UNIVERSITY OF CALIFORNIA AT BERKELEY

Department of Economics

Berkeley, California 94720

Working Paper No. 91-185

Incorporating Fairness into Game Theory

Matthew Rabin

Economics Department
University of California at Berkeley

December 1991

Key words: altruism, fairness, psychological games, public goods, revenge

JEL Classification: A12, A13, D63, C70

I thank Nikki Blasberg, Jeff Ely, and April Franco for research assistance, George Akerlof, Bill Dickens, Jonathan Feinstein, and Danny Kahneman for useful conversations on related topics, and Gadi Bar-Levy, Roland Benabou, Nikki Blasberg, Jeffy Ely, April Franco, Drew Fudenberg, Jim Ratliff, Halsey Rogers, Lones Smith, Jeff Zweibel, and especially Eddie Dekel-Tabak, Jim Fearon, and Vai-Lam Mui for very helpful comments on earlier drafts of this paper. I also thank the Institute of Business and Economic Research at the University of California at Berkeley for their funding of research assistance.

Abstract

Psychological evidence shows that, rather than pursuing solely their own material interests in group situations, people have additional "social" goals: They wish to help those who are helping them, and hurt those who are hurting them. In this paper, I model such behavior in non-cooperative game theory, and define the solution concept "Fairness Equilibrium" as those outcomes that constitute equilibrium behavior when such motives are added to material games. I apply the model to some well-known games and a model of monopoly pricing.

Applying the model shows the special role of "Mutual-Max" outcomes—in which each player maximizes the other's material payoffs—and "Mutual-Min" outcomes—in which each player minimizes the other's material payoffs. The following results hold: Any Nash equilibrium that is either a Mutual-Max outcome or Mutual-Min outcome is also a fairness equilibrium. If the material payoffs are small relative to the "psychological payoffs," then, roughly, an outcome is a fairness equilibrium if and only if it is a Mutual-Max or a Mutual-Min outcome. If the material payoffs are large, then an outcome is a fairness equilibrium if it is a strict Nash equilibrium, and only if it is a Nash equilibrium.

I. Introduction

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 theory and evidence indicate 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 altruistic 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 can have economic implications. If an employee has been exceptionally loyal to a company, then a manager may feel some obligation to treat that employee well, even when it is not in his self-interest to do so (For the related idea of workers and management giving "gifts" to each other, see Akerlof [1982]). 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 Train, McFadden, and Goett [1987]), donations to public television stations, and many forms of voluntary labor (Weisbrod [1988] estimates that, in the U.S., 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.

In this paper, I formally incorporate such social goals into non-cooperative game theory. By modeling these emotions formally, we can begin to understand their economic implications more rigorously and more generally.

In the next section, I briefly present some of the evidence from the psychological literature, and outline more specifically the stylized facts about behavior that I incorporate into my model. In Section III, I develop the solution concept "fairness equilibrium," and discuss its implications in some examples.

In Section IV, I present some general results about which outcomes in material games are likely to be fairness equilibria. These results demonstrate the special role of "Mutual-Max" outcomes--in which each player maximizes the other's material payoffs--and "Mutual-Min" outcomes--in which each player minimizes the other's material payoffs. The following results hold: Any Nash equilibrium that is either a Mutual-Max outcome or Mutual-Min outcome is also fairness equilibrium. If the material payoffs of a game are small, then, roughly, an outcome is a fairness equilibrium if and only if it is a Mutual-Max or a Mutual-Min outcome. If the material payoffs of a game are large, then an outcome is a fairness equilibrium if it is a strict Nash equilibrium, and only if it is a Nash equilibrium.

In Section V, I discuss the welfare implications of fairness. I believe that 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 unfairly, this makes him unhappy, and this should be a real consideration in judging the efficiency of a situation. Indeed, if we arm ourselves with well-founded psychological assumptions, we can start to address the non-material benefits and costs of the free market and other institutions.

I show in Section V that there exist situations in which the unique "fairness equilibrium" leaves both players "bitter"--they leave the situation feeling that they have been treated badly. This has strong implications: 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 "bitter equilibrium," so that there do *not* exist any situations in which players necessarily depart with positive feelings.

I hope eventually to use this or related models to formally study the implications of fairness in different economic situations. While I do not do so in depth in this paper, Section VI considers the role of fairness in monopoly pricing. I conclude in Section VII with a discussion of some of the shortfalls of my model, and an outline of possible revisions and extensions.

II. Fairness in Games: Some Evidence

In this section, I discuss some psychological research that demonstrates the importance of the emotions I shall incorporate into my model. My model will reflect the following stylized facts:

[A] People are willing to sacrifice their own material well-being to help those who are being fair;

[B] People are willing to sacrifice their own material well-being to punish those who are being unfair;

[C] Both motivations [A] and [B] have less effect on behavior as the material cost of sacrificing becomes larger;

[D] People determine the fairness of others according to their motives, not solely according to actions taken. For example, people differentiate between those who take a generous action by choice and those who are forced to do so.

Consider [A]. The attempt to provide public goods without coercion is an archetypical example where 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, Isaac, Walker, and Thomas [1984], Isaac, McCue, and Plott [1985], Isaac and Walker [1988a, 1988b], Kim and Walker [1984], Marwell and Ames [1981], van de Kragt, Dawes, and Orbell [1983], van de Kragt, Orbell, and Dawes [1982], Guth, Schmittberger, and Schwarze [1982], and Andreoni [1988]. These experiments typically involve subjects choosing how much to contribute towards a public good, where the optimal 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 Dawes and Thaler [1988], and they conclude that, for most experiments of one-shot public-good decisions in which the individually optimal contribution is close to 0%, the contribution rate ranges between 40% and 60% of the socially optimal level.[1]

These experiments indicate that contributions towards public goods are *not*, however, the result of "pure altruism," where people seek unconditionally to

[1] Further examples of Stylized Fact [A] can be found in Greenberg and Frisch [1972], Kahneman, Knetsch, and Thaler [1986a, 1986b], Hoffman and Spitzer [1982], and Goranson and Berkowitz [1966].

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.

Indeed, Stylized Fact [B] says 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 the Decider should accept any offer of at least a penny. Yet experiments clearly reject such behavior: Data show that Deciders are willing to punish unfair offers by rejecting them, and that Proposers tend to make fair offers.[2]

Some papers illustrating Stylized Fact [B] are Kahneman, Knetsch, and Thaler [1986a,1986b], Guth, Schmittberger, and Schwarze [1982], Greenberg [1978], Finn and Lee [1986], and Goranson and Berkowitz [1966].

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 Leventhal and Anderson [1970], but its validity is intuitive to most of us. If the ultimatum game were conducted with $1, then most Deciders would reject a proposed split of ($.90,$.10). If the ultimatum game were conducted with $10 million, the vast

[2] The decision by Proposers to make fair offers can come from at least two motivations: Self-interested Proposers should be fair because they know unfair offers will be rejected, and Proposers themselves have a preference for being fair.

majority of Deciders would *accept* a proposed split of ($9 million, $1 million).[3] Consider also the following example from Dawes and Thaler [1988]:

> 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.

