1 = a_1 =$ opera and $c_2 = b_2 = a_2 =$ boxing, then player 2 feels hostility, and $f_2 = -1$. Thus, player 1's utility from playing U is 0 (with $f_1 = -1$) and from playing boxing it is $X - 1$ (with $f_1 = 0$). Thus, if $X < 1$, player 1 prefers opera to boxing given these beliefs. Player 2 prefers boxing to opera. For $X < 1$, therefore, (opera, boxing) is an equilibrium. In this equilibrium, both players are hostile toward each other and unwilling to coordinate with the other if it means conceding to the other player.[12]

Because the players will feel no hostility if they coordinate, both (opera, opera) and (boxing, boxing) are also equilibria for all values of X. Again, these are conventional outcomes; the interesting implication of fairness in Example 1 is that the players' hostility may lead each to undertake costly punishment of the other. The game Prisoner's Dilemma shows, by contrast, that fairness may also lead each player to sacrifice to help the other player (see Fig. 2).

		Player 2: Cooperate	Player 2: Defect
Player 1	Cooperate	4X, 4X	0, 6X
	Defect	6X, 0	X, X

FIGURE 2. EXAMPLE 2: PRISONER'S DILEMMA

Consider the cooperative outcome, (cooperate, cooperate). If it is common knowledge to the players that they are playing (cooperate, cooperate), then each player knows that the other is sacrificing his own material well-being in order to help him. Each will thus want to help the other by playing cooperate, so long as the material gains from defecting are not too large. Thus, if X is small enough (less than $\frac{1}{4}$), (cooperate, cooperate) is a fairness equilibrium.

For any value of X, however, the Nash equilibrium (defect, defect) is also a fairness equilibrium. This is because if it is common knowledge that they are playing (defect, defect), then each player knows that the other is not willing to sacrifice X in order to give the other $6X$. Thus, both players will be hostile; in the outcome (defect, defect), each player is satisfying both his desire to hurt the other and his material self-interest.

The prisoner's dilemma illustrates two issues I discussed earlier. First, one cannot fully capture realistic behavior by invoking "pure altruism." In Example 2, both (cooperate, cooperate) and (defect, defect) are fairness equilibria, and I believe this prediction of the model is in line with reality. People sometimes cooperate, but if each expects the other player to defect, then they both will. Yet, having both of these as equilibria is inconsistent with pure altruism. Suppose that player 1's concern for player 2 were independent of player 2's behavior.

[11] GPS prove the existence of an equilibrium in all psychological games meeting certain continuity and convexity conditions. The kindness function used in the text does not yield utility functions that are everywhere continuous, so that GPS's theorem does not apply (although I have found no counterexamples to existence). As I discuss in Appendix A, continuous kindness functions that are very similar to the one used in the text, and for which all general results hold, can readily be constructed. Such kindness functions would guarantee existence using the GPS theorem.

[12] For $X < \frac{1}{2}$, (boxing, opera) is also an equilibrium. In this equilibrium, both players are with common knowledge "conceding," and both players feel hostile toward each other because both are giving up their best possible payoff in order to hurt the other player. The fact that, for $\frac{1}{2} < X \leq 1$, (opera, boxing) is an equilibrium, but (boxing, opera) is not, might suggest that (opera, boxing) is "more likely."

Then if he thought that player 2 was playing cooperate, he would play cooperate if and only if he were willing to give up $2X$ in order to help player 2 by $4X$; if player 1 thought that player 2 was playing defect, then he would play cooperate if and only if he were willing to give up X in order to help player 2 by $5X$. Clearly, then, if player 1 plays cooperate in response to cooperate, he would play cooperate in response to defect. In order to get the two equilibria, player 1 must care differentially about helping (or hurting) player 2 as a function of player 2's behavior.[13]

The second issue that the prisoner's dilemma illustrates is the role of intentionality in attitudes about fairness. Psychological evidence indicates that people determine the fairness of others according to their motives, not solely according to actions taken.[14] In game-theoretic terms, "motives" can be inferred from a player's choice of strategy from among those choices he has, so what strategy a player *could* have chosen (but did not) can be as important as what strategy he actually chooses. For example, people differentiate between those who take a generous action by choice and those who are forced to do so. Consider Example 3, depicted in Figure 3.

This is the "prisoner's dilemma" in which player 2 is forced to cooperate. It corresponds, for instance, to a case in which someone is forced to contribute to a public good. In this degenerate game, player 1 will always defect, so the unique fairness equilibrium is (defect, cooperate). This contrasts to the possibility of the (cooperate, cooperate) equilibrium in the prisoner's dilemma. The difference is that now player 1 will feel no positive regard for player 2's "decision" to cooperate, because player 2 is not voluntarily doing player 1 any favors; you are not grateful to somebody who is simply doing what he must.[15]

		Player 2: C
Player 1	C	4X, 4X
	D	6X, 0

FIGURE 3. EXAMPLE 3: PRISONER'S NON-DILEMMA

		Player 2: Dare	Player 2: Chicken
Player 1	Dare	-2X, -2X	2X, 0
	Chicken	0, 2X	X, X

FIGURE 4. EXAMPLE 4: CHICKEN

In both Examples 1 and 2, adding fairness creates new equilibria but does not get rid of any (strict) Nash equilibria. Example 4, the game "Chicken"[16] illustrates that fairness *can* rule out strict Nash equilibria (see Fig. 4).

This game is widely studied by political scientists, because it captures well situations in which nations challenge each other. Each country hopes to "dare" while the other country backs down [outcomes (dare, chicken) and (chicken, dare)]; but both dread most of all the outcome (dare, dare), in which neither nation backs down.

[13]Of course, I am ruling out "income effects" and the like as explanations; but that is clearly not what causes the multiplicity of equilibria in public-goods experiments.

[14]Greenberg and Frisch (1972) and Goranson and Berkowitz (1966) find evidence for this proposition, though not in as extreme a form as implied by my model.

[15]Player 1's complete indifference to player 2's plight here is because I have excluded any degree of pure altruism from my model. Indeed, many of the strong results throughout the paper are because I am ruling out pure altruism.

[16]While I will stick to the conventional name for this game, I note that it is extremely speciesist—there is little evidence that chickens are less brave than humans and other animals.

