UNIVERSITY OF CALIFORNIA AT BERKELEY

Department of Economics

Berkeley, California 94720-3880

Working Paper No. 97-252

Fairness in Repeated Games

Matthew Rabin

Department of Economics
University of California, Berkeley

January 1997

Key words: fairness, incremental games, psychological games, public goods, reciprocal altruism, repeated games, replicated games, revenge

Non-keywords: earthenware, kohlrabi, nozzle, tissue paper

JEL Classification: A12, A13, B49, C70, D63

I thank David Frankel and especially Eddie Dekel for helpful comments. Financial support from the National Science Foundation (Grant SES92-10323) is gratefully acknowledged.
Contact: Matthew Rabin, Department of Economics, University of California, Berkeley, Berkeley, CA 94720-3880. e-mail: rabin@econ. berkeley.edu

Fairness in Repeated Games

Abstract

In addition to pursuing their material self-interest, people are motivated to help those who are kind to them, and to hurt those who are mean to them. Such social preferences influence behavior most when material stakes are small. Rabin (1993) defines an outcome reflecting such preferences as *fairness equilibrium*. This paper applies a version of fairness equilibrium to repeated games. Some fairness-equilibrium outcomes in small-stakes, one-shot games are shown to be fairness-equilibrium outcomes every period in *incremental games*, which are finitely repeated games of large overall material stakes but very small per-period stakes. For instance, it is a fairness equilibrium for players to cooperate in *every* period of the finitely repeated Prisoners' Dilemma with arbitrarily high total payoffs, so long as the per-period material payoffs are small. I consider more generally whether fairness equilibria in small-stakes, one shot games can be the stationary fairness-equilibrium outcomes in incremental games, providing sufficient and (approximately equivalent) necessary conditions for this result to hold for *all* fairness preferences meeting my general assumptions. I also show that outcomes that yield either player below her minmax payoffs (which is often true of fairness equilibria in small-stakes, one-shot games) cannot be stationary fairness-equilibrium outcomes for any fairness preferences meeting the general assumptions in incremental games of sufficiently large overall payoffs.

1. Introduction

In Rabin (1993), I develop a game-theoretic model of some social components of preferences which have conventionally been ignored by economists. The solution concept *fairness equilibrium* assumes that people may prefer to sacrifice their material well-being both to help those who are being kind to them and to punish those who are being unkind.[1] Of course, people are also self-interested. Fairness equilibrium therefore assumes that players are also motivated by material self-interest, and that socially-oriented goals affects behavior less as the marginal cost of sacrificing becomes larger.

This paper explores the implications of fairness equilibrium in finitely repeated games, comparing the predicted behavior to both fairness equilibria in one-shot games and to classical predictions in finitely repeated games. I show that fairness equilibrium predicts that reciprocally-altruistic outcomes--such as cooperation in the Prisoners' Dilemma--can occur in *every* period of *incremental games*, which are finitely repeated games with arbitrarily large total payoffs, so long as per-period payoffs are negligible. More generally, I provide some conditions for fairness equilibria in small-stakes, one-shot games to be the stationary fairness-equilibrium outcomes in incremental games, providing sufficient and (approximately equivalent) necessary conditions for this result to hold for *all* fairness preferences meeting my general assumptions. I also show that outcomes that yield either player below her minmax payoffs (which fairness equilibria often do in small-stakes, one-shot games) cannot be stationary fairness-equilibrium outcomes for *any* fairness preferences meeting the general assumptions in incremental games of sufficiently large overall payoffs.

I begin in Section 2 by presenting the model and basic results from Rabin (1993). Two types of outcomes play a special role in the results: "mutual-max" outcomes--in which, given the other player's behavior, each player maximizes the other's material payoffs--and "mutual-min" outcomes--in which, given the other player's behavior, each player minimizes the other's material payoffs. Rabin (1993) shows that, when material payoffs are arbitrarily small, then, *roughly*, an outcome is a fairness equilibrium in a one-shot game if and only

---

[1] I will not here outline the evidence that these motivations are important. For introductions to these principles, see Dawes and Thaler (1988), Kahneman, Knetsch, and Thaler (1986a,1986b), Rabin (1993), and Thaler (1988).

if it is a mutual-max or a mutual-min outcome.

		P2: Cooperate	P2: Defect
P1	Cooperate	4X, 4X	0, 6X
	Defect	6X, 0	X, X

Prisoner's' Dilemma

Figure 1A

		P2: Chicken	P2: Dare
P1	Chicken	4X, 4X	2X, 5X
	Dare	5X, 2X	0, 0

Chicken

Figure 1B

		P2: Gift	P2: No Gift
P1	Gift	3X, 3X	0, 6X
	No Gift	6X, 0	4X, 4X

Gift Game

Figure 1C

Figure 1 -- In each game, payoffs are scaled by $X > 0$

To illustrate these ideas, consider Figure 1. If it is common knowledge to the players that they are playing the mutual-max outcome (Cooperate,Cooperate) in the *Prisoners' Dilemma*, for instance, then each player knows that the other is sacrificing his own material well-being in order to be nice. Because each player is being nice to the other, each will *want* to be nice to the other. If X is small enough, so that defecting is not too tempting, (Cooperate,Cooperate) is a fairness equilibrium. The mutual-max outcomes (Chicken,Chicken) in *Chicken* and (Gift,Gift) in *Gift* will likewise be equilibria for small enough stakes. The outcomes (Defect,Defect) in *Prisoners' Dilemma*, (Dare,Dare) in *Chicken*, and (No Gift,No Gift) in *Gift* are all mutual-min outcomes, in which each player is being mean to the other. Because these outcomes generate hostility and the desire to hurt each other, they will also be fairness equilibria if the stakes are small.[2]

By making some additional assumptions about preferences, I examine in Sections 3 and 4 the implications of fairness equilibrium in repeated games. In Section 3, I consider the implications of my model for *incremental games*, which are finitely repeated games with arbitrarily many repetitions and negligible per-period payoffs. The results largely pertain to *stationary incremental fairness equilibria* (*SIFEs*), which are outcomes that are played

---

[2] The outcomes (Defect,Defect) and (No Gift,No Gift) are fairness equilibria independent of scale; because they are both Nash equilibria and mutual-min outcomes, both material self-interest and desire for revenge lead the players to behave as they are. By contrast, the stakes must be low enough to make (Dare,Dare) a fairness equilibrium, because a player will be tempted by material self-interest to deviate from this outcome.

every period in some fairness equilibrium of the incremental game.

