UNIVERSITY OF CALIFORNIA AT BERKELEY

Department of Economics

Berkeley, California 94720

Working Paper No. 92-199

Incorporating Fairness Into Game Theory and Economics

Matthew Rabin

Economics Department
University of California at Berkeley

July 1992

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

JEL Classification: A12, A13, D63, C70

This paper is a revised version of "Incorporating Fairness into Game Theory," U.C. Berkeley Economics Working Paper No. 91-185, December 1991. I thank Gadi Barlevy, Nikki Blasberg, Jeff Ely, and April Franco for research assistance; George Akerlof, Colin Camerer, Joe Farrell, Jonathan Feinstein, Ruth Given, Danny Kahneman, Annette Montgomery, Richard Thaler, and especially Bill Dickens for useful conversations; and Gadi Barlevy, Roland Benabou, Nikki Blasberg, Gary Bolton, Jeff Ely, April Franco, Drew Fudenberg, Rachel Kranton, Michael Lii, James Montgomery, Jim Ratliff, Halsey Rogers, Lones Smith, Jeff Zweibel, and especially Eddie Dekel-Tabak, Jim Fearon, Berkeley's David Levine, and Vai-Lam Mui for helpful comments on earlier drafts of this paper. I also thank the Institute of Business and Economic Research at the University of California 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 to models of monopoly pricing and labor economics. I argue that the welfare implications of fairness can be large, both because concern for fairness affects behavior, and because it changes a person's utility for a given outcome.

Applying the model shows the special role of "Mutual-Max" outcomes – in which each person maximizes the other's material payoffs – and "Mutual-Min" outcomes – in which each person 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, roughly, an outcome is a fairness equilibrium if 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 evidence indicates that most altruistic behavior is more complex: people do not seek uniformly to help other people; rather, they do so according to how generous these other people are being. Indeed, *the same people who are 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 have economic implications. If an employee has been exceptionally loyal, then a manager may feel some obligation to treat that employee well, even when it is not in his self-interest to do so. Other examples of economic behavior induced by social goals are voluntary reductions of water-use during droughts, conservation of energy to help solve the energy crisis (as documented, for instance, in 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. An employee who feels she has been mistreated by a firm may engage in acts of sabotage. Members of a striking labor union may strike longer than is in their material interests because they want to punish a firm for being unfair.

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

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

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

[C] Both motivations [A] and [B] have a greater effect on behavior as the material cost of sacrificing becomes smaller.

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

In the next section, I briefly present some of the evidence from the psychological literature regarding these stylized facts. In Section III, I develop a game-theoretic solution concept "fairness equilibrium" that incorporates these stylized facts. Fairness equilibria do not in general constitute either a subset or a superset of Nash equilibria; that is, incorporating fairness considerations can both add new predictions to economic models and eliminate conventional predictions. In Section IV, I present some general results about which outcomes in economic situations are likely to be fairness equilibria. The results demonstrate the special role of "Mutual-Max" outcomes--in which, given the other person's behavior, each person maximizes the other's material payoffs--and "Mutual-Min" outcomes--in which, given the other person's behavior, each person minimizes the other's material payoffs. The following results hold: Any Nash equilibrium that is either a Mutual-Max outcome or Mutual-Min outcome is also fairness equilibrium. If material payoffs are small, then, roughly, an outcome is a fairness equilibrium if and only if it is a Mutual-Max or a Mutual-Min outcome. If material payoffs are large, then, roughly, an outcome is a fairness equilibrium if and only if it is a Nash equilibrium.

I hope this framework will eventually be used to study the implications of fairness in different economic situations. While I do not develop extended applications in this paper, Sections V and VI contain examples illustrating the economic implications of my model of fairness. In Section V, I develop a simple model of monopoly pricing, and show that fairness implies that goods can only be sold at below the classical monopoly level. In Section VI, I explore the implications of fairness in several examples from labor economics.

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

While my model suggests that the behavioral implications of fairness are greatest when the material consequences of an economic interaction are not too large, there are several reasons why this does not imply that the economic implications of fairness are minor. First, while it is true that fairness influences behavior *most* when material stakes are small, it is not clear that it makes little difference when material stakes are large. Little empirical research on the economic implications of fairness has been conducted, and much anecdotal evidence suggests that people sacrifice substantial amounts of money to reward or punish kind or unkind behavior.

Second, many major economic institutions--most notably decentralized markets--are best described as accumulations of minor economic interactions, so that the aggregate implications of departures from standard theory in these cases may be substantial. Third, the fairness component of a person's overall well-being can be influenced substantially by even small material changes.

Finally, even if material incentives in a situation are so large as to dominate behavior, fairness *still* matters. Welfare economics should be concerned not only with the efficient allocation of material goods, but also with designing institutions such that people are happy about the way they interact with others. For instance, if a person leaves an exchange in which he was treated unkindly, then his unhappiness at being so treated should be a consideration in evaluating the efficiency of that exchange. 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.[2]

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

II. Fairness in Games: Some Evidence

In this section, I discuss some psychological research that demonstrates the stylized facts outlined in the introduction. Consider [A]--"People are willing to sacrifice their own material well-being to help those who are being kind." The attempt to provide public goods without coercion is an archetypical example 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 self-interested contribution is small or zero. The evidence from these experiments is that

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

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.[3]

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

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, even in one-shot settings, Deciders are willing to punish unfair offers by rejecting them, and that Proposers tend to make fair offers.[4]

Some papers illustrating Stylized Fact [B] are Kahneman, Knetsch, and

[3] 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].