Consider the Nash equilibrium (dare, chicken), where player 1 "dares" and player 2 "chickens out." Is it a fairness equilibrium? In this outcome, it is common knowledge that player 1 is hurting player 2 to help himself. If X is small enough, player 2 would therefore deviate by playing dare, thus hurting both player 1 and himself. Thus, for small X, (dare, chicken) is not a fairness equilibrium; nor, obviously, is (chicken, dare). Both Nash equilibria are, for small enough X, inconsistent with fairness.

Whereas fairness does not rule out Nash equilibrium in Examples 1 and 2, it does so in Example 4. The next section presents several propositions about fairness equilibrium, including one pertaining to why fairness rules out Nash equilibria in Chicken, but not in Prisoner's Dilemma or Battle of the Sexes.

III. Some General Propositions

In the pure-strategy Nash equilibria of Battle of the Sexes, each taking the other player's strategy as given, each player is maximizing the other player's payoff by maximizing his own payoffs. Thus, each player can satisfy his own material interests without violating his sense of fairness. In the Nash equilibrium of Prisoner's Dilemma, each player is minimizing the other player's payoff by maximizing his own. Thus, bad will is generated, and "fairness" means that each player will try to hurt the other. Once again, players simultaneously satisfy their own material interests and their notions of fairness.

These two types of outcomes—where players mutually maximize each other's material payoffs, and where they mutually minimize each other's material payoffs—will play an important role in many of the results of this paper, so I define them formally:

Definition 4: A strategy pair $(a_1, a_2) \in (S_1, S_2)$ is a *mutual-max outcome* if, for $i = 1,2$, $j \neq i$, $a_i \in \arg\max_{a \in S_i} \pi_j(a, a_j)$.

Definition 5: A strategy pair $(a_1, a_2) \in (S_1, S_2)$ is a *mutual-min outcome* if, for $i = 1,2$, $j \neq i$, $a_i \in \arg\min_{a \in S_i} \pi_j(a, a_j)$.

The following definitions will also prove useful. Each of these definitions characterizes an outcome of a game in terms of the value of "kindness" f_i induced by each of the players.

Definition 6: (a) An outcome is *strictly positive* if, for $i = 1,2$, $f_i > 0$. (b) An outcome is *weakly positive* if, for $i = 1,2$, $f_i \geq 0$. (c) An outcome is *strictly negative* if, for $i = 1,2$, $f_i < 0$. (d) An outcome is *weakly negative* if, for $i = 1,2$, $f_i \leq 0$. (e) An outcome is *neutral* if, for $i = 1,2$, $f_i = 0$. (f) An outcome is *mixed* if, for $i = 1,2$, $j \neq i$, $f_i f_j < 0$.

Using these definitions, I state a proposition about two types of Nash equilibria that will necessarily also be fairness equilibria (all proofs are in Appendix B).

PROPOSITION 1: *Suppose that* (a_1, a_2) *is a Nash equilibrium, and either a mutual-max outcome or a mutual-min outcome. Then* (a_1, a_2) *is a fairness equilibrium.*

Note that the pure-strategy Nash equilibria in Chicken do not satisfy either premise of Proposition 1. In each, one player is maximizing the other's payoff, while the other is minimizing the first's payoff. If X is small enough, so that emotions dominate material payoffs, then the player who is being hurt will choose to hurt the other player, even when this action is self-destructive, and will play dare rather than chicken.

While Proposition 1 characterizes Nash equilibria that are necessarily fairness equilibria, Proposition 2 characterizes which outcomes—Nash or non-Nash—can possibly be fairness equilibria.

PROPOSITION 2: *Every fairness equilibrium outcome is either strictly positive or weakly negative.*

Proposition 2 shows that there will always be a certain symmetry of attitude in any fairness equilibrium. It will never be the case that, in equilibrium, one person is kind while the other is unkind.

While Propositions 1 and 2 pertain to all games, irrespective of the scale of material payoffs, I present in the remainder of this

section several results that hold when material payoffs are either arbitrarily large or arbitrarily small. To do so, I will consider classes of games that differ only in the scale of the material payoffs. Given the set of strategies $S_1 \times S_2$ and the payoff functions $(\pi_1(a_1, a_2), \pi_2(a_1, a_2))$, let $\mathcal{G}$ be the set of games with strategies $S_1 \times S_2$ and, for all $X > 0$, material payoffs

$$\left(X \cdot \pi_1(a_1, a_2), X \cdot \pi_2(a_1, a_2)\right).$$

Let $G(X) \in \mathcal{G}$ be the game corresponding to a given value of X.

Consider Chicken again. It can be verified that, if X is small enough, then both (dare, dare) and (chicken, chicken) are fairness equilibria. Note that, while these two outcomes are (respectively) mutual-min and mutual-max outcomes, they are not Nash equilibria. Yet, when X is small, the fact that they are not equilibria in the "material" game is unimportant, because fairness considerations will start to dominate. Proposition 3 shows that the class of "strict" mutual-max and mutual-min outcomes are fairness equilibria for X small enough.

PROPOSITION 3: *For any outcome (a_1, a_2) that is either a strictly positive mutual-max outcome or a strictly negative mutual-min outcome, there exists an $\overline{X}$ such that, for all $X \in (0, \overline{X})$, (a_1, a_2) is a fairness equilibrium in $G(X)$.*

While Proposition 3 gives sufficient conditions for outcomes to be fairness equilibria when material payoffs are small, Proposition 4 gives conditions for which outcomes will not be fairness equilibria when material payoffs are small.

PROPOSITION 4: *Suppose that $(a_1, a_2) \in (S_1, S_2)$ is not a mutual-max income, nor a mutual-min outcome, nor a Nash equilibrium in which either player is unable to lower the payoffs of the other player. Then there exists an $\overline{X}$ such that, for all $X \in (0, \overline{X})$, (a_1, a_2) is not a fairness equilibrium in $G(X)$.*

Together, Propositions 3 and 4 state that, for games with very small material payoffs, finding the fairness equilibria consists approximately of finding the Nash equilibria in each of the following two hypothetical games: (i) the game in which each player tries to maximize the other player's material payoffs and (ii) the game in which each player tries to minimize the other player's material payoffs.

There are two caveats to this being a general characterization of the set of fairness equilibria in low-payoff games. First, Proposition 3 does not necessarily hold for mutual-max or mutual-min outcomes in which players are giving each other the equitable payoffs (i.e., when the outcomes are neutral). Thus, "non-strict" mutual-max and mutual-min outcomes need to be double-checked. Second, it is also necessary to check whether certain types of Nash equilibria in the original game are also fairness equilibria, even though they are neither mutual-max nor mutual-min outcomes. The potentially problematic Nash equilibria are those in which one of the players has no options that will lower the other's material payoffs.