I show that every fairness equilibrium of any one-shot game is a SIFE of the corresponding incremental game whose total payoffs are the same as the one-shot game. That is, allowing players to make decisions incrementally never eliminates outcomes that are possible when players make once-and-for-all decisions. Additional outcomes may also be SIFEs in incremental games, however. I show that a class of mutual-max outcomes will be SIFEs whenever each player has some repeated-game strategy that guarantees *both* players get per-period payoffs below the mutual-max outcome. The argument relies on punishments hurting *both* players because if the *non*-deviating player would get a higher payoff by punishing the other player than in the mutual-max outcome, she may resent the other player for not deviating, and thus be unwilling to cooperate.

As applied to the repeated Prisoner's Dilemma, this result implies that no matter how large the total stakes, and no matter how small a component of overall preferences is fairness, there is a fairness equilibrium in which the players cooperate each period of a finitely repeated Prisoner's Dilemma, so long as the per-period payoffs are allowed to be arbitrarily small. Intuitively, punishment strategies that hurt a player induce him to cooperate out of self-interest through most of the game. But when per-period payoffs are negligible, even minimally fair players will be willing to reciprocally cooperate in the last few periods of the game. Even a little bit of concern for fairness means that the standard backwards-induction, "unraveling" problem (whereby players can never plan to reward cooperation with cooperation in the final periods) does not arise.

In games such the *Prisoner's Dilemma*, therefore, fairness equilibrium predicts that players might, in incremental games of arbitrarily large total material payoffs, behave the same way as they do in small-scale, one-shot games. This resemblance of behavior very much depends on the strategic structure of the game being considered, however. I show, in fact, that the condition that players have mutually harmful punishment strategies is *necessary* for such a resemblance, in the following sense: If there are no such "supporting strategies" for an outcome, there exists *some* fairness preferences meeting the general assumptions of my model such that the outcome will not be a SIFE. Moreover, I show that when an outcome yields either player below her minmax payoff, then the outcome cannot occur all or even close to all of the time in any fairness equilibrium of incremental games with sufficiently high overall payoffs. For instance, the mutual-max outcome (Gift,Gift) in *Gift* and

the mutual-min outcome (Dare,Dare) in *Chicken* will not be SIFEs for sufficiently large overall stakes for any fairness preferences meeting the basic assumptions of the model.

In Section 4, I consider the role of fairness in *replicated games*, which are repeated games with arbitrarily many periods whose per-period payoffs are non-negligible. I show that every fairness equilibrium in such games yields each player at least his minmax payoff, and give conditions (somewhat more restrictive than in the Nash folk theorem) under which all outcomes yielding more than minmax payoffs are fairness equilibria. I also argue that replication may *eliminate* some cooperative outcomes that one can get in one-shot games. Even if a preference for fairness induces players to cooperate in a one-shot Prisoners' Dilemma, for instance, they may be unwilling to cooperate in *either* period when that game is played twice. While Section 3 indicates that "incrementalizing" a game can help players cooperate, Section 4 illustrates that there is no presumption that repetition per se will help fairness-oriented players cooperate, any more than it helps self-interested people cooperate for standard folk-theorem reasons.

The analysis of this paper is complicated by some issues that are likely to arise more generally in the application of psychology to repeated games. First, when players' well-being reflects both material payoffs and emotional payoffs, the scale of payoffs in a game may be important, so that we must analyze repeated games separately depending on both the per-period and the overall scale of material payoffs. A second complication arises in considering rewards and punishments in repeated games. Classical folk theorems rely (more or less) solely on punishments involving stationary behavior, where a player is deterred from deviating with the threat that other players will play the same minmax stage-game strategies forever following a deviation. In the model presented below, the punishments needed to deter deviations from an equilibrium may *require* non-stationary strategies. Cooperation in the repeated Prisoner's Dilemma, for instance, is sustainable only with strategies such as tit-for-tat, and could not be sustained with an stationary punishment strategies such as "Defect forever".[3]

---

[3] Such non-stationary strategies are needed because the model does not suppose that players' overall utilities are quasi-concave in their material payoffs. If the utility functions were quasi-concave, then anything that could be detered by a non-stationary strategies could also be detered by players threatening to punish deviations with a stationary strategy that "minmaxes" the other player's utility defined in terms of some function of the material

The final complication reflects the fact that players' utilities may depend in my model not just on strategies played but directly on their expectations.[4] This can mean that the unpleasantness of a punishment depends on what equilibrium is being played, so we cannot necessarily find a universally "worst" punishment that will deter a player from deviating from all putative equilibria. Punishment A may deter a player from deviating from equilibrium X, and punishment B may deter him from deviating from equilibrium Y, even though neither works as a punishment for deviations from the other equilibrium.

I use these complications as a partial excuse for two big shortcomings of the model presented in this paper. First, I incorporate no notion of sequential rationality, applying only the normal-form solution concept developed in Rabin (1993). In lieu of a satisfactory model of sequential rationality applied to fairness issues, some of my results must therefore be treated as tentative.[5] A second shortcoming is that I do not provide a full characterization of the set of fairness equilibria in repeated games; rather, I concentrate one which outcomes that are fairness equilibria in small-stakes, one-shot games are also fairness-equilibrium outcomes in repeated games.

I conclude the paper in Section 5 with a brief discussion comparing the predictions of my model to the experimental evidence on repeated games, and to

---

payoffs to the players.