Greenberg and Frisch [1972] and Goranson and Berkowitz [1966] find evidence for Proposition [D], though not in as extreme a form as implied by my model.

From an economist's point of view, it matters not only whether stylized facts [A] to [D] are true, but whether they have important economic implications. Kahneman, Knetsch, and Thaler [1986a,1986b] discuss this at length, and are convincing that the general issues are indeed important. For those unconvinced by this empirically or intuitively, one purpose of this paper is to help us actually test the proposition *theoretically*: Will adding fairness to economic models substantially alter our conclusions? If so, in what situations will our conclusions be altered, and in what way?

[3] Clearly, however, a higher percentage of Deciders would turn down an offer of ($9,999,999.90,$.10) than turn down ($.90,$.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.

III. A Model

To formalize fairness, I adopt the framework developed by Geanakoplos, Pearce, and Stachetti [1989] (hereafter, GPS). They modify conventional game theory by allowing payoffs to depend on players' *beliefs* as well as on their actions.[4] While explicitly incorporating 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 analogs in these "psychological games."

In this paper, I extend the GPS approach with an additional step which I think will facilitate economic research: I *derive* psychological games from basic "material games." Whereas GPS provide a technique for analyzing games that already incorporate emotions into them, 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, where X is a positive number. (Throughout the paper, I shall represent games with the positive "scale variable" X. This allows us 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: both players prefer to play either (U,L) or (D,R) rather than not coordinating; but player 1 prefers (U,L) and player 2 prefers (D,R).

[4] See also Gilboa and Schmeidler [1988].

		Player 2	
		L	R
Player 1	U	2X,X	0,0
	D	0 , 0	X,2X

Example 1 -- Battle of the Sexes

The payoffs drawn 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 R, and b) that player 2 believes he is playing D. Then player 1 concludes that player 2 is choosing an action that helps both players (playing L would hurt both players). Because player 2 is not being either generous or mean, neither stylized fact [A] nor [B] apply. Thus, player 1 will be neutral about his effect on player 2, and pursue his material self interest by playing D. If we repeat this argument for player 2, we can show that, in the natural sense, (D,R) 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, (D,R) is a conventional Nash equilibrium in this game. To see the importance of fairness, suppose player 1 believes a) that player 2 will play R, and b) that player 2 believes that he is playing U. Now player 1 concludes that player 2 is lowering her own payoff in order to hurt him. Player 1 will therefore feel hostility towards player 2, and wish to harm her.

If this hostility is strong enough, player 1 may be willing to sacrifice his own material well-being, and play U rather than D. Indeed, if both players have a strong enough emotional reaction to each other's behavior, then (U,R) 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 we 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 (D,R), player 1 strictly prefers playing D to U. In the equilibrium (U,R), player 1 strictly prefers U to D. No matter what payoffs we choose, these statements would be contradictory if payoffs depended solely on the actions taken. To formalize these preferences, therefore, we need to develop a model that explicitly incorporates beliefs. 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 $\pi_i : S_1 \times S_2 \longrightarrow \mathbb{R}$ be player i's *material payoffs*.[5]

[5] I shall emphasize pure strategies in most of the paper, though formal definitions allow for mixed strategies. 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. Another reason I de-emphasize mixed strategies is that they are hard to solve for; Mathematica was used to find the two mixed-strategy equilibria in the Prisoner's Dilemma discussed below.

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: 1) his strategy, 2) what he believes the other player's strategy to be, and 3) what he believes the other player believes his strategy to be. 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.

The first step to incorporating fairness into our analysis is to define a "kindness" function, $f_i(a_i,b_j)$, which measures how kind player i is being to player j.[6]

If player i believes that player j is choosing strategy b_j, how kind is player i being by choosing a_i? Well, 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 define in Appendix A a relatively broad class of kindness functions for which all of the results of this paper are valid.

[6] I assume in this paper that players have a shared notion of kindness and fairness, and that they apply these standards symmetrically. While I believe that this is appropriate for modeling purposes, psychological evidence suggests that people do not all share notions of fairness, and--more importantly--they select notions of fairness with a strong bias towards those that justify pursuing their own material interests. I discuss in Appendix B how multiple kindness functions can be employed.

Let $\pi_j^h(b_j)$ be player j's highest payoff in $\Pi(b_j)$, and let $\pi_j^l(b_j)$ be player j's lowest payoff *among points that are Pareto-efficient in* $\Pi(b_j)$. Let the "equitable payoff" be $\pi_j^e(b_j) = (\pi_j^h(b_j) + \pi_j^l(b_j))/2$. In the case where 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 1.1:

Player i's kindness to player j is given by

$$f_i(a_i,b_j) \equiv [\pi_j(b_j,a_i) - \pi_j^e(b_j)]/[\pi_j^h(b_j) - \pi_j^{min}(b_j)];$$

if $\pi_j^h(b_j) - \pi_j^{min}(b_j) = 0$, then $f_i(a_i,b_j) = 0$.[7]

Note that $f_i = 0$ if and only if player i is trying to give player j her equitable payoff. 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 non-singleton; otherwise, $\pi_j^e = \pi_j^h$.

[7] 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 $f_i = 0$.

Abstract

Psychological evidence shows that, rather than pursuing solely their own material interests in group situations, people have additional "social" goals: They wish to help those who are helping them, and hurt those who are hurting them. In this paper, I model such behavior in non-cooperative game theory, and define the solution concept "Fairness Equilibrium" as those outcomes that constitute equilibrium behavior when such motives are added to material games. I apply the model to some well-known games and a model of monopoly pricing.

Applying the model shows the special role of "Mutual-Max" outcomes—in which each player maximizes the other's material payoffs—and "Mutual-Min" outcomes—in which each player minimizes the other's material payoffs. The following results hold: Any Nash equilibrium that is either a Mutual-Max outcome or Mutual-Min outcome is also a fairness equilibrium. If the material payoffs are small relative to the "psychological payoffs," then, roughly, an outcome is a fairness equilibrium if and only if it is a Mutual-Max or a Mutual-Min outcome. If the material payoffs are large, then an outcome is a fairness equilibrium if it is a strict Nash equilibrium, and only if it is a Nash equilibrium.

I. Introduction

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 theory and evidence indicate 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 altruistic 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 can have economic implications. If an employee has been exceptionally loyal to a company, then a manager may feel some obligation to treat that employee well, even when it is not in his self-interest to do so (For the related idea of workers and management giving "gifts" to each other, see Akerlof [1982]). 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 Train, McFadden, and Goett [1987]), donations to public television stations, and many forms of voluntary labor (Weisbrod [1988] estimates that, in the U.S., 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.

In this paper, I formally incorporate such social goals into non-cooperative game theory. By modeling these emotions formally, we can begin to understand their economic implications more rigorously and more generally.