I now turn to the case in which material payoffs are very large. Proposition 5 states essentially that as material payoffs become large, the players' behavior is dominated by material self-interest. In particular, players will play only Nash equilibria if the scale of payoffs is large enough.

PROPOSITION 5: *If (a_1, a_2) is a strict Nash equilibrium for games in $\mathcal{G}$, then there exists an $\overline{X}$ such that, for all $X > \overline{X}$, (a_1, a_2) is a fairness equilibrium in $G(X)$.*[17] *If (a_1, a_2) is not a Nash equilibrium for games in $\mathcal{G}$, then there exists an $\overline{X}$ such that, for all $X > \overline{X}$, (a_1, a_2) is not a fairness equilibrium in $G(X)$.*

The only caveat to the set of Nash equilibria being equivalent to the set of fairness equilibria when payoffs are large is that

[17] A Nash equilibrium is *strict* if each player is choosing his unique optimal strategy. Mixed-strategy Nash equilibria are, for instance, never strict, because they involve the players being indifferent among two or more actions.

some non-strict Nash equilibria are not fairness equilibria.

IV. Two Applications

One context in which fairness has been studied is monopoly pricing (see e.g., Thaler, 1985; Kahneman et al., 1986a,b). Might consumers see conventional monopoly prices as unfair and refuse to buy at such prices even when worth it in material terms? If this is the case, then even a profit-maximizing monopolist would price below the level predicted by standard economic theory. I now present a game-theoretic model of a monopoly and show that this intuition is an implication of fairness equilibrium.

I assume that a monopolist has costs c per unit of production, and a consumer values the product at v. These are common knowledge. The monopolist picks a price $p \in [c, v]$ as the consumer simultaneously picks a "reservation" price $r \in [c, v]$, above which he is not willing to pay. If $p \leq r$, then the good is sold at price p, and the payoffs are $p - c$ for the monopolist and $v - p$ for the consumer. If $p > r$, then there is no sale, and the payoffs are 0 for each player.

Though this is formally an infinite-strategy game, it can be analyzed using my model of fairness.[18] Applying Nash equilibrium allows any outcome. We might, however, further narrow our prediction, because the strategy $r = v$ for the consumer weakly dominates all other strategies (this would also be the result of subgame perfection if this were a sequential game, with the monopolist setting the price first). Thus, if players cared only about material payoffs, a reasonable outcome in this game is the equilibrium where $p = r = v$, so that the monopolist extracts all the surplus from trade.

[18]Note, however, that I have artificially limited the strategy spaces of the players, requiring them to make only mutually beneficial offers; there are problems with the definitions of this paper if the payoff space of a game is unbounded. Moreover, though I believe that all results would be qualitatively similar with more realistic models, the exact answers provided here are sensitive to the specification of the strategy space.

What is the highest price consistent with a fairness equilibrium at which this product could be sold? First, what is the function $f_C(r, p)$, how fair the consumer is being to the monopolist? Given that the monopolist sets p, the only question is whether the monopolist gets profits $p - c$ or profits 0. If $r \geq p$, then the consumer is maximizing both the monopolist's and his own payoffs, so $f_C(r, p) = 0$. If $r < p$, then the consumer is minimizing the monopolist's payoffs, so $f_C(r, p) = -1$. One implication of this is that the monopolist will always exploit its position, because it will never feel positively toward the consumer; thus, $r > p$ cannot be a fairness equilibrium.

Because $r < p$ leads to no trade, this means that the only possibility for an equilibrium with trade is when $p = r$. How fair is the monopolist being to the consumer when $p = r = z$? Calculations show that $f_M(z, z) = [c - z]/2[v - c]$. Because I am considering only values of z between c and v, this number is negative. Any time the monopolist is not setting a price equal to its costs, the consumer thinks that the monopolist is being unfair. This is because the monopolist is choosing the price that extracts as much surplus as possible from the consumer, given the consumer's refusal to buy at a price higher than z.

To see whether $p = r = z$ is a fairness equilibrium for a given z, one must see whether the consumer would wish to deviate by setting $r < z$, thus eliminating the monopolist's profits. The consumer's total utility from $r < z$ is

$$U_C = 0 + f_M(z, z) \cdot [1 + -1] = 0.$$

The consumer's total utility from sticking with strategy $r = z$ is

$$\begin{aligned} U_C &= v - z + f_M(z, z) \cdot [1 + 0] \\ &= v - z + [c - z]/2[v - c]. \end{aligned}$$

Calculations show that the highest price consistent with fairness equilibrium is given

by

$$z^* = [2v^2 - 2cv + c]/[1 + 2v - 2c].$$

This number is strictly less than v when $v > c$. Thus, the highest equilibrium price possible is lower than the conventional monopoly price when fairness is added to the equation. This reflects the arguments of Kahneman et al. (1986a,b): a monopolist interested in maximizing profits ought not to set price at "the monopoly price," because it should take consumers' attitudes toward fairness as a given.

I can further consider some limit results as the stakes become large in this game. Let the monopolist's costs and the consumer's value be $C \equiv cX$ and $V \equiv vX$, respectively. I represent the percentage of surplus that the monopolist is able to extract by $(z^* - C)/(V - C)$. Algebra shows that this equals $[2(V - C)]/[1 + 2(V - C)]$, and the limit of this as X becomes arbitrarily large is 1. That is, the monopolist is able to extract "practically all" of the surplus, because rejecting an offer for the sake of fairness is more costly for the consumer.

Another interesting implication of the model is that $dz^*/dc > 0$ for all parameter values. This means that the higher are the monopolist's costs, the higher the price the consumer will be willing to pay (assuming that the consumer knows the firm's costs). This is one interpretation of the results presented in Thaler (1985): consumers are willing to pay more for the same product from a high-cost firm than from a low-cost firm.