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

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 majority of Deciders would *accept* a proposed split of ($9 million, $1 million).[5] 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.

From an economist's point of view, it matters not only whether stylized facts [A] to [C] are true, but whether they have important economic implications. Kahneman, Knetsch, and Thaler [1986a,1986b] present strong arguments that these general issues are indeed important. For any unconvinced of the importance of social goals empirically or intuitively, one purpose of

[5] 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.

this paper is to help us 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?

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.[6] 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 developing my model of fairness, however, I extend the GPS approach with an additional step which I believe will prove essential for incorporating psychology into economic research: I *derive* psychological games from basic "material games." Whereas GPS provide a technique for analyzing games that already incorporate emotions 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

[6] See also Gilboa and Schmeidler [1988]. Outside the context of non-cooperative game theory, Akerlof and Dickens [1982] presented an earlier model incorporating beliefs directly into people's utility functions.

consider the effects of increasing or decreasing a game's stakes without changing its fundamental strategic structure.) This is a standard battle-of-the-sexes game: two people prefer to go to the same event together, but each prefers a different event. Formally, both players prefer to play either (Opera,Opera) or (Boxing,Boxing)rather than not coordinating; but player 1 prefers (Opera,Opera) and player 2 prefers (Boxing,Boxing).

		Player 2	
		Opera	Boxing
Player 1	Opera	2X,X	0,0
	Boxing	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 Boxing, and b) that player 2 believes player 1 is playing Boxing. Then player 1 concludes that player 2 is choosing an action that helps both players (playing Opera would hurt both players). Because player 2 is not being either generous or mean, neither stylized fact [A] nor [B] apply. Thus, player 1 will be neutral about his effect on player 2, and pursue his material self interest by playing Boxing. If we repeat this argument for player 2, we can show that, in the natural sense, (Boxing,Boxing) is an equilibrium: if it is common knowledge that this will be the outcome, then each player is maximizing his utility by playing his strategy.

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

Notice the central role of expectations: Player 1's payoffs do not depend simply on the actions taken, but also on his beliefs about player 2's *motives*. Could these emotions be directly modeled by transforming the payoffs, so that 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 (Boxing,Boxing), player 1 strictly prefers playing Boxing to Opera. In the equilibrium (Opera,Boxing) player 1 strictly prefers Opera to Boxing. No matter what payoffs 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.[7] I now construct such a model, applicable to

[7] My point here is that the results I get could not be gotten simply by

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.*[8]

From this "material game," I now construct a "psychological game" as defined in GPS. I assume that each player's subjective expected utility when he chooses his strategy will depend on three factors: 1) his strategy, 2) his beliefs about the other player's strategy choice, and 3) his beliefs about the other player's beliefs about his strategy. Throughout, I shall use the following notation: $a_1 \in S_1$ and $a_2 \in S_2$ represent the strategies chosen by the two players; $b_1 \in S_1$ and $b_2 \in S_2$ represent, respectively, player 2's beliefs about what strategy player 1 is choosing, and player 1's beliefs about what strategy player 2 is choosing; $c_1 \in S_1$ and $c_2 \in S_2$ represent player 1's beliefs about what player 2 believes player 1's strategy is, and player 2's beliefs about what player 1 believes player 2's strategy is.

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.[9]

respecifying the payoffs over the physical actions in the game. Kolpin [forthcoming] argues that we can apply conventional game theory to these games by including the choice of beliefs as additional parts of players' strategies.

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

[9] 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

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.

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

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

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

importantly--they select notions of fairness that tend to justify pursuing their own material interests. I discuss in Appendix B how multiple kindness functions can be employed.

Note that $f_i = 0$ if and only if player i is trying to give player j her equitable payoff.[10] 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, $\pi_j^e = \pi_j^h$.

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

Definition 1.2:

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

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

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

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'

[10] 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 $f_i = 0$.

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 an action a_i such that $f_i(\cdot)$ is low or negative. If player j is treating player i kindly, then $\tilde{f}_j(\cdot)$ will be positive, and player i will wish to treat player j kindly. Of course, the specified utility function is such that players will trade off their preference for fairness against their material well-being, and material pursuits may override concerns for fairness.

Because the kindness functions are bounded above and below, this utility function reflects stylized fact [C]: the bigger the material payoffs, the less the players' behavior reflects their concern for fairness. Thus, the behavior in these games is sensitive to the scale of material payoffs. Obviously, I have not precisely determined the relative power of fairness versus material interest, nor even given units for the material payoffs; my results in specific examples are, therefore, only qualitative.