In the next section, I briefly present some of the evidence from the psychological literature, and outline more specifically the stylized facts about behavior that I incorporate into my model. In Section III, I develop the solution concept "fairness equilibrium," and discuss its implications in some examples.

In Section IV, I present some general results about which outcomes in material games are likely to be fairness equilibria. These results demonstrate the special role of "Mutual-Max" outcomes--in which each player maximizes the other's material payoffs--and "Mutual-Min" outcomes--in which each player minimizes the other's material payoffs. The following results hold: Any Nash equilibrium that is either a Mutual-Max outcome or Mutual-Min outcome is also fairness equilibrium. If the material payoffs of a game are small, then, roughly, an outcome is a fairness equilibrium if and only if it is a Mutual-Max or a Mutual-Min outcome. If the material payoffs of a game are large, then an outcome is a fairness equilibrium if it is a strict Nash equilibrium, and only if it is a Nash equilibrium.

In Section V, I discuss the welfare implications of fairness. I believe that 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 unfairly, this makes him unhappy, and this should be a real consideration in judging the efficiency of a situation. Indeed, if we arm ourselves with well-founded psychological assumptions, we can start to address the non-material benefits and costs of the free market and other institutions.

I show in Section V that there exist situations in which the unique "fairness equilibrium" leaves both players "bitter"--they leave the situation feeling that they have been treated badly. This has strong implications: 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 "bitter equilibrium," so that there do *not* exist any situations in which players necessarily depart with positive feelings.

I hope eventually to use this or related models to formally study the implications of fairness in different economic situations. While I do not do so in depth in this paper, Section VI considers the role of fairness in monopoly pricing. I conclude in Section VII with a discussion of some of the shortfalls of my model, and an outline of possible revisions and extensions.

II. Fairness in Games: Some Evidence

In this section, I discuss some psychological research that demonstrates the importance of the emotions I shall incorporate into my model. My model will reflect the following stylized facts:

[A] People are willing to sacrifice their own material well-being to help those who are being fair;

[B] People are willing to sacrifice their own material well-being to punish those who are being unfair;

[C] Both motivations [A] and [B] have less effect on behavior as the material cost of sacrificing becomes larger;

[D] People determine the fairness of others according to their motives, not solely according to actions taken. For example, people differentiate between those who take a generous action by choice and those who are forced to do so.

Consider [A]. The attempt to provide public goods without coercion is an archetypical example where 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, Isaac, Walker, and Thomas [1984], Isaac, McCue, and Plott [1985], Isaac and Walker [1988a,1988b], Kim and Walker [1984], Marwell and Ames [1981], van de Kragt, Dawes, and Orbell [1983], van de Kragt, Orbell, and Dawes [1982], Guth, Schmittberger, and Schwarze [1982], and Andreoni [1988]. These experiments typically involve subjects choosing how much to contribute towards a public good, where the optimal 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 Dawes and Thaler [1988], and they conclude that, for most experiments of one-shot public-good decisions in which the individually optimal contribution is close to 0%, the contribution rate ranges between 40% and 60% of the socially optimal level.[1]

These experiments indicate that contributions towards public goods are *not*, however, the result of "pure altruism," where people seek unconditionally to

[1] Further examples of Stylized Fact [A] can be found in Greenberg and Frisch [1972], Kahneman, Knetsch, and Thaler [1986a,1986b], Hoffman and Spitzer [1982], and Goranson and Berkowitz [1966].

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.

Indeed, Stylized Fact [B] says 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 the Decider should accept any offer of at least a penny. Yet experiments clearly reject such behavior: Data show that Deciders are willing to punish unfair offers by rejecting them, and that Proposers tend to make fair offers.[2]

Some papers illustrating Stylized Fact [B] are Kahneman, Knetsch, and Thaler [1986a,1986b], Guth, Schmittberger, and Schwarze [1982], Greenberg [1978], Finn and Lee [1986], and Goranson and Berkowitz [1966].

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 Leventhal and Anderson [1970], but its validity is intuitive to most of us. If the ultimatum game were conducted with $1, then most Deciders would reject a proposed split of ($.90,$.10). If the ultimatum game were conducted with $10 million, the vast

[2] The decision by Proposers to make fair offers can come from at least two motivations: Self-interested Proposers should be fair because they know unfair offers will be rejected, and Proposers themselves have a preference for being fair.

majority of Deciders would *accept* a proposed split of ($9 million, $1 million).[3] Consider also the following example from Dawes and Thaler [1988]:

> 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.

Greenberg and Frisch [1972] and Goranson and Berkowitz [1966] find evidence for Proposition [D], though not in as extreme a form as implied by my model.

From an economist's point of view, it matters not only whether stylized facts [A] to [D] are true, but whether they have important economic implications. Kahneman, Knetsch, and Thaler [1986a,1986b] discuss this at length, and are convincing that the general issues are indeed important. For those unconvinced by this empirically or intuitively, one purpose of this paper is to help us actually test the proposition *theoretically*: Will adding fairness to economic models substantially alter our conclusions? If so, in what situations will our conclusions be altered, and in what way?

[3] Clearly, however, a higher percentage of Deciders would turn down an offer of ($9,999,999.90,$.10) than turn down ($.90,$.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.

III. A Model

To formalize fairness, I adopt the framework developed by Geanakoplos, Pearce, and Stachetti [1989] (hereafter, GPS). They modify conventional game theory by allowing payoffs to depend on players' *beliefs* as well as on their actions.[4] While explicitly incorporating 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 analogs in these "psychological games."

In this paper, I extend the GPS approach with an additional step which I think will facilitate economic research: I *derive* psychological games from basic "material games." Whereas GPS provide a technique for analyzing games that already incorporate emotions into them, 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, where X is a positive number. (Throughout the paper, I shall represent games with the positive "scale variable" X. This allows us 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: both players prefer to play either (U,L) or (D,R) rather than not coordinating; but player 1 prefers (U,L) and player 2 prefers (D,R).

[4] See also Gilboa and Schmeidler [1988].

		Player 2	
		L	R
Player 1	U	2X,X	0,0
	D	0 , 0	X,2X

Example 1 -- Battle of the Sexes

The payoffs drawn 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 R, and b) that player 2 believes he is playing D. Then player 1 concludes that player 2 is choosing an action that helps both players (playing L would hurt both players). Because player 2 is not being either generous or mean, neither stylized fact [A] nor [B] apply. Thus, player 1 will be neutral about his effect on player 2, and pursue his material self interest by playing D. If we repeat this argument for player 2, we can show that, in the natural sense, (D,R) 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, (D,R) is a conventional Nash equilibrium in this game. To see the importance of fairness, suppose player 1 believes a) that player 2 will play R, and b) that player 2 believes that he is playing U. Now player 1 concludes that player 2 is lowering her own payoff in order to hurt him. Player 1 will therefore feel hostility towards player 2, and wish to harm her.