An area of economics where fairness has been widely discussed (more so than in monopoly pricing) is labor economics.[19] I now present an extended example that resembles the "gift-exchange" view of the employment relationship discussed in Akerlof (1982). Consider the situation in which a worker chooses an effort level and the firm simultaneously chooses a benefit level for the worker.[20] Formally, the worker chooses either a high or low effort level: $e \in \{H, L\}$. If $e = H$, the firm receives revenue $R > 0$, and the worker receives disutility γ. If $e = L$, the firm receives no revenue, and the worker experiences no disutility. Simultaneously, the firm chooses a benefit level $b \in [0, R]$. Material payoffs are as follows:

$$\pi_W = \begin{cases} b^{1/2} - \gamma & \text{if } e = H \\ b^{1/2} & \text{if } e = L \end{cases}$$

$$\pi_F = \begin{cases} (R - b)^{1/2} & \text{if } e = H \text{ and } b \le R \\ 0 & \text{if } e = L \text{ or } b > R \end{cases}$$

where π_W is the worker's material payoffs, and π_F is the firm's material payoffs.[21]

This situation is essentially a continuous-strategy prisoner's dilemma, because each player has a dominant strategy: the worker maximizes his material payoffs by choosing $e = L$, and the firm maximizes its material payoffs by choosing $b = 0$. Thus, the unique Nash equilibrium is the nasty one in which $e = L$ and $b = 0$. Because this outcome is also a mutual-min outcome, this will be a fairness equilibrium in which the players feel negatively toward each other.

I now consider the possibility of a positive fairness equilibrium. First observe that the kindness of the worker to the firm is $f_W = \frac{1}{2}$ if the worker puts in high effort, and $f_W = -\frac{1}{2}$ if the worker puts in low effort. This is because $e = H$ involves the worker fully yielding along the Pareto frontier to the firm, and $e = L$ means that the worker is choosing the best Pareto-efficient point for himself, given the firm's choice of b.

Given the worker's choice of effort, the kindest the firm can be to the worker is to choose $b = R$; the least kind is clearly to choose $b = 0$. Therefore the equitable material payoff to the worker is $R^{1/2}/2 - \gamma$ if $e = H$, and $R^{1/2}/2$ if $e = L$. Using this, one

[19] For some examples discussing the role in labor economics of fairness and related issues, see Akerlof (1982), John Bishop (1987), James N. Baron (1988), David I. Levine (1991, 1993), and Rotemberg (1992). In Rabin (1992), I applied this model of fairness to several more examples from labor economics.

[20] This model is a version of one suggested to me by James Montgomery (pers. comm.).

[21] The assumptions that the parties are risk-averse and that the firm's payoff is 0 (rather than negative) if $e = L$ are made for convenience and are not essential.

can calculate that the kindness of the firm to the worker is given by $f_F = (b/R)^{1/2} - \frac{1}{2}$.

Using this, consider the possibility of a positive fairness equilibrium. What is the firm's utility if it is commonly known that the worker is setting $e = \text{H}$? It is given by

$$U_F = (R-b)^{1/2} + \tfrac{1}{2}\left[\tfrac{1}{2} + (b/R)^{1/2}\right].$$

Thus, the firm will maximize its utility by setting $\partial U_F / \partial b = 0$, and one gets the result that $b^* = R/(1+4R)$. With this level of b, the firm's kindness to the worker is $f_F^* = [1/(1+4R)]^{1/2} - \frac{1}{2}$.

Finally, in order for this to constitute a fairness equilibrium, it must be that the worker would wish to set $e = \text{H}$ rather than $e = \text{L}$. The two possible utility levels are:

$$U_W(e = \text{H}) = b^{1/2} - \gamma + \left\{\left[1/(1+4R)\right]^{1/2} - \tfrac{1}{2}\right\}\left(\tfrac{1}{2}\right)$$

$$U_W(e = \text{L}) = b^{1/2} + \left\{\left[1/(1+4R)\right]^{1/2} - \tfrac{1}{2}\right\}\left(-\tfrac{1}{2}\right).$$

Algebra yields the conclusion that the worker would not strictly prefer to choose $e = \text{L}$ if and only if

$$R \le 0.25\left[1/(0.5+\gamma)^{1/2} - 1\right].$$

For all such combinations of R and γ, therefore, there exists a "gift-giving" equilibrium in which the worker sets $e = \text{H}$, and the firm gives the worker a bonus of $b^* = R/(1+4R)$. Note that the larger is γ, the smaller R must be for there to exist a gift-giving equilibrium. The reason for this is roughly as follows. If γ is large, the worker is very tempted to "cheat" the firm by not working hard. The only way he will not cheat is if the firm is being very kind. But the firm's material costs to yielding a given percentage of profits to the worker increases as R increases; thus, only if R is very small will the firm give the worker a generous enough share of profits to induce the worker to be kind.

		Player 2: Grab	Player 2: Share
Player 1	Grab	X, X	2X, 0
	Share	0, 2X	X, X

FIGURE 5. EXAMPLE 5: THE GRABBING GAME

In fact, if $\gamma \ge \frac{1}{2}$, then there is no gift-giving equilibrium, no matter how small is R. This is because the firm's material incentives are such that it will choose to be unkind to the worker, so that the worker will choose to be unkind to the firm. Thus, overall the model says that workers and firms will cooperate if neither is too tempted by material concerns to cheat.

V. Fairness and Welfare

I consider now some welfare implications of fairness.[22] My perspective here is that the full utility functions (combining material payoffs and "fairness payoffs") are the utility functions with which to determine social welfare. As such, I believe one should care not solely about how concerns for fairness support or interfere with material efficiency, but also about how these concerns affect people's overall welfare.

Consider Example 5 (see Fig. 5). In this game, two people are shopping, and there are two cans of soup left. Each person can either try to grab both cans, or not try to grab. If both grab or both do not grab, they each get one can; if one grabs, and the

[22] Robert H. Frank (1988, 1990) and others have explored how the existence of various emotions are understandable as adaptive evolutionary features of humans. While this view of emotions as "adaptive" may be broadly correct, Frank himself emphasizes that emotions can also be destructive in many situations. People's propensity for revenge can be harmful as well as helpful. My model of people's preferences for fairness will help economists do exactly what is done with "material" preferences—study how these preferences play out in different economic settings.

other does not, then the grabber gets both cans. This is a constant-sum version of the prisoner's dilemma: each player has a dominant strategy, and the unique Nash equilibrium is (grab, grab). As in the prisoner's dilemma, the noncooperative (grab, grab) outcome is a fairness equilibrium, no matter the value of X. For small X, however, the positive, mutual-max outcome (share, share) is also a fairness equilibrium. Moreover, because these two fairness equilibria yield the same material payoffs, (share, share) always Pareto-dominates (grab, grab).