Notice that the preferences $V_i(a_i,b_j,c_i) \equiv \pi_i(a_i,b_j)+\tilde{f}_j(b_j,c_i)\cdot f_i(a_i,b_j)$ would yield precisely the same behavior as the utility function $U_i(a_i,b_j,c_i)$. I have made the preferences slightly more complicated so as to capture one bit of realism: whenever player j is treating player i unkindly, player i's overall utility will be lower than his material payoffs. That is, $\tilde{f}_j(\cdot) < 0$ implies $U_i(\cdot) \leq \pi_i(\cdot)$. If a person is treated badly, he leaves the situation bitter, and his ability to take revenge only partly makes up for the loss in welfare.[11]

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 implies that there always exists a fairness equilibrium.

Definition 2:

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

1) $a_i \in \text{argmax}_{a \in S_i} U_i(a,b_j,c_i)$, 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 (Opera,Boxing) a fairness equilibrium? If $c_1 = b_1 = a_1 =$ Opera and $c_2 = b_2 = a_2 =$ Boxing, then player 2 feels hostility, and $f_2 = -1$. Thus, player 1's utility from playing U is 0 (with $f_1 = -1$) and from playing Boxing it is X-1 (with $f_1 = 0$). Thus, if $X < 1$, player 1 prefers Opera to Boxing given these beliefs. Player 2 prefers Boxing to Opera. For $X < 1$, therefore, (Opera,Boxing) is an equilibrium. In

[11] 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.

this equilibrium, both players are hostile towards each other, and unwilling to coordinate with the other if it means conceding to the other player.[12]

Because the players will feel no hostility if they coordinate, both (Opera,Opera) and (Boxing,Boxing) are also equilibria for all values of X. 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:

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

Example 2 -- Prisoners' Dilemma

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

[12] For $X < 1/2$, (Boxing,Opera) 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 $1/2 < X \leq 1$--(Opera,Boxing) is an equilibrium, but (Boxing,Opera) is not, might suggest that (Opera,Boxing) is "more likely."

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

The Prisoner's Dilemma illustrates two issues I discussed earlier. First, we cannot fully capture realistic behavior by invoking "pure altruism." In Example 2, both (Cooperate,Cooperate) and (Defect,Defect) are fairness equilibria, and I believe this prediction of the model is in line with reality. People sometimes cooperate, but if each expects the other player to defect, then they both will. Yet, having both of these as equilibria is inconsistent with pure altruism. Suppose that player 1's concern for player 2 were independent of player 2's behavior. Then if he thought that player 2 was playing Cooperate, he would play Cooperate if and only if he were willing to give up 2X in order to help player 2 by 4X; if player 1 thought that player 2 were playing Defect, then he would play Cooperate if and only if he were willing to give up X in order to help player 2 by 5X. Clearly, then, if player 1 plays Cooperate in response to Cooperate, he would play Cooperate in response to Defect. In order to get the two equilibria, player 1 *must* care differentially about helping (or hurting) player 2 as a function of player 2's behavior.[13]

The second issue that the Prisoner's Dilemma illustrates is the role of intentionality in attitudes about fairness. Namely, psychological evidence indicates that people determine the fairness of others according to their

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

motives, not solely according to actions taken.[14] In game-theoretic terms, "motives" can be inferred from a player's choice of strategy among those choices he has, so that it can be as important what strategy a player *could* have chosen but didn't as with what strategy he actually did choose. For example, people differentiate between those who take a generous action by choice and those who are forced to do so. 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 fairness equilibrium is (Defect,Cooperate). This contrasts to the possibility of the (Cooperate,Cooperate) 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.[15]

In both Examples 1 and 2, adding fairness creates new equilibria, but

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

[15] Player 1's complete indifference to player 2's plight here is because I have excluded any degree of pure altruism from my model.

does not get rid of any (strict) Nash equilibria. Example 4--the game "Chicken"--illustrates that fairness *can* rule out strict Nash equilibria.

		Player 2	
		Dare	Chicken
Player 1	Dare	-2X, -2X	2X, 0
	Chicken	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 (Dare,Chicken) and (Chicken,Dare)); but both dread most of all the outcome (Dare,Dare), in which neither nation backs down.

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

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

IV. Some General 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 $i = 1,2$, $j \neq i$, $a_i \in \mathrm{argmax}_{a \in S_i}\ \pi_j(a,a_j)$.

Definition 3.2:

A strategy pair $(a_1,a_2) \in (S_1,S_2)$ is a *Mutual-Min Outcome* if, for $i = 1,2$, $j \neq i$, $a_i \in \mathrm{argmin}_{a \in S_i}\ \pi_j(a,a_j)$.[16]

[16] It is immediate 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. Each of these definitions characterizes an outcome of a game in terms of the value of "kindness" f_i induced by each of the players.

Definition 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 ≠ 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 fairness equilibria:

Proposition 1[17]:

Suppose that (a_1,a_2) is a Nash equilibrium, and either a Mutual-Max outcome or a Mutual-Min outcome. Then (a_1,a_2) is a fairness equilibrium.