I believe that an assumption of quasi-concavity would not be a natural restriction in this context. Consider the "fairness payoffs" that a player gets from sharing a pie with the other player if she feels positively towards that other player. Quasi-concavity would imply that the player gets more marginal satisfaction from giving the other player 20% rather than 10% of the pie than she would get from giving the player 50% rather than 40% of the pie. The opposite is likely to be the case in most situations--giving only 20% of the pie to somebody who is being kind to you will be unlikely to make you feel very good.

[4] Formally, my model uses the framework of *psychological games* developed by Geanokoplos, Pearce, and Stacchetti (1989).

[5] While Geanokoplos, Pearce, and Stacchetti (1989) develop definitions of sequential rationality in psychological games, the solution concepts they develop in essence require that players must maintain the same emotional disposition at all contingencies in the game. This seems inappropriate when applied to my model. In a repeated prisoners' dilemma, for instance, the GPS approach would require that players involved in a cooperative equilibrium ought maintain a friendly disposition following deviations to non-cooperative behavior. I am currently working (with Jim Fearon) on an extensive-form version of fairness equilibrium which tries to capture sequential rationality, allowing unexpected moves by one player to change the emotional disposition of the other.

incomplete-information models of cooperation such as that presented in Kreps *et al* (1982).

2. Fairness Equilibrium

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. Throughout the paper I shall consider only games G whose material payoffs are generic--for each player, I assume that his material payoffs from any two outcomes is not the same. Let $\pi_i: S_1 \times S_2 \rightarrow \mathbb{R}$ be player i's *material payoffs*. To formalize fairness, I adopt the framework developed by Geanakoplos, Pearce, and Stacchetti (1989) (hereafter, GPS) who modify conventional game theory by allowing payoffs to depend on players' beliefs as well as on their actions. I assume that each player's subjective expected utility depends 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.[6]

I begin the specification of fairness equilibria by defining a *kindness function*, $f_i(a_i, b_j)$, which measures how kind player i is being to player j. This function will be positive when player i believes he is treating player j kindly and negative when player i believes he is being mean to player j. Likewise, I define $\tilde{f}_j(b_j, c_i)$, which represents player i's beliefs about how kindly player j is treating him; this function is positive when he believes player j is trying to treat him kindly and negative when he believes that player j is trying to treat him meanly.

How kind is player i being to player j if he chooses strategy a_i in response to his beliefs that player j is choosing strategy b_j? To consider

---

[6] It is because a player's utility depends in this model on his second-order beliefs that we require the apparatus of psychological games.

this, note that player i believes that he is choosing some payoff pair from the set $\Pi(b_j) \equiv \{(\pi_i(a,b_j),\pi_j(b_j,a))|a\in S_i\}$. The nicest player i can be to player j is to choose a strategy that yields player j his highest payoff among the set $\Pi(b_j)$: $\pi_j^h(b_j) \equiv \text{Max}_{a\in S_1}\{\pi_j(a,b_j)\}$. The meanest that player i could be to player j is to choose player j's lowest payoff in $\Pi(b_j)$: $\pi_j^{\min}(b_j) \equiv \text{Min}_{a\in S_1}\{\pi_j(a,b_j)\}$.

Note that even selfish behavior by player i may often involve giving player j a payoff higher than $\pi_j^{\min}(b_j)$, because a strategy by player i that yields player j $\pi_j^{\min}(b_j)$ may also in fact hurt player i. We may consider player i's most relevant distributional choice to be his choice among payoff pairs on the Pareto frontier, and we may consider it relatively mean for player i to grab his best payoff on the Pareto frontier. We thus define the payoff $\pi_j^1(b_j)$ as player j's lowest payoff *among points that are strictly Pareto-efficient in* $\Pi(b_j)$. Formally, $\pi_j^1(b_j) \equiv \text{Max}_{a\in S_1}\{\pi_j(a,b_j)|a \in \text{argmax } \pi_i(a,b_j)\}$.

Rather than using the specific form of kindness functions used in the test of Rabin (1993), I shall in this paper work directly with the more general class of kindness functions introduced in the Appendix of that paper, adding some additional properties needed for the results of this paper.

Assumption 1 requires that 1) 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, and 2) the utility and marginal utility derived from fairness is bounded.[7]

Assumption 1:

The kindness functions are *Bounded and Increasing*. For every game,

1) $f_i(a_i,b_j) > f_i(a_i',b_j)$ iff $\pi_j(b_j,a_i) > \pi_j(b_j,a_i')$; and
2) There exists a number $N_1 > 0$ such that
   a) For all $a_i \in S_i$ and $b_j \in S_j$, $f_i(a_i,b_j) \in [-N_1,N_1]$, and
   b) For $i = 1,2$ and $j \neq i$, for all a, a', α, $\alpha' \in S_i$ and $b \in S_j$ such that $\pi_j(a,b) - \pi_j(a',b) = \pi_j(\alpha,b) - \pi_j(\alpha',b) \neq 0$,
   $$[f_i(a,b) - f_i(a',b)]/[f_i(\alpha,b) - f_i(\alpha',b)] \leq N_1.$$

---

[7] In the formulation of Assumption 1 in the Appendix of Rabin (1993), I omit the restriction on the marginal value of the kindness function here incorporated into part 2b The proof of Proposition 5 (Proposition A5 below) given in that paper essentially assumes such a restriction, so that (though it does apply to example of kindness functions used in the text of that paper) it is erroneous as applied to the class of kindness functions presented in the Appendix.

Assumption 1 incorporates the idea that how kind player i is being to player j is determined by the payoff he is giving to player j. With this perspective, it is natural to define an "equitable payoff," $\pi_j^e(b_j)$, as the morally neutral--neither kind nor mean--level for player i to give to player j. Then, giving j a higher payoff than $\pi_j^e(b_j)$ is nice, and giving player j a lower payoff than $\pi_j^e(b_j)$ is mean. Assumption 2 requires that $\pi_j^e(b_j)$ be strictly between player j's worst and best Pareto-efficient payoff, so long as the Pareto frontier is not a singleton.