If this hostility is strong enough, player 1 may be willing to sacrifice his own material well-being, and play U rather than D. Indeed, if both players have a strong enough emotional reaction to each other's behavior, then (U,R) 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 we 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 (D,R), player 1 strictly prefers playing D to U. In the equilibrium (U,R), player 1 strictly prefers U to D. No matter what payoffs we choose, these statements would be contradictory if payoffs depended solely on the actions taken. To formalize these preferences, therefore, we need to develop a model that explicitly incorporates beliefs. 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 $\pi_i : S_1 \times S_2 \longrightarrow \mathbb{R}$ be player i's *material payoffs.*[5]

[5] I shall emphasize pure strategies in most of the paper, though formal definitions allow for mixed strategies. 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. Another reason I de-emphasize mixed strategies is that they are hard to solve for; Mathematica was used to find the two mixed-strategy equilibria in the Prisoner's Dilemma discussed below.

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: 1) his strategy, 2) what he believes the other player's strategy to be, and 3) what he believes the other player believes his strategy to be. 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.

The first step to incorporating fairness into our analysis is to define a "kindness" function, $f_i(a_i,b_j)$, which measures how kind player i is being to player j.[6]

If player i believes that player j is choosing strategy b_j, how kind is player i being by choosing a_i? Well, 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 define in Appendix A a relatively broad class of kindness functions for which all of the results of this paper are valid.

[6] I assume in this paper that players have a shared notion of kindness and fairness, and that they apply these standards symmetrically. While I believe that this is appropriate for modeling purposes, psychological evidence suggests that people do not all share notions of fairness, and--more importantly--they select notions of fairness with a strong bias towards those that justify pursuing their own material interests. I discuss in Appendix B how multiple kindness functions can be employed.

Let $\pi_j^h(b_j)$ be player j's highest payoff in $\Pi(b_j)$, and let $\pi_j^l(b_j)$ be player j's lowest payoff *among points that are Pareto-efficient in* $\Pi(b_j)$. Let the "equitable payoff" be $\pi_j^e(b_j) = (\pi_j^h(b_j) + \pi_j^l(b_j))/2$. In the case where 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.

<u>Definition</u> <u>1.1</u>:

Player i's kindness to player j is given by

$$f_i(a_i,b_j) \equiv [\pi_j(b_j,a_i) - \pi_j^e(b_j)]/[\pi_j^h(b_j) - \pi_j^{min}(b_j)];$$

if $\pi_j^h(b_j) - \pi_j^{min}(b_j) = 0$, then $f_i(a_i,b_j) = 0$.[7]

Note that $f_i = 0$ if and only if player i is trying to give player j her equitable payoff. 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 non-singleton; otherwise, $\pi_j^e = \pi_j^h$.

[7] 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 $f_i = 0$.

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 separate, this function is formally equivalent to the function $f_j(a_j,b_i)$.

Definition 1.2:

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

$$\tilde{f}_j(b_j,c_i) \equiv [\pi_i(c_i,b_j) - \pi_i^e(c_j)]/[\pi_i^h(c_i) - \pi_i^{min}(c_i)];$$

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

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

These kindness functions can now be used to fully specify 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:

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

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 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, however, I have not precisely determined the relative power of fairness versus material interest, nor even given units for the material payoffs; my results 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.[8]

Because these preferences form a psychological game, we can use the concept *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 beliefs match actual behavior. I shall call the solution concept thus defined *fairness equilibrium.* GPS prove the existence of an equilibrium in all psychological games, which obviously implies that there

[8] 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 we ignore the implication for material payoffs. While this is perhaps correct if people enjoy making suckers of others, it is more likely a player will feel guilty if he is mean to somebody who is nice to him.

always exists a fairness equilibrium.

Definition 2:

The pair of strategies $(a_1,a_2) \in (S_1,S_2)$ is a *Fairness Equilibrium (FE)* if, for $i = 1,2$, $j \neq i$,

1) $a_i \in \text{argmax}_{a \in S_i} U_i(a,b_j,c_i)$, and

2) $c_i = b_i = a_i$.

Is this solution concept consistent with the earlier discussion of Example 1? In particular, is the "hostile" outcome (U,R) a FE? If $c_1 = b_1 = a_1 = U$ and $c_2 = b_2 = a_2 = R$, then $f_2 = -1$. Thus, player 1's utility from playing U is 0 (with $f_1 = -1$) and from playing D it is X-1 (with $f_1 = 0$). Thus, if $X < 1$, player 1 prefers U to D given these beliefs. Player 2 likewise prefers R to L. For $X < 1$, therefore, (U,R) is an equilibrium. In this equilibrium, both players are hostile towards each other, and unwilling to coordinate with the other if it means conceding to the other player.[9]

Because the players will feel no hostility if they coordinate, both (U,L) and (D,R) are also equilibria for all values of X. But, 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 Prisoners' Dilemma shows, by contrast, that fairness may also lead each player to sacrifice to *help* the other player:

[9] For $X < 1/2$, (D,L) is also an equilibrium. In this equilibrium, both players are with common knowledge "conceding", and both players feel hostile towards each other because both are giving up their best possible payoff in order to hurt the other player. The fact that, for $1/2 < X \leq 1$, (U,R) is an equilibrium but (D,L) is not perhaps suggests that (U,R) is "more likely."

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

Example 2 -- Prisoners' Dilemma

Consider the cooperative outcome, (C,C). If it is common knowledge to the players that they are playing (C,C), 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 C, so long as the material gains from defecting are not too large. Thus, if X is small enough (less than 1/18), (C,C) is a fairness equilibrium.

For any value of X, however, the Nash equilibrium (D,D) is also a FE. This is because if it is common knowledge that they are playing (D,D), 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 (D,D), 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, we cannot fully capture realistic behavior by invoking "pure altruism." In Example 2, both (C,C) and (D,D) are FE, 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. Then if he thought that player 2 was playing C, he would play C 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 were playing D, then he would play C 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 C in response to C, he would play C in response to D. 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.[10]

The second issue that the Prisoner's Dilemma illustrates is the role of intentionality in attitudes about fairness, as articulated by stylized fact [D]. Consider Example 3:

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

Example 3 -- Prisoners' Non-Dilemma

This is the Prisoners' Dilemma where player 2 is forced to cooperate. It corresponds, for instance, to a case where somebody is forced to contribute to a public good. In this degenerate game, player 1 will always defect, so the unique FE is (D,C). This contrasts to the possibility of the (C,C) equilibrium in the Prisoners' 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.[11]

[10] Of course, I am ruling out "income effects" and similar stuff as an explanation; but that is not what causes this multiplicity of equilibria in public-goods experiments and elsewhere.

[11] Of course, player 1's complete indifference to player 2's plight here is

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"--illustrates that fairness *can* rule out strict Nash equilibria.[12]

		Player 2	
		D	C
Player 1	D	-2X,-2X	2X,0
	C	0,2X	X,X

Example 4 -- Chicken

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 (D,C) and (C,D)); but both dread most of all the outcome (D,D), in which neither nation backs down.