Shopping for minor items is a situation in which people definitely care about material payoffs, and this concern drives the nature of the interaction; but they probably do not care a great deal about individual items. If two people fight over a couple of cans of goods, the social grief and bad tempers are likely to be of greater importance to the people than whether they get the cans. Indeed, while both (grab, grab) and (share, share) are fairness equilibria when material payoffs are arbitrarily small, the overall utility in each equilibrium is bounded away from zero.[23] *As the material payoffs involved become arbitrarily small, equilibrium utility levels do not necessarily become arbitrarily small*. This is realistic: no matter how minor the material implications, people's well-being is affected by the friendly or unfriendly behavior of others.

In Example 5, as with many examples in this paper, there is both a strictly positive and a strictly negative fairness equilibrium. Are there games that contain only positive or only negative fairness equilibria? If there are, this could be interpreted as saying that there are some economic situations that endogenously determine the friendliness or hostility of the people involved. More generally, one could consider the question of which types of economic structures are likely to generate which types of emotions.

The prisoner's dilemma illustrates that there do exist situations that endogenously generate hostility. Applying Proposition 5, the only fairness equilibrium of the prisoner's dilemma with very large material payoffs is the Nash equilibrium, where both players defect. This fairness equilibrium is strictly negative. Interpreting a negative fairness equilibrium as a situation in which parties become hostile to each other, this implies that if mutual cooperation is beneficial, but each person has an irresistible incentive to cheat when others are cooperating, then people will leave the situation feeling hostile.

Are there opposite, happier situations, in which the strategic logic of a situation dictates that people will depart on *good* terms? In other words, are there games for which all fairness equilibria yield strictly positive outcomes? Proposition 6 shows that the answer is no.[24]

PROPOSITION 6: *In every game, there exists a weakly negative fairness equilibrium*.

Proposition 6 states that it is never guaranteed that people will part with positive feelings. It implies a strong asymmetry in my model of fairness: there is a bias toward negative feelings. What causes this asymmetry? Recall that if a player is maximizing his own material payoffs, then he is being either mean or neutral to the other player, because being "nice" inherently involves sacrificing material well-being. Thus, while there are situations in which material self-interest tempts a player to be mean even if other players are being kind, material self-interest will never tempt a player to be kind when other players are being mean, because the

[23]In particular, the utility from (share, share) is positive for each player, and the utility from (grab, grab) is negative for each player: (share, share) Pareto-dominates (grab, grab). This again highlights the fact that social concerns take over when material payoffs are small.

[24]The proof of Proposition 6 invokes the existence theorem of GPS, which applies only if the kindness functions are continuous, so that technically I have established this result only when applying variants of the kindness functions that are continuous. See Appendix A for a discussion of the continuity assumption.

only way to be kind is to go against one's material self-interest.

VI. Discussion and Conclusion

The notion of fairness in this paper captures several important regularities of behavior but leaves out other issues. Evidence indicates, for instance, that people's notions of fairness are heavily influenced by the status quo and other reference points. For instance, Kahneman et al. (1986a,b) illustrate that the consumer's view of the fairness of prices charged by a firm can be heavily influenced by what that firm has charged in the past.

Extending the model to more general situations will create issues that do not arise in the simple two-person, normal-form, complete-information games discussed in this paper. The central distinction between two-person games and multiperson games is likely to be how a person behaves when he is hostile to some players but friendly toward others. The implications are clear if he is able to choose whom to help and whom to hurt; it is more problematic if he must choose either to help everybody or to hurt everybody, such as when choosing the contribution level to a public good. Does one contribute to reward those who have contributed or not contribute to punish those who have not contributed.

Extending the model to incomplete-information games is essential for applied research, but doing so will lead to important new issues. Because the theory depends so heavily on the motives of other players, and because interpreting other players' motives depends on beliefs about their payoffs and information, incomplete information is likely to have a dramatic effect on decision-making. Extending the model to sequential games is also essential for applied research. In conventional game theory, observing past behavior can provide information; in psychological games, it can conceivably change the motivations of the players. An important issue arises: can players "force" emotions; that is, can a first-mover do something that will compel a second player to regard him positively? One might imagine, for instance, that an analogue to Proposition 6 might no longer be true, and sequential games could perhaps be used as mechanisms that guarantee positive emotions.

		Player 2	
		Share	Grab
Player 1	Trust	6X, 6X	0, 12X
	Dissolve	5X, 5X	5X, 5X

FIGURE 6. EXAMPLE 6: LEAVING A PARTNERSHIP

Finally, future research can also focus on modeling additional emotions. In Example 6, for instance, my model predicts no cooperation, whereas it seems plausible that cooperation would take place (see Fig. 6).[25]

This game represents the following situation. Players 1 and 2 are partners on a project that has thus far yielded total profits of $10X$. Player 1 must now withdraw from the project. If player 1 dissolves the partnership, the contract dictates that the players split the profits fifty-fifty. But total profits would be higher if player 1 leaves his resources in the project. To do so, however, he must forgo his contractual rights and trust player 2 to share the profits after the project is completed. So, player 1 must decide whether to "dissolve" or to "trust"; if he trusts player 2, then player 2 can either "grab" or "share."

What will happen? According to the notion of fairness in this paper, the only (pure-strategy) equilibrium is for player 1 to split the profits now, yielding an inefficient solution. The desirable outcome (trust, share) is not possible because player 2 will deviate. The reason is that he attributes no positive motive to player 1—while it is true that player 1 trusted player 2, he did so simply to increase his own expected material payoff. No kindness was involved.

One might think that (trust, share) is a reasonable outcome. This would be the outcome, for instance, if it is assumed that

[25]A related example was first suggested to me by Jim Fearon (pers. comm.).

players wish to be kind to those who trust them. If player 1 plays "trust" rather than "split," he is showing he trusts player 2. If player 2 feels kindly toward player 1 as a result of this trust, then he might not grab all the profits. If it is concluded that the idea that people are motivated to reward trust is psychologically sound, then it could be incorporated into formal models.

APPENDIX A: THE KINDNESS FUNCTION CAN BE GENERALIZED

There is a broad class of kindness functions for which all of the results of this paper hold. Indeed, the proofs of all results contained in the body of the paper are general enough that they establish the results for the kindness functions that I now define.

Definition A1 requires that (i) fairness cannot lead to infinitely positive or infinitely negative utility, and (ii) how kind player i is being to player j is an increasing function of how high a material payoff player i is giving player j.