Assumption 2:

The kindness functions are *Pareto Splits*. There exists some $\pi_j^e(b_j)$ such that:

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

Assumption 3 guarantees that notions of the fairness of particular outcomes do not dramatically change when all payoffs are (say) doubled.[8]

Assumption 3:

The kindness functions are *Affine*. Changing all payoffs for both players by the same positive affine transformation does not change the value of $f_i(a_i,b_j)$.

Assumption 3 is crucial to the model, because it is what guarantees that behavior will be more influenced by material self-interest as the material payoffs at stake increase, and more influenced by the taste for fairness as the material payoffs at stake decrease. Finally, a continuity assumption is needed to apply GPS's general existence theorem to the model of this paper:

---

8 Assumption 3 does, however, allow the kindness functions to be sensitive to affine transformations of *one* player's payoffs. If we double player 2's payoffs, then it may be that fairness dictates that he get more--or less--than before.

Assumption 4:

The kindness functions are continuous. For $i = 1,2$ and $j \neq i$, for all $a_i \in S_i$ and $b_j \in S_j$, $f_i(a_i,b_j)$ is continuous in both a_i and b_j.

Given these kindness functions (with the functions $\tilde{f}_j(b_j,c_i)$ assumed to have the same properties), I represent player i's utility by the function $U_i(a_i,b_j,c_i)$, which incorporates both his material utility and the players' shared notion of fairness:

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

These preferences reflect the assumptions outlined in the introduction: If player i believes that player j is treating him badly--which will correspond to $\tilde{f}_j(\cdot) < 0$--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 i believes that player j is treating him kindly, then $\tilde{f}_j(\cdot)$ will be positive, and player i will wish to treat player j kindly.

While the kindness functions are insensitive to positive affine transformations of the material payoffs, the overall utility *is* sensitive to such transformations. Because the kindness functions are bounded above and below, 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.

Because these preferences form a psychological game, we can use the concept *psychological Nash equilibrium* defined by GPS; this solution concept is the natural analog of Nash equilibrium for psychological games, imposing the additional condition that all higher-order beliefs match actual behavior. I call the solution concept thus defined *fairness equilibrium*.

Definition 1:

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) $c_i = b_i = a_i$.
2) $a_i \in \text{argmax}_{a \in S_i} U_i(a,b_j,c_i)$, and

Characterizing the set of fairness equilibria in a game revolves around two types of outcomes. A *mutual-max* outcome is one where players mutually

maximize each other's material payoffs; a *mutual-min* outcome is one where they mutually minimize each other's material payoffs.

Definition 2:

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

Definition 3:

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

In the Prisoner's Dilemma, (Cooperate,Cooperate) is a mutual-max outcome and (Defect,Defect) is a mutual-min outcome. In Chicken, (Chicken,Chicken) is a mutual-max outcome and (Dare,Dare) is a mutual-min outcome.

Definition 4 provides some useful characterizations of outcomes in terms of the level of kindness induced by each of the players.

Definition 4:

An outcome (a_1,a_2) is *strictly positive* if, for $i = 1,2$, $f_i(a_i,a_j) > 0$. An outcome is *strictly negative* if, for $i = 1,2$, $f_i(a_i,a_j) < 0$. An outcome is *weakly negative* if, for $i = 1,2$, $f_i(a_i,a_j) \leq 0$.

Finally, Definition 5 characterizes those outcomes in which each player chooses the action that maximizes his own material payoffs among those actions that yield the other player the same material payoff. Because players are motivated in part by self-interest, and because each player's desire to reward or punish the other is manifested only through his effect on the material payoffs of the other, only such strategies will be played.

Definition 5:

An outcome (a_1,a_2) involves *No Pointless Sacrifice (NPS)* if, for $i = 1,2$ and $j \neq i$, there does not exist a_i' such that both $\pi_j(a_i',a_j) = \pi_j(a_i,a_j)$ and $\pi_i(a_i',a_j) > \pi_i(a_i,a_j)$.

I now present Propositions A1-A5, which reproduce (with one correction) Propositions 1-5 from Rabin (1993). (The proofs are omitted, and can be found in the Appendix to Rabin (1993).) Proposition A1 states that any Nash equilibrium that is also a mutual-max or a mutual-min outcome is also a

fairness equilibrium. Intuitively, these are outcomes in which each player is, given the other's behavior, simultaneously maximizing both his material and his fairness in payoffs.

Proposition A1:

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

Proposition A1 guarantees, for instance, that the outcome (Defect,Defect) is a fairness equilibrium in the Prisoner's Dilemma. Proposition A2 states that, in all fairness equilibria, players have the same disposition towards each other; either they both sacrifice to help the other, or neither sacrifices to help the other.

Proposition A2:

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

While Propositions A1 and A2 pertain to all games, additional results pertain to the limit cases where the scale of material payoffs is made either arbitrarily large or arbitrarily small. For every positive X, let the game G(X) be the game G where all the material payoffs are multiplied by X. Then:

Proposition A3:

For any NPS 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).[9]

While Proposition A1 guarantees that (Defect,Defect) is a fairness equilibrium in the Prisoner's Dilemma irrespective of the scale of material payoffs, Proposition A3 establishes that, if material payoffs are very small, then (Cooperate,Cooperate) is also a fairness equilibrium in the Prisoner's

---

[9] The statement of Proposition A3 is a slightly modified and corrected version of Proposition 3 in Rabin (1993), adding the criterion that the outcome must be NPS. The proof of Proposition 3 in Rabin (1993) explicitly but incorrectly asserts that the NPS condition must hold in any strictly positive mutual-max and strictly negative mutual-min outcomes. With the corrected form of the Proposition, the original proof holds.