Consider the Nash equilibrium (D,C), where player 1 "dares" and player 2 "chickens out." Is it a FE? 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 D, thus hurting both player 1 and himself. Thus, for small X, (D,C) is not a FE. Nor, obviously, is (C,D). 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 Prisoners' Dilemma or Battle of the Sexes.

because I have excluded any degree of pure altruism from my model.

[12] 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.

IV. Some Propositions about Fairness Equilibria

In the pure-strategy Nash equilibria of Battle of the Sexes, each player is--taking the other player's strategy as given--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 Prisoners' 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 3.1:

A strategy pair $(a_1,a_2) \in (S_1,S_2)$ is a *Mutual-Max Outcome* if, for $i = 1,2$, $j \neq i$, $a_i \in \text{argmax}_{a \in S_i} \pi_j(a,a_j)$.

Definition 3.2:

A strategy pair $(a_1,a_2) \in (S_1,S_2)$ is a *Mutual-Min Outcome* if, for $i = 1,2$, $j \neq i$, $a_i \in \text{argmin}_{a \in S_i} \pi_j(a,a_j)$.[13]

[13] It is trivial that at least one Mutual-Max and at least one Mutual-Min outcome exists in every game, because we know that a Nash equilibrium exists in every game; A Mutual-Max outcome is simply a Nash equilibrium in a game where each player is trying to maximize the other's material payoff, and a Mutual-Min outcome is simply a Nash equilibrium in which each player is trying to minimize the other player's material payoff.

The following definitions will also prove useful:

Definition 4:

4.1 An outcome is *strictly positive* if, for $i = 1,2$, $f_i > 0$.

4.2 An outcome is *weakly positive* if, for $i = 1,2$, $f_i \geq 0$.

4.3 An outcome is *strictly negative* if, for $i = 1,2$, $f_i < 0$.

4.4 An outcome is *weakly negative* if, for $i = 1,2$, $f_i \leq 0$.

4.5 An outcome is *neutral* if, for $i = 1,2$, $f_i = 0$.

4.6 An outcome is *mixed* if, for $i = 1,2$, $j \neq i$, $f_i \cdot f_j < 0$.

Using these definitions, I state a Proposition about two types of Nash equilibria that will necessarily also be FE:

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 FE.[14]

Proof:

Suppose that (a_1,a_2) is a Mutual-Max outcome. Then both f_1 and f_2 must be non-negative. 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.

Suppose that (a_1,a_2) is a Mutual-Min outcome. Then f_1 and f_2 will both be

[14] Proposition 1 is actually a bit misleading, if interpreted as finding Nash equilibria in which players are being kind towards each other: It follows trivially from definitions that any Mutual-Max Nash equilibrium must be neutral in the sense of Definition 4.5.

non-positive, 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.

Q.E.D.

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 self-destructive, and play D rather than C.

While Proposition 1 characterizes types of Nash equilibria that are necessarily also FE, Proposition 2 characterizes types of outcomes--Nash or non-Nash--that can possibly be FE:

Proposition 2:

Every FE outcome is either strictly positive or weakly negative.

Proof:

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 FE. The only outcomes consistent with FE, therefore, are those for which both f_i and f_j are strictly positive, or neither are. This establishes the proposition.

Q.E.D.

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.[15] 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 the set of games with strategies $S_1 \times S_2$ and, for all $X > 0$, material payoffs $(X\cdot\pi_1(a_1,a_2),X\cdot\pi_2(a_1,a_2))$. 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 (D,D) and (C,C) are FE. 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 FE for X small enough.

[15] While the importance of what happens in games with large material payoffs is clear, I also believe that understanding behavior of games with "small" material payoffs is important. 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. Moreover, I will show in Section V that the *welfare* effects of a situation need not be infinitesimal, even if material payoffs are infinitesimal.

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 $\bar{X}$ such that, for all $X \in (0,\bar{X})$, (a_1,a_2) is a FE in G(X).

Proof:

As $X \longrightarrow 0$, 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 FE. If (a_1,a_2) is a strictly negative Mutual-Max outcome, each player would strictly prefer to play a_i, since this uniquely maximizes the fairness product. Thus, this too would be a FE. Q.E.D.

While Proposition 3 gives sufficient conditions for outcomes to be FE when material payoffs are small, Proposition 4 gives conditions for which outcomes will *not* be FE when material payoffs are small:

Proposition 4:

Suppose that $(a_1,a_2) \in (S_1,S_2)$ is not a Mutual-Max outcome, 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 $\bar{X}$ such that, for all $X \in (0,\bar{X})$, (a_1,a_2) is *not* a FE in G(X).

Proof:

See Appendix D.

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: 1) the game in which each player tries to maximize the other player's material payoffs, and 2) the game in which each player tries to minimize the other player's material payoffs.

There are only 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 doubled-checked. Second, we must also check for whether certain types of Nash equilibria in the original game are also FE, 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 where 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 (S_1,S_2) is a *strict* Nash equilibrium for games in $\mathcal{G}$, then there exists an $\bar{X}$ such that, for all $X > \bar{X}$, (a_1,a_2) is a FE in $G(X)$. If (a_1,a_2) is *not* a Nash equilibrium for games in $\mathcal{G}$, then there exists an $\bar{X}$ such that, for all $X > \bar{X}$, (a_1,a_2) is not a FE in $G(X)$.

Proof:

If (a_1,a_2) is a strict Nash equilibrium, then the difference in material payoffs from playing the equilibrium strategy versus a non-equilibrium 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 a dominated by some other strategy with respect to a_j as X becomes large. Q.E.D.

The only caveat to the set of Nash equilibria being equivalent to the set of FE when payoffs are large is that some non-strict Nash equilibria are not FE.[16]

V. Fairness and Welfare

I consider now some welfare implications of fairness.[17] Consider Example 6. 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 they either both do not grab, or both grab, they each get one can; if one grabs,

[16] This suggests that the definitions of this paper can be used to "refine" Nash equilibrium, by eliminating only those (non-strict) Nash equilibria that are not FE no matter how large are material payoffs.

[17] While it is coherent and sometimes plausible to assume that rational people maximize one "goal utility function" while their well-being corresponds to a different "welfare utility function," I assume here that the two coincide.

and the other does not, then the grabber gets both cans. This is a zero-sum version of the prisoners' dilemma: each player has a dominant strategy, and the unique Nash equilibrium is (grab,grab).

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

Example 6 -- The Grabbing Game

Shopping for minor items is a situation in which people 1) definitely care about material payoffs, and this concern "drives" the nature of the interaction, but they 2) 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, both (grab,grab) and (share,share) are FE when material payoffs are arbitrarily small, but the overall utility in each equilibrium is bounded away from zero.[18] *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 are affected by the observable efforts of others to be friendly or unfriendly.

In Example 6, as with many examples in this paper, there is both a strictly

[18] 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 the social concerns take over when material payoffs are small. A general principle is that, for any game with arbitrarily small material payoffs, every strictly positive FE Pareto-dominates every weakly negative FE.

positive and a strictly negative FE. Are there games that contain only positive, or only negative, FE? 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, we could consider the question of which types of economic structures are likely to generate which types of emotions.