Definition A1: A kindness function is *bounded and increasing* if:

(i) there exists a number N such that $f_i(a_i, b_j) \in [-N, N]$ for all $a \in S_i$ and $b_j \in S_j$; and
(ii) $f_i(a_i, b_j) > f_i(a_i', b_j)$ if and only if $\pi_j(b_j, a_i) > \pi_j(b_j, a_i')$.

Definition A2 requires that the payoff that player j "deserves" is strictly between player j's worst and best Pareto-efficient payoffs, so long as the Pareto frontier is not a singleton.

Definition A2: Consider $\Pi(b_j)$, $\pi_j^h(b_j)$, and $\pi_j^\ell(b_j)$ as defined in the paper. A kindness function $f_i(a_i, b_j)$ is a *Pareto split* if there exists some $\pi_j^e(b_j)$ such that:

(i) $\pi_j(b_j, a_i) > \pi_j^e(b_j)$ implies that $f_i(a_i, b_j) > 0$; $\pi_j(b_j, a_i) = \pi_j^e(b_j)$ implies that $f_i(a_i, b_j) = 0$; and $\pi_j(b_j, a_i) < \pi_j^e(b_j)$ implies that $f_i(a_i, b_j) < 0$;
(ii) $\pi_j^h(b_j) \geq \pi_j^e(b_j) \geq \pi_j^\ell(b_j)$; and
(iii) if $\pi_j^h(b_j) > \pi_j^\ell(b_j)$, then $\pi_j^h(b_j) > \pi_j^e(b_j) > \pi_j^\ell(b_j)$.

Propositions 1, 2, and 6 are all true for any kindness function meeting Definitions A1 and A2. Propositions 3, 4, and 5, however, pertain to when material payoffs are made arbitrarily large or arbitrarily small. In order for these results to hold, one must guarantee that notions of the fairness of particular outcomes do not dramatically change when all payoffs are doubled (say). Definition A3 is a natural way to do so.

Definition A3: A kindness function $f_i(a_i, b_j)$ is *affine* if changing all payoffs for both players by the same affine transformation does not change the value of $f_i(a_i, b_j)$.

All the propositions in this paper hold for any kindness function meeting Definitions A1, A2, and A3. One substantial generalization allowed for here is that the kindness function can be sensitive to affine transformations of one player's payoffs. If all of player 2's payoffs are doubled, then it may be that fairness dictates that he get more—or less—than before. The definition and all of the limit results simply characterize what happens if *both* players' payoffs are comparably changed.

Knowing that the general results of this paper hold for a large class of kindness functions is also important should existence be problematic. While fairness equilibria exist in all of the examples of this paper, I have proved no general existence result and cannot invoke the existence theorem of GPS, because of possible discontinuities.

The kindness function in the text can be discontinuous in b_j at points where $\pi_j^h(b_j) = \pi_j^{\min}(b_j)$; at such points, $\Pi(b_j)$ is a single point, and $f_i(a_i, b_j)$ is set equal to zero independent of a_i. The discontinuity comes from the fact that, by normalizing the kindness function by $[\pi_j^h(b_j) - \pi_j^{\min}(b_j)]$, the kindness function can be bounded away from zero even when $\Pi(b_j)$ is arbitrarily small. While I chose this kindness function so as to emphasize that kindness or meanness can be large issues even when the stakes are small, this property could be made less extreme. For instance, one could choose

the kindness function as

$$g_i(a_i, b_j) \equiv \frac{\pi_j(b_j, a_i) - \pi_j^e(b_j)}{(1-\gamma)[\pi_j^h(b_j) - \pi_j^{\min}(b_j) + \gamma(\pi_j^{\max} - \pi_j^{\min})]}$$

where $\pi_j^{\max}$ and $\pi_j^{\min}$ are player j's maximum and minimum payoffs in the entire game. This kindness function is well-defined for all $\gamma \in (0,1]$, so long as $\pi_j^{\max} \neq \pi_j^{\min}$ (which is true unless one has a game in which no decisions by either player could possibly affect player j's payoff). A second type of discontinuity in the kindness functions is that $\pi_2^e(b_2)$ can be discontinuous in b_2. This discontinuity can be smoothed out with the following definition: for $D > 0$, let

$$\pi_2^e(b_2, D) = \max_{b^* \in B_2}\{\pi_2^e(b^*) + D\|b_2 - b^*\|\}.$$

It can be shown that $\pi_2^e(b_2, D)$ is a well-defined function and is continuous in b_2. To construct a continuous kindness function (and thus allow the application of the GPS existence proof), one need merely replace π_2^e by $\pi_2^e(b_2, D)$ in the above definition. It can be shown (proof available from the author upon request) that there exists a $D > 0$ defined for each game such that the resulting kindness function satisfies Definitions A1, A2, and A3 for all γ. Moreover, by choosing γ arbitrarily close to 0 and D arbitrarily large, one essentially defines kindness functions that are "smoothed" versions of that used in the paper.

While the precise kindness function used is not important to the qualitative results of this paper, the way I specify the overall utility function is perhaps more restrictive. One aspect that clearly determines some of the results in this paper is the fact that I completely exclude "pure altruism"; that is, I assume that unless player 2 is being kind to player 1, player 1 will have no desire to be kind to player 2. Psychological evidence suggests that, while people are substantially motivated by the type of "contingent altruism" I have incorporated into the model, pure altruism can also sometimes be important.

One natural way to expand the utility function to incorporate pure altruism would be as follows:

$$\tilde{U}_i(a_i, b_j, c_i) \equiv \pi_i(a_i, b_j) + \left[\alpha + (1-\alpha)\tilde{f}_j(b_j, c_i)\right][1 + f_i(a_i, b_j)]$$

where $\alpha \in [0,1]$.

In this utility function, if $\alpha > 0$, then the player i will wish to be kind to player j even if player j is being "neutral" to player i. The relative importance of pure versus contingent altruism is captured by the parameter α; if α is small, then outcomes will be much as in the model of this paper; if α is close to 1, then pure altruism will dominate behavior.

As discussed above with regard to the kindness function, my model assumes that the fairness utility is completely independent of the scale of the material payoffs. Consider a situation in which a proposer's offer to split \$1 evenly is rejected by a decider. My model says that the proposer will leave the situation unhappy not only because he has no money, but because he was badly treated. Yet my model implies that the proposer will be as unhappy, *but no more so*, when leaving a situation in which the decider rejected an offer to split \$1 million evenly.