Dilemma, and that both (Chicken,Chicken) and (Dare,Dare) are fairness equilibria in Chicken. While Proposition A3 establishes that generally mutual-max and mutual-min outcomes are fairness equilibria when material stakes are small, Proposition A4 establishes that typically mutual-max and mutual-min outcomes are the *only* fairness equilibria when material stakes are small.

Proposition A4:

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).

Finally, Proposition A5 shows that only Nash equilibria can be fairness equilibria when material stakes are large:

Proposition A5:

If (a_1,a_2) is a *strict* Nash equilibrium, 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, 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).

3. Fairness in Incremental Games

Propositions A3 and A4 of the previous section help characterize the set of fairness equilibria in small-stakes, one-shot games. In this section, I consider repeated games in which the material stakes each period are negligible, and develop some results corresponding to and contrasting with Propositions A3 and A4. Consider a finite-strategy, one-shot game G. As in the previous section, let G(X) be the game G where the payoffs for each player is multiplied by X, Let $G^T(X)$ be the repeated game (without discounting) consisting of T repetitions of G(X). To simplify analysis, I assume that even mixed strategies are observed by players at the end of each period.

I define a class of games called *incremental games,* which is a catch-phrase for finitely repeated games where the payoffs in each period are negligible. Formally, the *T-incremental game* is the game $G^T(1/T)$. Note that the T-incremental game G^T is a finitely repeated game whose *total* payoffs is

equivalent to the payoffs from the game G. Thus, when T is very large the material payoffs each period of $G^T(1/T)$ are very small.

A couple of additional assumptions about the kindness functions are needed to analyze incremental games. First, because I wish to compare games with differing strategic structures (one-shot games vs. incremental games), I assume the kindness function does not depend in an "ad hoc" manner on the strategic structure of a game. Assumption 5 guarantees that if a decision by player i involves a choice among precisely the same set of payoffs as does a decision in the corresponding incremental game, then he should measure the kindness of the payoffs he chooses in the two situations in the same way. This assumption seems natural insofar as emotional implications of behavior depend on payoff consequences and not on the precise physical actions associated with the behavior.

Assumption 5:

If the convex hull of payoffs in games G and G′ are identical, then for any strategies b_j, b'_j by player j in games G and G′ such that $\Pi(b_j) \equiv \Pi(b'_j)$, $f_i(a_i, b_j) = f_i(a'_i, b'_j)$ if and only if $\pi_j(a_i, b_j) = \pi_j(a'_i, b'_j)$.

Assumption 5 (even without maintaining Assumptions 1-4) trivially implies a simple result relating the set of fairness equilibria in one-shot games to the fairness equilibria in the corresponding incremental games:

Proposition B1:

Suppose the kindness functions meet Assumption 5. Let σ_i be the strategy in which player i plays a_i in each period of the game $G^T(1/T)$ irrespective of what happens in earlier periods. Then (σ_1, σ_2) is a fairness equilibrium in $G^T(1/T)$ if and only if (a_1, a_2) is a fairness equilibrium in the game G.

Proof:

Given that player j is playing a_j in each period, player i's set of payoffs he chooses from is exactly $\Pi(a_j)$ in the original game G. Moreover, the choice of the stationary strategy a_i in the incremental game selects exactly the same payoff pair in this set as does the strategy a_i in the game G. By Assumption 5, therefore, (a_1, a_2) is a fairness equilibrium in G iff (σ_1, σ_2) is a fairness equilibrium in G^T. Q.E.D.

Proposition B1 says that we do not eliminate any fairness equilibria when

we change a one-shot game into an incremental game. This is because if each player j is choosing to play a_j every period no matter what happens in the game, then each player i faces exactly the same choice among payoff pairs as he does given the same strategy in the one-shot game. This result is similar to the fact that any Nash equilibrium of a game can also be the outcome each period in a Nash equilibrium of the finite repetition of that game. But the analogy breaks down for two reasons. First, Proposition B1 very much depends on incremental games having the same total payoffs as the one-shot game, and does not extend to repeated games where the one-shot game is repeated without reducing the scale each period.[10] Second, while it is always a Nash equilibrium of a repeated game to play a different stage-game Nash equilibrium each period, the analog here is not true: Playing one stage-game fairness equilibrium in period 1 and a different stage-game fairness equilibrium in period 2 typically will not constitute a fairness equilibrium in a two-period game.[11]

Proposition B1 identifies all fairness equilibria consisting of stationary, history-independent strategies in incremental games. But there may of course be other fairness equilibria. In considering what these additional equilibria might be, I specify a further assumption about kindness functions. Assumption 3 of the previous section says that if we make the material stakes of a given game negligible, behavior will typically be determined by fairness considerations. Assumption 6 is a generalization of Assumption 3: It says, roughly, that when a player is choosing among a very small set of efficient outcomes, concerns for fairness will typically determine his behavior. Assumption 6 requires this in a brute-force way: I assume that, for any sequence of Pareto-frontiers which converge to zero in total size, the limit of potential kindness payoffs for the players are infinitely greater than the

---

[10] For instance, for a given pair of kindness functions and scale of payoffs, the outcome (Dare,Dare) may be a fairness equilibrium in the one-shot game of Chicken. Yet I will show in Section 4 that if per-period payoffs are non-negligible, the outcome (Dare,Dare) *cannot* be a frequent outcome in any fairness equilibrium of repeated Chicken with enough repetitions.

[11] For instance, the strategies of playing a strictly positive stage-game fairness equilibrium in period 1 and, irrespective of first-period play, playing a strictly negative, non-Nash stage-game fairness equilibrium in period 2 will never be a fairness equilibrium to a two-period game. If the emotions generated by these strategies are positive or neutral, they will generate a period 2 deviation; if they are negative, they will generate a period 1 deviation.

size of the Pareto frontiers.[12] Formally, for any payoff set $\Pi(b_j)$, let $\Pi^P(b_j)$ be defined as the set of points that are not Pareto dominated by other points in $\Pi(b_j)$. Let $f_i^h(b_j)$ be the nicest player i can be to player j, and let $\pi_i^h(b_j)$ and $\pi_i^l(b_j)$ be player i's highest and lowest payoff among points in $\Pi^P(b_j)$.