The Prisoners' Dilemma illustrates that there *do* exist situations that endogenously generate hostility. Applying Proposition 5, the only FE of the Prisoners' Dilemma with very large material payoffs is the Nash equilibrium, where both players defect. This FE is strictly negative. Interpreting a negative FE 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 almost surely leave the situation feeling unfriendly.

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 FE yield strictly positive outcomes? Proposition 6 shows that the answer is *No*: there exists in every game a weakly negative FE.

Proposition 6:

In every game, there exists a weakly negative FE.

Proof:

See Appendix D.

Proposition 6 states that it is never guaranteed that people will part with positive feelings.[19] It implies a strong asymmetry in my model of fairness--there is a bias towards 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* your material well-being. Thus, while there are situations in which a player is tempted by material gains 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 only way to be kind is to go *against* your material self-interest.

Of course, in games where there are both positive and negative FE, there may be reasons--such as efficient communication--to expect that the positive equilibria will prevail.

VI. A Model of Monopoly Pricing

One context in which fairness has been studied is monopoly pricing (see, e.g., Thaler [1985] and Kahneman, Knetch, and Thaler [1986a,1986b]). Might consumers see conventional monopoly prices as unfair, and refuse to buy at that price 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 reflected in my model.

I assume that a monopolist has costs c per unit of production, and a

[19] Note, however, that "matching pennies" and other games contain only neutral outcomes, so that people are guaranteed to be emotionally neutral after the play of the game.

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.[20]

Though this is formally an infinite-strategy game, it can be analyze using the model of fairness.[21] Applying Nash equilibrium allows any outcome. We might narrow down the prediction further, however, because the strategy $r = v$ for Consumer weakly dominates all other strategies (this would also be the resulting price of subgame perfection if we made this a sequential game, with Monopolist setting the price first). Thus, if players cared only about material payoffs, the most reasonable outcome from this game is the equilibrium of $p = r = v$, so that the monopolist extracts all the surplus from trade.

What is the highest price consistent with a FE at which this produce could be sold? First, what is the function $f_C(p,r)$, how fair Consumer is being to Monopolist? Given that Monopolist sets p, the only question is whether Monopolist gets profits p-c or profits 0. If $r \geq p$, then Consumer is maximizing both Monopolist's and his own payoffs, so $f_C(p,r) = 0$. If $r < p$, then Consumer is minimizing Monopolist's payoffs, so $f_C(p,r) = -1$. One implication of this is that Monopolist will always exploit its position,

[20] Note that this is essentially the ultimatum game, where the monopolist is the Proposer, and the consumer is the Decider.

[21] 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.

because it will never feel positively towards Consumer; thus, $r > p$ cannot be a FE.

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 Monopolist being to Consumer when $p = r = z$? Calculations show that $f_M(z,z) = [c-z]/2[v-c]$. Because we are considering only values of z between c and v, this number is negative: Anytime the monopolist is not setting a price equal to its costs, the consumer thinks that the monopolist is being unfair. This is because Monopolist is choosing the price that extracts as much surplus as possible from the consumer given the consumer's reservation price.

To see whether $p = r = z$ is a FE for a given z, we must see whether Consumer would wish to deviate by setting $r < z$, thus eliminating Monopolist's profits. Consumer's total utility from $r < z$ is $U_C = 0 + f_M(z,z)\cdot[1+-1] = 0$. Consumer's total utility from sticking with strategy $r = z$ is $U_C = v-z + f_M(z,z)\cdot[1+0] = v-z + [c-z]/2[v-c]$.

Calculations show that the highest price consistent with FE 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 then the conventional monopoly price when fairness is added to the equation. This reflects the arguments of Kahneman, Knetsch, and Thaler [1986a,b]: A monopolist interested in maximizing profits ought not set price at "the monopoly price," because it ought take consumers' attitude towards fairness as a given.

We can further consider some limit results as the stakes become large in this game. Let the monopolist's costs and consumer's value be $C \equiv c\cdot X$ and $V \equiv v\cdot X$. We can 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 fairness's sake is more costly for the consumer.

VII. Discussion and Conclusion

The notion of fairness in this paper captures several important regularities of behavior, but leaves out other issues.

As an example, problems arise because the definition of fairness I use is very "local": in judging each other's fairness, players consider only each other's willingness to resist unilateral deviations, rather than taking into account possible outcomes in which both players change their strategies. For instance, in the battle of the sexes a more "global" notion of fairness would allow a player to have different emotions in his preferred efficient equilibrium than in his less-preferred efficient equilibrium; my definition makes no distinction. The "local" notion of fairness also allows a sort of paradox: players may feel more positive towards each other with one outcome than with an alternative outcome that gives them both higher material payoffs.[22] While I feel this focus on "local" fairness is often valid (and very much in the spirit of non-cooperative game theory more generally), further research could consider how players incorporate broader aspects of a game into their emotions.

Future research can also focus on modeling additional emotions. In Example 7, for instance, my model predicts no cooperation, whereas it seems plausible

[22] Drew Fudenberg pointed this out. Formally, a strictly positive, Mutual-Max outcome can be Pareto-dominated by a neutral Nash equilibrium.

that cooperation would take place.[23]

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

Example 7 -- Leaving a Partnership

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, the 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.

We might think that (Trust,Share) *is* a reasonable outcome. This would be

[23] A related example was first pointed out to me by Jim Fearon.

the outcome, for instance, if we assumed that players wish to be kind to those that trust them: If player 1 plays "Trust" rather than "Split", he is showing he trusts player 2. If player 2 feels kindly towards player 1 as a result of this trust, then he might not grab all the profits. If we concluded that the idea that people are motivated to reward was psychologically sound, we could incorporate it into formal models.

Even if we wanted to keep the basic theory as is, 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 multi-person games is likely to be how a person behaves when he is hostile to some players, but friendly towards 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 to either help everybody or to hurt everybody. This, for instance, would be the case when choosing the contribution level to a public good. Do you contribute to reward those who have contributed, or not contribute to punish those who have not contributed?

Extending the model to Bayesian games is likely to be 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 will enter dramatically into decision-making. This is similar to extending the model to the multi-person case; instead of facing a known number of kind and unkind people, a player faces probabilities that a given player is unkind or unkind.

Extending the model to sequential games is also likely to be 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 analog to Proposition 6 might no longer be true, and sequential games could perhaps be used as mechanisms that guarantee positive emotions.

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 1) fairness cannot lead to infinitely positive or infinitely negative utility, and 2) 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:

1) There exists a number N such that $f_i(a_i,b_j) \in [-N,N]$ for all (a_i,b_j); and

2) $f_i(a_i,b_j) > f_i(a_i',b_j)$ iff $\pi_i(a_i,b_j) > \pi_i(a_i',b_j)$.