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

While Proposition 1 characterizes Nash equilibria that are necessarily

[17] All proofs are in Appendix D.

fairness equilibria, Proposition 2 characterizes which outcomes--Nash or non-Nash--can possibly be fairness equilibria:

Proposition 2:

Every fairness equilibrium outcome is either strictly positive or weakly negative.

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

While Propositions 1 and 2 pertain to all games--irrespective of the scale of material payoffs--I present in the remainder of this section several results that hold when material payoffs are either arbitrarily large or arbitrarily small. To do so, I will consider classes of games that differ only in the scale of the material payoffs. Given the set of strategies $S_1 \times S_2$, and the payoff functions $(\pi_1(a_1,a_2),\pi_2(a_1,a_2))$, let $\mathcal{G}$ be the 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 (Dare,Dare) and (Chicken,Chicken) are fairness equilibria. Note that, while these two outcomes are (respectively) Mutual-Min and Mutual-Max outcomes, they are *not* Nash equilibria. Yet, when X is small, the fact that they are not equilibria in the "material" game is unimportant, because fairness considerations will start to dominate. Proposition 3 shows that the class of "strict" Mutual-Max and Mutual-Min outcomes are fairness equilibria for X small enough.

Proposition 3:

For any outcome (a_1,a_2) that is either a strictly positive Mutual-Max outcome or a strictly negative Mutual-Min outcome, there exists an $\bar{X}$ such that, for all $X \in (0,\bar{X})$, (a_1,a_2) is a fairness equilibrium in $G(X)$.

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

Proposition 4:

Suppose that $(a_1,a_2) \in (S_1,S_2)$ is not a Mutual-Max 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 fairness equilibrium in $G(X)$.

Together, Propositions 3 and 4 state that, for games with very small material payoffs, finding the fairness equilibria consists *approximately* of finding the *Nash* equilibria in *each* of the following two hypothetical games: 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 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 fairness equilibria, even though they are neither Mutual-Max nor Mutual-Min outcomes. The potentially problematic Nash equilibria are those in which one of the players has no options that will lower the other's material payoffs.

I now turn to the case 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 (a_1,a_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 fairness equilibrium 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 fairness equilibrium in $G(X)$.

The only caveat to the set of Nash equilibria being equivalent to the set of fairness equilibria when payoffs are large is that some non-strict Nash equilibria are not fairness equilibria.[18]

[18] 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 fairness equilibria no matter how large are material payoffs.

V. Application to 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 an implication of fairness equilibrium.

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

Though this is formally an infinite-strategy game, it can be analyzed using my model of fairness.[19] Applying Nash equilibrium allows any outcome. We might, however, further narrow our prediction, because the strategy $r = v$ for Consumer weakly dominates all other strategies (this would also be the result 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 where $p = r = v$,

[19] 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.

so that the monopolist extracts all the surplus from trade.

What is the highest price consistent with a fairness equilibrium at which this product could be sold? First, what is the function $f_C(r,p)$, how fair 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 $f_C(r,p) = 0$. If $r < p$, then Consumer is minimizing Monopolist's payoffs, so $f_C(r,p) = -1$. One implication of this is that Monopolist will always exploit its position, because it will never feel positively towards Consumer; thus, $r > p$ cannot be a fairness equilibrium.

Because $r < p$ leads to no trade, this means that the only possibility for an equilibrium with trade is when $p = r$. How fair is 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 the monopolist is choosing the price that extracts as much surplus as possible from the consumer--given the consumer's refusal to buy at a price higher than z.

To see whether $p = r = z$ is a fairness equilibrium for a given z, 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 fairness equilibrium is given by $z^* = [2v^2 - 2cv + c] / [1 + 2v - 2c]$. This number is strictly less than v when $v > c$. Thus, the highest equilibrium price possible is lower 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.

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

VI. Applications to Labor Economics

In this section I discuss several applications of fairness to labor economics. While these examples are not very detailed, I believe they suggest the potential of this model in formalizing issues of fairness that many authors have argued are important in labor economics.[20]

[20] For some examples discussing the role in labor economics of fairness and

I begin with an extended example that resembles the "gift exchange" view of the employment relationship discussed in Akerlof [1982]. Consider the situation where a worker chooses an effort level, and the firm simultaneously chooses a benefit level for the worker.[21] Formally, Worker chooses either a high or low effort level: $e \in \{W,L\}$. If $e = H$, Firm receives revenue $R > 0$, and Worker receives disutility γ. If $e = L$, Firm receives no revenue, and Worker experiences no disutility. Simultaneously, Firm chooses a benefit level $b \in [0,R]$. Material payoffs are as follows[22]:

$$\begin{array}{lll} \pi_W & = & b^{1/2} - \gamma \text{ if } e = H \\ & & b^{1/2} \qquad \text{if } e = L \\ \pi_F & = & (R-b)^{1/2} \text{ if } e = H \text{ and } b \leq R \\ & & 0 \qquad\quad \text{if } e = L \text{ or } b > R, \end{array}$$

where π_W is Worker's material payoffs, and π_F is Firm's material payoffs.