Assumption 6:

Consider any sequence of games, $\{G_n\}_{n\to\infty}$, and any $c > 0$. Let $(b_i(G_n), b_j(G_n))$ be strategies such that $\Pi^P(b_j(G_n))$ and $\Pi^P(b_i(G_n))$ each converge to a single point, but for all $n < \infty$, $\Pi^P(b_j(G_n))$ and $\Pi^P(b_i(G_n))$ are a non-singleton locus of points such that the slope $(d\pi_j/d\pi_i)$ and inverse slope $(d\pi_i/d\pi_j)$ at all points is less than $-c$. Then for $i = 1,2$ and $j \neq i$,

$$\text{Limit}_{n\to\infty} f_i^h(b_j)\cdot f_j^h(b_i)/[\pi_i^h(b_j)-\pi_i^l(b_j)] \to \infty.$$

Assumption 6 directly implies that the limit of a sequence of mutual-max strategies which involve almost no scope for material gains from deviating will be a fairness equilibrium. This provides one key to the next proposition. This and other propositions will refer to fairness equilibria where behavior on the equilibrium path is the same each period. For ease of reference, I define such equilibria for incremental games in the limit case as the number of periods, T, becomes arbitrarily large:

Definition 6:

An outcome (a_1, a_2) in the game G is a *stationary incremental fairness equilibrium (SIFE)* if there exists a $\bar{T}$ such that for all $T \geq \bar{T}$ there exists a fairness equilibrium in the game $G^T(1/T)$ that involves the players playing (a_1, a_2) in every period.

Translating into the language of this section, I argued in the Introduction that (Cooperate,Cooperate) in the incremental Prisoners' Dilemma is a SIFE. I also noted that this result relied on the existence of punishment strategies that gave both players lower payoffs than in the Cooperative outcome. Definition 7 lists variants of the condition that such *supporting* strategies exist:

---

[12] I do not require this property when the Pareto frontiers are arbitrarily close to horizontal or vertical, where the costs of kindness are either trivial or exorbitant.

Definition 7:

The supergame strategies (σ_1,σ_2) *support* the payoff pair $(\hat{\pi}_1,\hat{\pi}_2)$ if, for i = 1,2, j ≠ i, and every strategy $\gamma_j \in \Sigma_j$, (σ_i,γ_j) yields a path of (expected) payoffs (π_1^t,π_2^t) such that: a) for all $\tau > 0$, $\sum_{t=1}^{\tau} \pi_i^t/\tau \leq \hat{\pi}_i$, and b) $\text{Lim}_{\tau\to\infty} \sum_{t=1}^{\tau} \pi_j^t/\tau \leq \hat{\pi}_j$, and for i = 1, 2, the slope of $\Pi^P(\sigma_i)$ is bounded in the sense of Assumption 6.

i) An outcome (a_1,a_2) in the one-shot game is *supported* if there exist supergame strategies (σ_1,σ_2) that support the payoffs $(\pi_1(a_1,a_2),\pi_2(a_2,a_1))$.

ii) An outcome (a_1,a_2) in the one-shot game is *strictly supported* if there exists $\varepsilon > 0$ and supergame strategies (σ_1,σ_2) such that (σ_1,σ_2) support the payoffs $(\pi_1(a_1,a_2)-\varepsilon,\ \pi_2(a_2,a_1)-\varepsilon)$.

iii) An outcome (a_1,a_2) is *strictly not supported* if, for either i = 1 or i = 2, there exists $\varepsilon > 0$ such that for all σ_i, σ_i does not support $(\pi_1(a_1,a_2)+\varepsilon,\ \pi_2(a_1,a_2)+\varepsilon)$.

An outcome (a_1,a_2) is supported if each player has a repeated-game strategy that both guarantees that the other player gets a worse average payoff than she gets in the outcome (a_1,a_2), and that he himself will never be forced, even in the short run, to get a higher average payoff than he gets from (a_1,a_2). The outcome (Cooperate,Cooperate) in the Prisoner's Dilemma is supported by each player by the strategy tit-for-tat; *any* response by the other player to this strategy will yield each player less than or equal to the payoffs from cooperation. Note that the grim strategies "Defect forever" do *not* support the cooperative payoffs. If player 1 played Defect forever, then player 2 could respond by playing Cooperate forever, which gives player 1 a payoff higher than he gets from (Cooperate,Cooperate).

Proposition B2 establishes that almost all supported mutual-max outcomes are SIFEs:

Proposition B2:

Suppose the kindness functions meet Assumptions 1-6. Then every supported, strictly positive, NPS mutual-max outcome (a_1,a_2) is a SIFE.

Proof:

Let σ_i be some supergame strategy by player i that supports the outcome (a_1,a_2). Then in a T-incremental game, consider the strategies (γ_1^*,γ_2^*) which are defined as follows. Each player i plays a_i in period 1, and plays a_i in period t if player j has player a_j in each period 1,...,t-1. If in any period

t player j deviates from a_j, then player i responds by playing the supergame strategy σ_i from period t+1 to T.

The strategies (γ_1^*, γ_2^*) constitute a strictly positive NPS mutual-max outcome in the repeated game: it is mutual-max because any deviation by player 2 (say) will yield player 1 no higher payoff than if they play (γ_1^*, γ_2^*); it is strictly positive because they are playing the strictly positive outcome (a_1, a_2) in the final period; it is NPS because (a_1, a_2) is NPS, and no deviation by a player can give that player a higher payoff.