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

Definition A2:

Consider $\Pi(b_j)$, $\pi_j^h(b_j)$, and $\pi_j^l(b_j)$ 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:

1) $\pi_j(b_j,a_i) > \pi_j(b_j^e)$ implies that $f_i(a_i,b_j) > 0$; and

$\pi_j(b_j,a_i) = \pi_j^e(b_j)$ implies that $f_i(a_i,b_j) = 0$; and

$\pi_j(b_j,a_i) < \pi_j^e(b_j)$ implies that $f_i(a_i,b_j) < 0$.

2) $\pi_j^h(b_j) \geq \pi_j^e(b_j) \geq \pi_j^l(b_j)$

3) If $\pi_j^h(b_j) > \pi_j^l(b_j)$, then $\pi_j^h(b_j) > \pi_j^e(b_j) > \pi_j^l(b_j)$

The reader can verify that 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, we 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 we double all of player 2's payoffs, 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 we comparably change *both* players' payoffs.

Appendix B: Players Can Have Different Notions of Kindness

In the paper, I assumed that players share a notion of fairness, and that they apply this notion of fairness to themselves and each other. Yet people sometimes choose self-serving notions of fairness; they may also in good faith disagree about standards of fairness. Can the lessons of this paper be extended to such situations?

I believe the answer is, to a limited extent, yes. Suppose, for instance, that we allowed each of f_i, $\tilde{f}_j$, f_j, and $\tilde{f}_i$ to have different functional forms, so long as they all meet Definitions A1, A2, and A3. Then all propositions of the paper would hold.

One natural way to incorporate the "self-serving" type of fairness may be to assume that there are two natural fairness functions, f_i and g_i, from which a player chooses the one that is most convenient for him in terms of what can yield him the larger utility. That is,

$$U_i(a_i,b_j,c_i) \equiv \text{Max}\ \{\pi_i(a_i,b_j) + \tilde{f}_j(b_j,c_i)\cdot[1+f_i(a_i,b_j)],\\ \pi_i(a_i,b_j) + \tilde{g}_j(b_j,c_i)\cdot[1+g_i(a_i,b_j)]\}$$

<u>Appendix</u> <u>C</u>: <u>The</u> <u>Utility</u> <u>Function</u> <u>Can</u> <u>Be</u> <u>Generalized</u>

The precise way I specify the utility function is limited in many ways. 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. Evidence suggests that, while people are substantially motivated by the type of "contingent altruism" I have incorporated into the model, pure altruism is also sometimes a motive.

We could readily expand the utility function to incorporate pure altruism:

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

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. (Moreover, note that if $\alpha = 1$, then this utility function will no longer lead to a psychological game, because second-order beliefs would no longer be relevant.)

Another unrealistic feature of the utility function is the linear separation of material payoffs from fairness payoffs. Furthermore, the fairness utility is independent of the scale of the material payoffs. Consider a situation in which a Proposer has an offer to split $1 evenly 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 evenly split $1 million. This seems unrealistic--the bitterness he feels should be larger the greater the harm done.

We 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.[24]

[24] This specification, and one of the conditions mentioned below to maintain the limit results, were suggested by Roland Benabou.

Of course, this might create problems for the limit results of the paper. However, the conditions that 1) $G(X)/X \longrightarrow 0$ as $X \longrightarrow \infty$ and 2) $G(X)$ bounded away from 0 as $X \longrightarrow 0$ would suffice for all propositions to hold. These conditions simply allow for a generalization of stylized fact [C].

Appendix D: Proofs of Propositions 4 and 6

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 $f_2 > 0$. But player 2 will choose $f_2 > 0$ only if $f_1 > 0$. Thus, both $f_1, f_2 > 0$. But if the material payoffs are small, this means 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, 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 FE 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, and $f_1 < 0$. But 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. But this condition on player 2's choices directly contradicts the assumption that she *could* lower player 1's payoffs.

This establishes the Proposition. Q.E.D.

Proof of Proposition 6:

From the material game, consider the psychological game from the preferences $V_i \equiv \pi_i(a_i,b_j) + \text{Min}[f_j(c_i,b_j),0]\cdot\text{Min}[f_i(a_i,b_j),0]$. By GPS's general existence result, this game has at least one equilibrium, (a_1^*,a_2^*). I will now argue that any such equilibrium is also a FE.

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 $a_i' \in \text{argmax}_{a\in S_i}\pi_i(a,a_j^*)$. 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, $f_i(a_1^*,a_2^*) \leq 0$. But this implies that each player maximizing $V_i(a_i,a_j^*,a_i^*)$ is the same as his maximizing $U_i(a_i,a_j^*,a_i^*)$. Thus, (a_i^*,a_j^*) is a FE. Q.E.D.

REFERENCES

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

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

Andreoni, James, "Why Free Ride? Strategies and Learning in Public Goods Experiments," Journal of Public Economics 37, 291-304, 1988b.

Batson, Daniel C., "Prosocial Motivation: Is it Ever Truly Altruistic?" Advances in Experimental Psychology 75, 73-98, 1982.

Bergstrom, Theodore, Lawrence Blume and Hal Varian, "On the Private Provision of Public Goods," Journal of Public Economics 29, 25-49, 1986.

Bercheid, Ellen, David Boye, and Elaine Walster, "Retaliation as a Means of Restoring Equity," Journal of Personality and Social Psychology 10, 370-376.

Binmore, Ken, Avner Shaked, and John Sutton, "Testing Noncooperative Bargaining Theory: A preliminary Study," American Economic Review 75, 1178-1180, 1985.

Dawes, Robyn M. and Richard H. Thaler, "Anomalies: Cooperation," Journal of Economic Perspectives 2, 187-198, Summer 1988.

Finn, R.H., and Sang Lee, "Salary Equity: Its Determination, Analysis and Correlates," Journal of Public Economics 29, 25-49, 1986.

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

Gilboa, Itzhak and David Schmeidler, "Information Dependent Games: Can Common Sense be Common Knowledge?", Economic Letters 27, 215-221, 1988.

Goetze, David and John M. Orbell, "Understanding and Cooperation," Public Choice

Goranson, Richard E., and Leonard Berkowitz, "Reciprocity and Responsibility Reactions To Prior Help," Journal of Personality and Social Psychology 3, 227-232, 1966.

Greenberg, Jerald, "Effects of Reward Value and Retaliative Power on Allocation Decisions: Justice, Generosity, of Greed?" Journal of Personality and Social Psychology 36, 367-379, 1978.

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

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze, "An Experimental Analysis of Ultimatum Bargaining," Journal of Economic Behavior and Organization 3, 367-388, 1982.

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

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

Isaac, R. Mark and James Walker, "Group Size Effects in Public Goods Provision: The Voluntary Contribution Mechanism," Quarterly Journal of Economics

Isaac, Mark, and James Walker, "Communication and Free-Riding Behavior: The Voluntary Contribution Mechanism," Economic Inquiry 26, 585-608, 1988b.