This situation is essentially a continuous-strategy Prisoners' Dilemma, because each player has a dominant strategy--Worker maximizes his material payoffs by choosing $e = L$, and Firm maximizes its material payoffs by choosing

related issues, see Akerlof [1982], Baron [1988], Bishop [1987], Finn and Lee [1986], Levine [1991a,1991b], and Rotemberg [1992].

[21] This model is a version of one suggested to me by James Montgomery. Several of the models in this section are more naturally modeled as sequential-move games, rather than the simultaneous-move games. Moreover, most industrial-relations issues clearly involve repeated interactions among the parties, with a corresponding evolution of attitudes by the parties. While I believe that much of the intuition of these examples would carry over, I also think it is important to extend the model of fairness presented in this paper to better incorporate such dynamic issues.

[22] I make the assumption that both parties are risk averse in money to conveniently de-linearize the utility functions--if we used non-linear kindness functions, a comparable model would work with risk-neutral agents. Also, the assumption that Firm's payoff is 0 (rather than negative) if $e = L$ is made for convenience, and is not essential.

b = 0. Thus, the unique Nash equilibrium is the nasty one in which e = L and b = 0. Because this outcome is also a mutual-min outcome, this will be a fairness equilibrium in which the players feel negatively towards each other.

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

Given Worker's choice of effort, the kindest Firm can be to Worker is by choosing b = R; the least kind is clearly to choose b = 0. Therefore the equitable material payoff to Worker is $R^{1/2}/2 - \gamma$ if e = H, and $R^{1/2}/2$ if e = L. Using this, we can calculate that the kindness of Firm to Worker is given by $f_F = (b/R)^{1/2} - 1/2$.

Using this, let us consider the possibility of a positive fairness equilibrium. What is Firm's utility if it is commonly known that Worker is setting e = H? It is given by:

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

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

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

$$U_W(e{=}H) = b^{1/2} - \gamma + [(1/[1+4R])^{1/2}-1/2]\cdot(1/2)$$
$$U_W(e{=}L) = b^{1/2} + [(1/[1+4R])^{1/2}-1/2]\cdot(-1/2)$$

Algebra yields the conclusion that Worker would not strictly prefer to choose e = L if and only if $R \le .25\cdot[1/(.5+\gamma)^{1/2} - 1]$. For all such combinations of R and γ, therefore, there exists a "gift-giving" equilibrium in which Worker sets e = H, and Firm gives Worker a bonus of $b^* = R/[1+4R]$. Note that the larger is γ, the smaller must be R for there to exist a gift-giving equilibrium. The reason for this is roughly as follows. If γ is large, Worker is very tempted to "cheat" Firm by not working hard. The only way he will not cheat is if Firm is being very kind. But Firm's material costs to yielding a given percentage of profits to Worker increases as R increases; thus, only if R is very small will Firm give Worker a generous enough share of profits to induce worker to be kind.

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

The above model emphasizes the potential for goodwill by worker and firm, and contrasts such goodwill with the possibility that the two parties will pursue their own self-interest. Yet, as we have seen in earlier examples, one implication of my model of fairness is that players may sacrifice their material well-being so as to *hurt* other players that are being unkind. In the context of labor economics, this translates into the possibility that workers will actively punish employers that they believe are being unfair. Consider Example 5:

		Firm	
		high wages	low wages
Worker	high effort	4X, 5X	2X, 7X
	low effort	6X, 2X	4X, 4X
	sabotage	3X, 0	X, 2X

Example 5 -- Worker Sabotage

Example 5 is a simplified model that adds the potential for worker sabotage to the underlying Prisoner's Dilemma of an employment relationship. If the firm pays low wages, the worker can not only exert low effort, but could exert effort to sabotage the firm. Examples of sabotage include any activity by a worker that is costly to both him and to his employer. For instance, a worker might try to ruin computer files of an employer, even if he cannot possibly profit by doing so. Moreover, given some possibility of being punished for this destructive act, such sabotage would result in an expected material cost to the worker.

What will be the fairness equilibria in such a situation? If the material payoffs are very small, then (high effort, high wages) will be the mutual-max fairness equilibrium, as in the Prisoner's Dilemma. But in this game the mutual-min fairness equilibrium will be (sabotage, low wages), rather than (low effort, low wages); the "negative" fairness equilibrium is more negative than in the Prisoner's Dilemma.[23]

Of course, while Example 5 realistically expands the worker's strategy space, it does not allow the firm to undertake a strategy that is quite prevelant in the real world--firing the worker. While firing a worker may be simply because the worker is inherently unproductive, there are also many cases where firms fire workers as an alternative to lowering wages, *because firing a worker represents a way to prevent a disgruntled worker from harming the firm.* Thus, a firm might prefer to fire a worker rather than lower his wages if the firm knows that lowering the wages would anger the worker, even if firing the worker is a very costly thing for the firm to do. Consider Example 6:

		Firm		
		high wages	low wages	fire
	high effort	4X, 5X	2X, 7X	0, 6X
Worker	low effort	6X, 2X	4X, 4X	X, 3X
	sabotage	3X, 0	X, 2X	X, 3X

Example 6 -- The Firing Game

While this type of situation quite inherently involves sequential moves, the normal-form of Example 6 captures *roughly* the following timing: the worker

[23] If the material payoffs are quite large, then there will be only one fairness equilibrium, which is also the unique Nash equilibrium--(low effort, low wages).

chooses high or low effort. Then, without having observed the effort, the firm decides on the policy of high wages, or low wages, or firing the worker. After this, the worker decides whether or not to sabotage. Importantly, sabotage only has an effect on the firm if it has not chosen to fire the worker.