It remains to show that, as $T \to \infty$, the set $\Pi^P(\gamma_j^*)$ for each j becomes arbitrarily small, with its slope bounded. It's slope is bounded by the assumption on $\Pi^P(\sigma_j)$ for j = 1,2. $\Pi^P(\gamma_j^*)$ becomes arbitrarily small, because for any supporting strategy σ_j, it is the case that for all $\mu > 0$, there exists $N > 0$ such that playing the first N periods of σ_j guarantees that player i's average per-period payoff will be less than or equal to $(1+\mu)\cdot\pi_i(a_1, a_2)$. As $T \to \infty$, this implies that the highest payoff that player i can get from deviating becomes arbitrarily small. Assumption 6 therefore implies that there exists a $\bar{T}$ such that, for all $T \geq \bar{T}$, these strategies constitute a fairness equilibrium to the game $G^T(1/T)$. Q.E.D.

The intuition for Proposition B2 resembles that presented in the Introduction: If players threaten to punish deviations from an outcome with supporting strategies, then the potential payoff gains to a player from deviating are minimal. If the stage-game outcome is a mutual-max outcome, and deviations from this outcome are punished by supporting strategies, then the overall strategies also form a mutual-max outcome.[13] Moreover, the Pareto frontier of each player's choice set will be small, because the fact that deviations are punished by supporting strategies means that the Pareto frontier for each player given the overall strategies is simply her Pareto

---

[13] It is here that we need for punishments that are guaranteed to hurt both players. Suppose a deviation by player 1 (say) leads to a punishment strategy by player 2 in which player 1 could respond with a strategy that yields player 2 a higher payoff than does the putative equilibrium. Then player 2 may consider it unkind of player 1 not to "deviate" and then choose a strategy to help player 2 after the deviation. For instance, the "grim" strategies of playing Defect forever following a deviation do not support the cooperative outcome in repeated Prisoners' Dilemma, because either player could help the other by Defecting once and Cooperating from then on. Indeed, when overall payoffs are small enough, we know from Proposition A4 that cooperation via the grim strategies will not constitute a fairness equilibrium to the incremental game.

frontier of the (scaled-down) stage-game payoff set. The fact that the mutual-max outcome is strictly positive means that the Pareto frontier is not a singleton. All this implies that the overall strategies constitute a strictly positive mutual-max outcome where the material benefits of deviating are small; thus, they constitute a fairness equilibrium.

Proposition B2 is usefully contrasted with Proposition A5, which states that if a mutual-max outcome is not a Nash equilibrium then it is not a fairness equilibrium in a game with very large material stakes. Proposition B2 shows that, if the mutual-max outcome is supported, then it can happen in every period as part of a fairness equilibrium of the incremental version of the game *with arbitrarily large overall payoffs.*

Because Propositions A1 and B1 together imply that all mutual-max Nash equilibria are SIFEs, Proposition B2 implies that that only if exactly one player is maximizing his own material payoffs might a supported mutual-max NPS outcome not be a SIFE. Thus, "typically" a supported mutual-max NPS outcome is a SIFE. Proposition B2 can be usefully applied, for instance, to games with the strategic structure of the Prisoners' Dilemma or Chicken, because tit-for-tat strategies (where players start out playing C, and play C in period t+1 if and only if the other player played C in period t) support the cooperative outcome in each game if the cooperative outcome is efficient. Proposition B2 implies therefore that (Cooperate,Cooperate) in Prisoners' Dilemma and (Chicken,Chicken) in Chicken are SIFEs.

Proposition B2 says that the "supportability" of a mutual-max outcome is sufficient for it to be a SIFE. Is it necessary? The answer to this appears to be no; there exist kindness functions meeting all the general assumptions such that unsupported mutual-max outcomes can be SIFEs. Proposition B3 indicates, however, that supportability is (nearly) necessary to establish that a NPS strictly positive mutual-max outcome is a SIFE for *all* kindness functions meeting the general assumptions:

Proposition B3:

If a strictly positive mutual-max outcome (a_1,a_2) is strictly not supported, then there exist kindness functions meeting Assumptions 1-6 such that (a_1,a_2) is not a SIFE.

Proof:

By Proposition B6 below, we know that there are kindness functions meeting Assumptions 1-6 such that, for all T, there is no strictly positive

fairness equilibrium yielding payoffs $(\pi_1(a_1,a_2),\pi_2(a_1,a_2))$. But a strictly positive outcome (a_1,a_2) can occur in the last period of a repeated game *only* as part of a strictly positive fairness equilibrium, because an outcome is strictly positive only if each player has the potential of increasing his own material payoffs while hurting the other player; if $f_i \leq$ for either player, the other player would therefore deviate in the last period. This proves that (a_1,a_2) cannot be a SIFE. Q.E.D.

Whereas Proposition B2 shows that a class of supported mutual-max outcomes can be SIFEs for *all* kindness functions, Proposition B3 shows that strictly unsupported outcomes will not be SIFEs for *some* kindness functions. The kindness functions for which unsupported outcomes are not SIFEs do not appear to be particularly exotic or extreme. The proof merely invoked the possibility that the kindness functions were such that each player 1) would not be willing to sacrifice too much of their own payoffs to help the other player, even if positively disposed to the other player, and 2) would not be willing to sacrifice at all unless the other player sacrifices some minimal amount for her. Such kindness functions can rule out cooperation for strictly unsupported outcomes, because such outcomes involve at least one player either sacrificing a non-trivial amount to help the other player, or refusing to "deviate" to increase the other player's material payoff by a non-trivial mount.

Propositions B1-B3 pertain to those outcomes that can possibly occur *every* period of a fairness equilibrium in an incremental game. Further results can be developed if we consider the set of *near* SIFEs. An outcome is a *near SIFE* if it occurs an arbitrarily large percentage of the time in some fairness equilibrium of the T-incremental game as T becomes large.

Definition 8:

An outcome (a_1,a_2) is a *Near SIFE* if, for all $\alpha \in (0,1)$, there exists $\bar{T}$ such that, for all $T \geq \bar{T}$, there exists a fairness equilibrium in the game $G^T(1/T)$ in which (a_1,a_2) occurs for more than αT periods.