This seems unrealistic—the bitterness he feels should be larger the greater the harm done. The assumption could, however, be relaxed while maintaining all the general results of the paper. I could specify the utility function as:

$$U_i(a_i, b_j, c_i) \equiv \pi_i(a_i, b_j) + G(X) \cdot \tilde{f}_j(b_j, c_i) \cdot [1 + f_i(a_i, b_j)]$$

where $G(X)$ is positive and increasing in X.[26]

[26]This specification and one of the conditions mentioned to maintain the limit results were suggested by Roland Benabou (pers. comm.).

This might create problems for the limit results of the paper. However, the conditions that $G(X)/X \to 0$ as $X \to \infty$ and that $G(X)$ is bounded away from 0 as $X \to 0$ would suffice for all propositions to hold. In this case, I am assuming that a person's fairness utility is less sensitive to the scale of payoffs than is his material utility, not that it is totally insensitive.

Appendix B: Proofs

PROOF OF PROPOSITION 1:

Since (a_1, a_2) is a Nash equilibrium, both players must be maximizing their material payoffs. First, suppose that (a_1, a_2) is a mutual-max outcome. Then both f_1 and f_2 must be nonnegative. Thus, both players have positive regard for the other. Since each player is choosing a strategy that maximizes both his own material well-being and the material well-being of the other player, this must maximize his overall utility.

Next, suppose that (a_1, a_2) is a mutual-min outcome. Then f_1 and f_2 will both be nonpositive, so that each player will be motivated to decrease the material well-being of the other. Since he is doing so while simultaneously maximizing his own material well-being, this must maximize his utility.

PROOF OF PROPOSITION 2:

Suppose that an outcome has one player being positive ($f_i > 0$), while the other player is not being positive ($f_j \leq 0$). If $f_i > 0$, then it must be that player i could increase his payoff in such a way that player j would be harmed, simply by changing his strategy to maximize his own material interest. If $f_j \leq 0$, it is inconsistent with utility maximization for player i not to do so; therefore, this outcome cannot be a fairness equilibrium. The only outcomes consistent with fairness equilibrium, therefore, are those for which both f_i and f_j are strictly positive, or neither is. This establishes the proposition.

PROOF OF PROPOSITION 3:

As $X \to 0$, the gain in material payoffs from changing a strategy approaches zero, and eventually it is dominated by the fairness payoffs. If (a_1, a_2) is a strictly positive mutual-max outcome, each player would strictly prefer to play a_i, since this uniquely maximizes the fairness product. Thus, this is a fairness equilibrium. If (a_1, a_2) is a strictly negative mutual-min outcome, each player would strictly prefer to play a_i, since this uniquely maximizes the fairness product. Thus, this too would be a fairness equilibrium.

PROOF OF PROPOSITION 4:

Suppose that (a_1, a_2) is not a Nash equilibrium. Then (without loss of generality) player 1 is not maximizing his material payoffs.

Suppose that player 1 is not minimizing player 2's payoffs. Then he is not minimizing f_1. Given that player 1 is also not maximizing his own material payoffs, this can be maximizing behavior only if $f_2 > 0$. Player 2 will choose $f_2 > 0$ only if $f_1 > 0$. Thus, both f_1 and f_2 are greater than 0; but if the material payoffs are small, this means that the players must choose to maximize f_1 and f_2, so that this must be a mutual-max outcome.

Suppose that player 1 is not maximizing player 2's payoffs. Then he is not maximizing f_1. If the payoffs are small, and given that player 1 is not maximizing his own payoffs, this implies that $f_2 < 0$. This means, as payoffs are small, that player 1 will minimize player 2's payoffs, so that $f_1 < 0$. If he does so, player 2 will in turn minimize player 1's payoffs. Thus, this outcome is a min-min outcome. This establishes that if (a_1, a_2) is not a mutual-max, mutual-min, or Nash equilibrium, then it will not be a fairness equilibrium for small enough X.

Now suppose that (a_1, a_2) *is* a Nash equilibrium, but one in which each player could lower the other player's material payoffs by changing his strategy. Suppose that (a_1, a_2) is not a mutual-max outcome. Then (without loss of generality) player 1 could increase player 2's material payoffs. Since player 1 is maximizing his own material payoffs in a way that hurts player 2, it is known that $f_1 < 0$. This can be optimal for small X only if $f_2 \leq 0$. If $f_2 < 0$, then earlier arguments imply that this must be a mutual-min outcome. Suppose $f_2 = 0$. Then this can be

optimal for player 2 only if she has no choice of lowering player 1's payoffs; otherwise, the fact that $f_1 < 0$ would compel her to change strategies. This condition on player 2's choices directly contradicts the assumption that she *could* lower player 1's payoffs. This establishes the proposition.

PROOF OF PROPOSITION 5:

If (a_1, a_2) is a strict Nash equilibrium, then the difference in material payoffs from playing the equilibrium strategy versus a nonequilibrium strategy becomes arbitrarily large as X becomes arbitrarily large. Because the fairness gains and losses are independent of X, a_i eventually becomes a strict best reply to a_j as X becomes large.

If (a_1, a_2) is not a Nash equilibrium, then, for at least one player, the benefit in material payoffs from deviating from (a_1, a_2) becomes arbitrarily large as X becomes arbitrarily large. Because the fairness gains and losses are independent of X, a_i is eventually dominated by some other strategy with respect to a_j as X becomes large.

PROOF OF PROPOSITION 6:

From the material game, consider the psychological game from the preferences $V_i \equiv \pi_i(a_i, b_j) + \min[f_j(c_i, b_j), 0] \cdot \min[f_i(a_i, b_j), 0]$. When the kindness functions are continuous, GPS's general existence result means that this game has at least one equilibrium, (a_1^*, a_2^*). I will now argue that any such equilibrium is also a fairness equilibrium.

First, I show that, for $i = 1, 2$, $f_i(a_i^*, a_j^*) \leq 0$. Suppose $f_i(a_i^*, a_j^*) > 0$. Let a_i' be such that $a_i' \in \arg\max_{a \in S_i} \pi_i(a, a_j^*)$. Then

$$V_i(a_i', a_j^*, a_i^*) > V_i(a_i^*, a_j^*, a_i^*)$$

which contradicts the premise. This is because the material payoff to i is higher with a_i' than with a_i^*, and because $f_i(a_i', a_j^*) \leq 0$, so that the fairness payoff cannot be any lower than from a_i^*.