In this game, (sabotage, fire) is a mutual-min Nash equilibrium, so that it is a fairness equilibrium for all values of X. Notice the important role of the potential for sabotage here: For values of X > 3/10, (low effort, fire) is not a fairness equilibrium, because the firm would be too tempted to not fire the employee--paying a low wage is less costly to the firm. Thus, for values of X > 3/10, the firm is firing the worker *not* as retaliation against the worker's low effort, but rather as a means of preventing the worker from sabotaging the firm.

Of course, there are different ways to fire an employee. An obvious way to fire an employee to prevent sabotage is to give him no notice. I have been told that computer trade journals often discuss issues of how to fire somebody. High-tech firms are generally more susceptible to sabotage than most firms, and thus tend to be very meticulous about firing people. People are told they are fired, and then, with security guards on hand, watched as they pack up their belongings, return all keys, and leave the premises. An employment specialist working for a large corporation has told me about a similar tactic--before an employee is told that he is fired, all his computer access codes, etc., are cut off. Example 7 helps illustrate these issues:

		Firm	
		Slow	Fast
Worker	No Sabotage	2X,2X	0,X
	Sabotage	-X,-X	0,X

Example 7 -- The Fast Firing Game

If we think of this game as following previous events that have left the worker feeling hostile, then the worker will aim to sabotage the firm. Firing the worker fast is the firm's optimal response.

Of course, one interpretation of such treatment is that a "repeated game" is coming to an end, and parties will take more effort to avoid opportunism by the other. But this is somewhat inadequate to explain things. It would suggest that non-disgruntled employees who must resign for reasons not indicating dissatisfaction with a firm, would similiarly be be treated in this way. While precautions are probably taken, they would seem not to be as strong in this case as general.

Finally, I wish to illustrate an example where treating an employee well is important *not* because he needs material incentives to work harder, but because he will feel ill-treated if he does not get paid much, and therefore not work hard. Consider Example 8, where a worker's choice of whether or not to be productive has absolutely no effect on his material well-being.

		Firm	
		high wages	low wages
Worker	be productive	2X, 2X	0, 4X
	be unproductive	2X, 0	0, 2X

Example 8 -- Worker Productivity

In this example, there does not exist a fairness equilibrium in which the worker is productive while being paid low wages, no matter the scale of X. In the equilibrium (be productive, low wages), the firm is grabbing all the surplus; this will anger the worker, who will--at no material cost--punish the firm by not being productive.

Baron [1988] argues, for instance, that "there is considerable contemporary evidence that casts doubt on the preeminence of effort aversion as an important problem that needs to be solved by labor contracts." That is, the notion that high productivity by workers involves an inherently unpleasant "effort" by them is much over-played in economic theory. Example 8 illustrates, however, that high pay may be needed to induce high productivity by workers, even if high productivity causes no special effort. High pay will induce high productivity because workers need to be convinced that firms are equitably distributing the gains in profits resulting from the workers' "effort."

VII. Fairness and Welfare

I consider now some welfare implications of fairness.[24] While it is sometimes plausible to assume that rational people behave as if they are maximizing a "goal utility function" while their well-being corresponds to a different "welfare utility function," I assume here that the two coincide. That is, I assume the full utility functions--combining material payoffs *and* "fairness payoffs"--are the utility functions with which to determine social welfare. As such, I believe we should care not solely about how concerns for fairness supports or interferes with material efficiency, but also consider how these concerns affect people's overall welfare.

Consider Example 9:

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

Example 9 -- The Grabbing Game

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

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

they either both do not grab, or both grab, they each get one can; if one grabs, 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). As in the Prisoners' Dilemma, the non-cooperative (grab,grab) outcome is a fairness equilibrium no matter the value of X. For small X, however, the positive, mutual-max outcome (share,share) is also a fairness equilibrium. Moreover, because these two fairness equilibria yield the same material payoffs, (share,share) always Pareto-dominates (grab,grab).

Shopping for minor items is a situation in which people 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, while both (grab,grab) and (share,share) are fairness equilibria when material payoffs are arbitrarily small, the overall utility in each equilibrium is bounded away from zero.[25] *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 9, as with many examples in this paper, there is both a strictly positive and a strictly negative fairness equilibrium. Are there

[25] 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.

games that contain only positive, or only negative, fairness equilibria? If there are, this could be interpreted as saying that there are some economic situations that endogenously determine the friendliness or hostility of the people involved. More generally, 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 fairness equilibrium of the Prisoners' Dilemma with very large material payoffs is the Nash equilibrium, where both players defect. This fairness equilibrium is strictly negative. Interpreting a negative fairness equilibrium as a situation in which parties become hostile to each other, this implies that if mutual cooperation is beneficial, but each person has an irresistible incentive to cheat when others are cooperating, then people will leave the situation feeling hostile.