Isaac, R. Mark, James M. Walker, and Susan H. Thomas, "Divergent Evidence on Free Riding: An Experimental Examination of Possible Explanations," Public Choice 43, 113-149, 1984.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, "Fairness as a Constraint on Profit Seeking: Entitlements in the Market," American Economic Review 76, 728-741, 1986.

Kahneman, Daniel, Jack L. Knetsch, and Richard H. Thaler, "Fairness and the Assumptions of Economics," Journal of Business 59, S285-S300, 1986.

Kim, Oliver and Mark Walker, "The Free Rider Problem: Experimental Evidence," Public Choice 43, 3-24, 1984.

Krebs, Dennis L., "Altruism--An Examination of the Concept and a Review of the Literature," Psychological Bulletin 73, 258-302, 1970.

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

Marwell, Gerald and Ruth Ames, "Economists Free Ride, Does Anyone Else?" Journal of Public Economics 15, 295-310, 1981.

Orbell, John M., Robyn M. Dawes, and Alphons J. C. van de Kragt, "Explaining Discussion Induced Cooperation," Journal of Personality and Social Psychology

Thaler, Richard, "Toward a Positive Theory of Consumer Choice," Journal of Economic Behavior and Organization 1, 39-60, 1980.

Thaler, Richard, "Mental Accounting and Consumer Choice," Marketing Science 4, 199-214, Summer 1985.

Thaler, Richard H., "Anomalies: The Ultimatum Game," Journal of Economic Perspectives 2, 195-207, Fall 1988.

Train, Kenneth E., Daniel L. McFadden, and Andrew A. Goett, "Consumer Attitudes and Voluntary Rate Schedules For Public Utilities," The Review of Economics and Statistics 64, 383-391, August 1987.

van de Kragt, Alphons J. C., John M. Orbell, and Robyn M. Dawes, "The Minimal Contributing Set as a Solution to Public Goods Problems," American Political Science Review 77, 112-122, 1983.

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

December 5, 1991

Working Paper Series
Department of Economics
University of California, Berkeley

Individual copies are available for $3.50 in the USA or Canada, $6.00 to Europe/South America, $7.00 to Japan/Middle East. Papers may be obtained from the Institute of Business and Economic Research: send requests to IBER, 156 Barrows Hall, University of California, Berkeley CA 94720. Prepayment is required. Make checks or money orders payable to "The Regents of the University of California."

90-149 "The 1933 World Economic Conference as an Instance of Failed International Cooperation." Barry Eichengreen and Marc Uzan. October 1990.

90-150 "Costs and Benefits of European Monetary Unification." Barry Eichengreen. October 1990.

90-151 "Is Europe an Optimum Currency Area?" Barry Eichengreen. October 1990.

90-152 "Major Fiscal Trends in the 1980s and Implications for the 1990s." George Break. October 1990.

90-153 "Historical Research on International Lending and Debt." Barry Eichengreen. December 1990.

91-154 "Risktaking, Capital Markets, and Market Socialism." Pranab Bardhan. January 1991.

91-155 "Is Inequality Harmful for Growth? Theory and Evidence." Torsten Persson and Guido Tabellini. January 1991.

91-156 "The Origins and Nature of the Great Slump, Revisited." Barry Eichengreen. March 1991.

91-157 "The Making of Exchange Rate Policy in the 1980s." Jeffrey Frankel. March 1991.

91-158 "Exchange Rate Forecasting Techniques, Survey Data, and Implications for the Foreign Exchange Market." Jeffrey Frankel and Kenneth Froot. March 1991.

91-159 "Convertibility and the Czech Crown." Jeffrey Frankel. March 1991.

91-160 "The Obstacles to Macroeconomic Policy Coordination in the 1990s and an Analysis of International Nominal Targeting (INT)." Jeffrey A. Frankel. March 1991.

91-161 "Highway Safety, Economic Behavior, and Driving Environment." Theodore E. Keeler. March 1991.

91-162 "Can Informal Cooperation Stabilize Exchange Rates? Evidence from the 1936 Tripartite Agreement." Barry Eichengreen and Caroline R. James. March 1991.

91-163 "Reneging and Renegotiation." Matthew Rabin. April 1991.

91-164 "A Model of Pre-game Communication." Matthew Rabin. April 1991.

91-165 "Contracting Between Sophisticated Parties: A More Complete View of Incomplete Contracts and Their Breach." Benjamin E. Hermalin and Michael L. Katz. May 1991.

91-166 "The Stabilizing Properties of a Nominal GNP Rule in an Open Economy." Jeffrey A. Frankel and Menzie Chinn. May 1991.

91-167 "A Note on Internationally Coordinated Policy Packages Intended to Be Robust Under Model Uncertainty or Policy Cooperation Under Uncertainty: The Case for Some Disappointment." Jeffrey A. Frankel. May 1991.

91-168 "Managerial Preferences Concerning Risky Projects." Benjamin Hermalin. June 1991.

91-169 "Information and the Control of Productive Assets." Matthew Rabin. July 1991.

91-170 "Rational Bubbles: A Test." Roger Craine. July 1991.

91-171 "The Eternal Fiscal Question: Free Trade and Protection in Britain, 1860-1929." Barry Eichengreen. July 1991.

91-172 "Game-Playing Agents: Unobservable Contracts as Precommitments." Michael L. Katz. July 1991.

91-173 "Taxation, Regulation, and Addiction: A Demand Function for Cigarettes Based on Time-Series Evidence." Theodore E. Keeler, Teh-wei Hu, and Paul G. Barnett. July 1991

91-174 "The Impact of a Large Tax Increase on Cigarette Consumption: The Case of California." Teh-wei Hu, Jushan Bai, Theodore E. Keeler and Paul G. Barnett. July 1991.

91-175 "Market Socialism: A Case for Rejuvenation." Pranab Bardhan and John E. Roemer. July 1991.

91-176 "Designing A Central Bank For Europe: A Cautionary Tale from the Early Years of the Federal Reserve." Barry Eichengreen. July 1991.

91-177 "Restructuring Centrally-Planned Economies: The Case of China in the Long Term." John M. Letiche. September 1991.

91-178 "Willingness to Pay for the Quality and Intensity of Medical Care: Evidence from Low Income Households in Ghana." Victor Lavy and John M. Quigley. September 1991.

91-179 "Focal Points in Pre-Game Communication." Matthew Rabin. September 1991.

91-180 "Cognitive Dissonance and Social Change." Matthew Rabin. September 1991.

91-181 "European Monetary Unification and the Regional Unemployment Problem." Barry Eichengreen. October 1991.

91-182 "The Effects of Competition on Executive Behavior." Benjamin E. Hermalin. October 1991.

91-183 "The Use of an Agent in a Signalling Model." Bernard Caillaud and Benjamin Hermalin. October 1991.

91-184 "The Marshall Plan: History's Most Successful Structural Adjustment Program." J. Bradford De Long and Barry Eichengreen. November 1991.

91-185 "Incorporating Fairness into Game Theory." Matthew Rabin. December 1991.