Thus, for $i = 1, 2$, $f_i(a_i^*, a_2^*) \leq 0$; but this implies that, for each player, maximizing $V_i(a_i, a_j^*, a_i^*)$ is the same as maximizing $U_i(a_i, a_j^*, a_i^*)$. Thus, (a_i^*, a_j^*) is a fairness equilibrium.

REFERENCES

Akerlof, George, "Labor Contracts as a Partial Gift Exchange," *Quarterly Journal of Economics*, November 1982, *97*, 543–69.

Akerlof, George and Dickens, William T., "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, June 1982, *72*, 307–19.

Andreoni, James, (1988a) "Privately Provided Public Goods in a Large Economy: The Limits of Altruism," *Journal of Public Economics*, February 1988, *35*, 57–73.

_____, (1988b) "Why Free Ride? Strategies and Learning in Public Goods Experiments," *Journal of Public Economics*, December 1988, *37*, 291–304.

Baron, James N., "The Employment Relation as a Social Relation," *Journal of the Japanese and International Economies*, December 1988, *2*, 492–525.

Bishop, John, "The Recognition and Reward of Employee Performance," *Journal of Labor Economics*, October 1987, *5*, S36–56.

Dawes, Robyn M. and Thaler, Richard H., "Anomalies: Cooperation," *Journal of Economic Perspectives*, Summer 1988, *2*, 187–98.

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

_____, "A Theory of Moral Sentiments," in Jane J. Mansbridge, ed., *Beyond Self-Interest*, Chicago: University of Chicago Press, 1990, pp. 71–96.

Geanakoplos, John, Pearce, David and Stacchetti, Ennio, "Psychological Games and Sequential Rationality," *Games and Economic Behavior*, March 1989, *1*, 60–79.

Gilboa, Itzhak and Schmeidler, David, "Information Dependent Games: Can Common Sense Be Common Knowledge?" *Economics Letters*, 1988, *27* (3), 215–21.

Goranson, Richard E. and Berkowitz, Leonard, "Reciprocity and Responsibility Reac-

tions to Prior Help," *Journal of Personality and Social Psychology*, February 1966, *3*, 227–32.

Greenberg, Jerald, "Effects of Reward Value and Retaliative Power on Allocation Decisions: Justice, Generosity or Greed?" *Journal of Personality and Social Psychology*, April 1978, *36*, 367–79.

Greenberg, Martin S. and Frisch, David, "Effect of Intentionality on Willingness to Reciporocate a Favor," *Journal of Experimental Social Psychology*, March 1972, *8*, 99–111.

Güth, Werner, Schmittberger, Rolf and Schwarze, Bernd, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, December 1982, *3*, 367–88.

Hoffman, Elizabeth and Spitzer, Matthew, "The Coase Theorem: Some Experimental Tests," *Journal of Law and Economics*, April 1982, *75*, 73–98.

Huang, Peter H. and Wu, Ho-Mou, "Emotional Responses in Litigation," *International Review of Law and Economics*, March 1992, *12*, 31–44.

Isaac, R. Mark, McCue, Kenneth F. and Plott, Charles, "Public Goods Provision in an Experimental Environment," *Journal of Public Economics*, February 1985, *26*, 51–74.

_____ **and Walker, James,** (1988a) "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism," *Quarterly Journal of Economics*, February 1988, *103*, 179–99.

_____ **and** _____, (1988b) "Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism," *Economic Inquiry*, October 1988, *26*, 585–608.

_____, _____ **and Thomas, Susan H.,** "Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations," *Public Choice*, 1984, *43* (2), 113–49.

Kahneman, Daniel, Knetsch, Jack L. and Thaler, Richard H., (1986a) "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," *American Economic Review*, September 1986, *76*, 728–41.

_____, _____ **and** _____, (1986b) "Fairness and the Assumptions of Economics," *Journal of Business*, October 1986, *59*, S285–S300.

Kim, Oliver and Walker, Mark, "The Free Rider Problem: Experimental Evidence," *Public Choice*, 1984, *43* (1), 3–24.

Kolpin, Van, "Equilibrium Refinements in Psychological Games," *Games and Economic Behavior*, 1993 (forthcoming).

Leventhal, Gerald and Anderson, David, "Self-Interest and the Maintenance of Equity," *Journal of Personality and Social Psychology*, May 1970, *15*, 57–62.

Levine, David I., "Cohesiveness, Productivity, and Wage Dispersion," *Journal of Economic Behavior and Organization*, March 1991, *15*, 237–55.

_____, "Fairness, Markets, and Ability To Pay: Evidence from Compensation Executives," *American Economic Review*, December 1993, *83* (5), 1241–59.

Marwell, Gerald and Ames, Ruth, "Economists Free Ride, Does Anyone Else?: Experiments on the Provision of Public Goods, IV," *Journal of Public Economics*, June 1981, *15*, 295–310.

Mui, Vai-Lam, "Two Essays in the Economics of Institutions: I. Envy," Ph.D. dissertation, Department of Economics, University of California-Berkeley, 1992.

Orbell, John M., Dawes, Robyn M. and van de Kragt, Alphons J. C., "Explaining Discussion Induced Cooperation," *Journal of Personality and Social Psychology*, May 1978, *54*, 811–19.

Rabin, Matthew, "Incorporating Fairness Into Game Theory and Economics," Department of Economics Working Paper No. 92-199, University of California-Berkeley, July 1992.

Rotemberg, Julio J., "Human Relations in the Workplace," mimeo, Massachusetts Institute of Technology, January 1992.

Roth, Alvin E., Prasnikar, Vesna, Okuno-Fujiwara, Masahiro and Zamir, Shmuel, "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, December 1991, *81*, 1068–95.

Thaler, Richard H., "Mental Accounting and

Consumer Choice," *Marketing Science*, Summer 1985, *4*, 199–214.

_____, "Anomalies: The Ultimatum Game," *Journal of Economic Perspectives*, Fall 1988, *2*, 195–207.

Train, Kenneth E., McFadden, Daniel L. and Goett, Andrew A., "Consumer Attitudes and Voluntary Rate Schedules for Public Utilities," *Review of Economics and Statistics*, August 1987, *64*, 383–91.

van de Kragt, Alphons J. C., Orbell, John M. and Dawes, Robyn M., "The Minimal Contributing Set as a Solution to Public Goods Problems," *American Political Science Review*, March 1983, *77*, 112–22.

Weisbrod, Burton A., *The Nonprofit Economy*, Cambridge, MA: Harvard University Press, 1988.