Are there opposite, happier situations, in which the strategic logic of a situation dictates that people will depart on *good* terms? In other words, are there games for which all fairness equilibria yield strictly positive outcomes? Proposition 6 shows that the answer is *No*: there exists in every game a weakly negative fairness equilibrium.

<u>Proposition</u> <u>6</u>:

In every game, there exists a weakly negative fairness equilibrium.

Proposition 6 states that it is never guaranteed that people will part with positive feelings.[26] It implies a strong asymmetry in my model of

[26] 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.

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 material self-interest tempts a player to be mean even if other players are being kind, material self-interest will never tempt a player to be kind when other players are being mean, because the only way to be kind is to go *against* your material self-interest.[27]

There is another interesting feature of fairness and welfare that deserves comment. Namely, my model has a seemingly paradoxical feature: players may feel more positive towards each other with one outcome than with an alternative outcome that gives them both higher material payoffs.[28]

Consider Example 10:

		Player 2	
		Gift	No Gift
Player 1	Gift	4X,4X	2X,6X
	No Gift	6X,2X	5X,5X

Example 10 -- The Gift Game

This game is, like the previous one, strategically similar to Prisoner's Dilemma--each player has a dominant strategy. Yet notice that the would-be "cooperative" outcome of (Gift,Gift) is Pareto-dominated (in material payoffs)

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

[28] Drew Fudenberg first pointed this out to me.

by the "non-cooperative" outcome of (No Gift, No Gift).

What will happen in this game? As with the outcome (Defect,Defect) in the Prisoners Dilemma, the Nash equilibrium (No Gift,No Gift) is a fairness equilibrium in this game for all levels of material payoffs. Moreover, this is a mutual-min outcome, and generates negative emotions.

For X < 1/4, however, (Gift,Gift) is also a fairness equilibrium, and generates positive emotions between the players. This may be considered odd--the players are feeling better about each other, though the material payoffs are lower than in the negative fairness equilibrium. Yet this is pretty realistic--our image of the exchange of gifts is not that we are increasing overall material surplus, but rather that we generated goodwill. In this game, each player is sacrificing 4X of his own material well-being to increase the material well-being of the other player by only 2X.[29] More generally, standard consumer theory says that exchanging equally costly presents must be a losing proposition. As most of us recognize, gift-giving has an appeal beyond its material benefits.

[29] The *overall* well-being of the players, however, is higher in the gift-giving equilibrium than in the non-gift-giving equilibrium, because the good will outweighs the material costs. We may wonder: are there cases where even the overall utilities are lower in positive fairness equilibria than in negative equilibria? The answer is yes. In essence, there can be a coordination problem, where players are stuck in a goodwill world, when they would prefer at some emotional cost but greater material benefit to be in a bad-will world.

Nevertheless, in any symmetric 2 × 2 gift-giving game that is strategically analogous to the Prisoner's Dilemma, the gift-giving fairness equilibrium always (when it exists) Pareto-dominates the non-gift-giving fairness equilibrium. This is because in order for the incentive to cheat on the cooperative equilibrium to be resistable, the material payoff to cheating has to be small compared to the material payoff of cooperation. But a player's payoff from not giving a gift when the other player is giving a gift must be higher than his material payoff when neither gives a gift (otherwise, "giving a gift" does not help the other player). But this puts an upper bound on how much higher the non-gift-giving payoffs can be relative to the gift-giving payoffs, and this upper bound is small enough that the non-gift-giving equilibrium cannot have higher overall utility.

VIII. Discussion and Conclusion

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

For instance, problems arise because my definition of a player's fairness is very contingent on the other player's actions. Players consider only each other's willingness to resist unilateral deviations, rather than taking into account the desirability of an outcome relative to all possible outcomes in a game. This focus on unilateral deviations is very much in the spirit of non-cooperative game theory, and highlights the coordination problems that can arise in any strategic situation. It is because of this approach, for instance, that the model predicts that differing expectations can yield very different material and emotional outcomes in the same physical situation.

Yet there are situations in which a more "global" model of fairness would add realism. In Battle of the Sexes, for instance, 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 direct distinction; in fact, in examining the outcome (Boxing,Boxing), the players' emotions do not depend at all on the payoffs they would get in (Opera,Opera). Further research could consider how players incorporate broader aspects of a game into their emotions.

Another way in which the current model is limited is that it has players judging the fairness of outcomes in a situation without any reference to what they might consider to be the "expected" outcome. This is clearly unrealistic. Evidence indicates that people's notions of fairness are heavily influenced by the status quo, and other reference points. For instance, Kahneman, Knetsch, and Thaler [1986] illustrate that the consumer's view of the fairness of prices charged by a firm can be heavily influenced by what that firm has charged in the past.

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 incomplete-information games is essential for applied research, but doing so will lead to important new issues. Because the theory depends so heavily on the motives of other players, and because interpreting other players' motives depends on beliefs about their payoffs and information, incomplete information is likely to have a dramatic effect on decision-making.

Extending the model to sequential games is also essential for applied research. In conventional game theory, observing past behavior can provide information; in psychological games, it can conceivably change the motivations of the players. An important issue arises: can players "force" emotions--that is, can a first mover do something that will compel a second player to regard him positively? One might imagine, for instance, that an analog to Proposition 6 might no longer be true, and sequential games could perhaps be used as mechanisms that guarantee positive emotions.

Finally, future research can also focus on modeling additional emotions. In Example 11, for instance, my model predicts no cooperation, whereas it seems plausible that cooperation would take place.[30]

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

Example 11 -- 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

[30] A related example was first suggested to me by Jim Fearon.

involved.

We might think that (Trust,Share) *is* a reasonable outcome. This would be 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 trust was psychologically sound, we could incorporate it into formal models.

Appendix A: The Kindness Function Can Be Generalized

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

Definition A1 requires that 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 $\pi_j(b_j,a_i) > \pi_j(b_j,a_i')$.

Definition A2 requires that the payoff that player j "deserves" is strictly between player j's worst and best Pareto-efficient 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 $\pi_j^h(b_j) > \pi_j^e(b_j) > \pi_j^l(b_j)$

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?

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,

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

Appendix C: The Utility Function Can Be Generalized

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:

$$\tilde{U}_i(a_i,b_j,c_i) \equiv \pi_i(a_i,b_j) + [\alpha + (1-\alpha)\tilde{f}_j(b_j,c_i)]\cdot[1+f_i(a_i,b_j)]$$
$$\text{where } \alpha \in [0,1].$$

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.[31]

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

Proof of Proposition 1:

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

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

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.

Proof of Proposition 2:

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

Proof of Proposition 3:

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 fairness equilibrium. If (a_1,a_2) is a strictly negative Mutual-Min outcome, each player would strictly prefer to play a_i, since this uniquely maximizes the fairness product. Thus, this too would be a fairness equilibrium. Q.E.D.

Proof of Proposition 4:

Suppose that (a_1,a_2) is not a Nash equilibrium. Then (without loss of

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.

Levine, David I., "Cohesiveness, Productivity, and Wage Dispersion," *Journal of Economic Behavior and Organization 15*, 237-255, 1991.

Levine, David I., "Fairness, Markets, and Ability to Pay: Evidence from Compensation Executives," mimeo, Haas School of Business, University of California--Berkeley, July 1991.

Mansbridge, Jane J., ed., *Beyond Self-Interest*, Chicago, IL: The University of Chicago Press, 1990.

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

Mui, Vai-Lam, *Two Essays in the Economics of Institutions: I. Envy*, PhD Dissertation, Department of Economics, University of California--Berkeley, 1992.

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

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir, "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review 81*, 1068-1095, 1991.

Rotemberg, Julio J., "Human Relations in the Workplace," mimeo, MIT, January 1992.

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.

Tversky, Amos and Daniel Kahneman, "Loss Aversion and Riskless Choice: A Reference Dependent Model," *Quarterly Journal of Economics*, 1991.

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.

July 7, 1992

Working Paper Series
Department of Economics
University of California, Berkeley

Individual copies are available for $3.50 within the USA and Canada; $6.00 for Europe and South America; and $7.00 for all other areas. 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."

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.

91-186 "How Pervasive Is the Product Cycle? The Empirical Dynamics of American and Japanese Trade Flows." Joseph E. Gagnon and Andrew K. Rose. December 1991.

92-187 "Shocking Aspects of European Monetary Unification." Tamim Bayoumi and Barry Eichengreen. January 1992.

92-188 "Is There a Conflict Between EC Enlargement and European Monetary Unification?" Tamim Bayoumi and Barry Eichengreen. January 1992.

92-189 "The Marshall Plan: Economic Effects and Implications for Eastern Europe and the Soviet Union." Barry Eichengreen and Marc Uzan. January 1992.

92-190 "Exploring the Relationship between R&D and Productivity at the Firm Level in French Manufacturing." Bronwyn H. Hall and Jacques Mairesse. February 1992.

92-191 "Three Perspectives on the Bretton Woods System." Barry Eichengreen. February 1992.

92-192 "Are Futures Margins Adequate?" Roger Craine. April 1992.

92-193 "Heterogeneity in Organizational Form: Why Otherwise Identical Firms Choose Different Incentives for Their Managers." Benjamin E. Hermalin. May 1992.

92-194 "Investment and Research and Development at the Firm Level: Does the Source of Financing Matter?" Bronwyn H. Hall. May 1992.

92-195 "The Determinants of Efficiency and Solvency in Savings and Loans." Benjamin E. Hermalin and Nancy E. Wallace. May 1992.

92-196 "Economics of Development and the Development of Economics." Pranab Bardhan. June 1992.

92-197 "The Core and the Hedonic Core: Equivalence and Comparative Statics." Suzanne Scotchmer and Greg Engl. July 1992.

92-198 "Incorporating Behavioral Assumptions into Game Theory." Matthew Rabin. July 1992.

92-199 "Incorporating Fairness into Game Theory and Economics." Matthew Rabin. July 1992.