Proposition B4 characterizes a class of non-mutual-max outcomes that can be near SIFEs:

Proposition B4:

Suppose that there exist supergame strategies (σ_1,σ_2) that strictly

support *both* the NPS outcome (a_1,a_2) *and* some strictly positive, mutual-max NPS outcome (d_1,d_2). Then for all kindness functions meeting Assumptions 1-6, (a_1,a_2) is a near SIFE.[14]

Proof:

Let σ_i be some supergame strategy by player i that supports the outcome (a_1,a_2) and the outcome (d_1,d_2). Then in a T-incremental game, consider the strategies (γ_1^*,γ_2^*) defined as follows. Each player i plays a_i in period 1, and plays a_i in period $t < t^*$ if player j has player a_j in each period $1,\ldots,t-1$. From periods t^* to T, each player i plays d_i in period t if player j has player a_j in each period $1,\ldots,t^*-1$ and has played d_j in each period $t^*,\ldots,t-1$. If in any period t player j deviates from a_j or d_j (whichever is relevant), then player i responds by playing σ_i forever.

Claim: There exists a positive integer $\mathcal{T}$ such that the above strategies are a NPS, strictly positive mutual-max outcome to the game $G^T(1/T)$ for all T and t^* such that $T - t^* \geq \mathcal{T}$. This is because, so long as $T - t^*$ is made large enough, any deviation by player j in periods $1,\ldots t^*-1$ will induce the strategy such that player i will be hurt relative to the non-deviation outcome. Player j cannot improve player i's payoff in any period $t^*,\ldots,T$ because (d_1,d_2) is a mutual-max outcome. These strategies are strictly positive in $G^T(1/T)$ because (d_1,d_2) is assumed to be strictly positive, so that player j playing d_j in period T by definition involves sacrifice.

Now consider the strategies (γ_1^*,γ_2^*) defined for each T with respect to t^* such that $T - t^* = \mathcal{T}$. Then as $T \longrightarrow \infty$, we can see that these strategies constitute a fairness equilibrium using the same arguments as in the proof of Proposition B2. Moreover, as $T \longrightarrow \infty$, the strategies yield the outcome (a_1,a_2) all but $\mathcal{T}/T$ proportion of the time, so that these strategies yield a near SIFE. Q.E.D.

The outcomes identified by Proposition B4 as near SIFEs are those that are supported by some pair of strategies that support some mutual-max outcome. Players can then play such outcomes most of the game, and then in the last few periods play the mutual-max outcomes. Just as in Proposition B2, the resulting

---

[14] Although my emphasis here has been on particular outcomes in the stage game G, Proposition B3 straightforwardly generalizes to any feasible payoff pair that is strictly supported, whether or not that payoff pair corresponds to particular strategies in the one-shot game.

overall strategies constitute a mutual-max outcome in the overall game. Note that Proposition B4 requires the stronger criterion of *strict* support, because deviations from non mutual-max outcomes can strictly improve both players' payoffs.

I have not shown that *only* outcomes with the payoffs described in Proposition B4 are near SIFEs. I have no complete characterization theorems along these lines. Proposition B5 indicates, however, that there exist no fairness equilibria yielding either player below his minmax payoffs in incremental games if the scale of the overall game is made arbitrarily large. With the natural additional assumption, Proposition B5 is a trivial analog of Proposition A5, because any Nash equilibrium in any repeated game must yield a player at least his minmax payoff. As we assume that the total material stakes of the repeated game become arbitrarily large, this means that every fairness equilibrium should yield the player at least (arbitrarily close to) his minmax payoff.[15] Let (π_1^*, π_2^*) be the players' minmax payoffs.

Proposition B5:

Suppose that the kindness functions meet Assumption 7. Then for all outcomes (a_1, a_2) that yield at least one of the players less than his minmax payoff there exists an $\bar{X}$ such that for all $X > \bar{X}$, (a_1, a_2) is not a near SIFE in $G(X)$.

Proof:

Assume (without loss of generality) that $\pi_1(a_1, a_2) < \pi_1^*$. Then there exists $\kappa > 0$ such that $\pi_1(a_1, a_2) = \pi_1^* - \kappa$. As the scale, X, of the game becomes arbitrarily large, then player 1 can improve his material payoff by arbitrarily much. Assumption 7 tells us that there exists some $\bar{X}$ such that for all $X > \bar{X}$, for *all* T, no fairness equilibrium to the game $G^T(X/T)$ can yield player 1 the payoff $\pi_1(a_1, a_2)$. Q.E.D.

Proposition B5 tells us that certain outcomes, including those that are fairness equilibria for small-payoff, one-shot games, are not fairness

---

[15] The Proposition invokes Assumption 7 presented in the next section, because neither Assumption 3 nor Assumption 5 imply that one cannot make the scale of the kindness functions themselves arbitrarily large as we increase the number of periods T. Using overkill, part 1 of Assumption 7 places a universal bound on the value of the kindness functions.

equilibria in incremental games if overall stakes are large. Consider, for instance, the outcome (Dare,Dare) in *Chicken*. For small enough X, we know that this is a fairness equilibrium in the one-shot game. (Dare,Dare) cannot too often occur as part of any fairness equilibrium in incremental *Chicken*, however, when total stakes are very high. Because (Dare,Dare) yields each player below his minmax payoff, players cannot punish deviations, so that a player could gain *each* period by deviating from this outcome. The cumulative benefit of deviating will therefore be substantial so long as the overall material payoffs are large, so that a (bounded) taste for revenge won't induce a player to keep playing Dare.

While Propositions A1 and B1 combine to mean that all mutual-min Nash equilibria (such as (Defect,Defect) in the Prisoners' Dilemma)) are SIFEs. The following lemma establishes that any non-Nash mutual-min outcome in which neither player plays all of his actions with positive probability yields at least one player below his minmax payoff. Thus, Proposition B4 implies that almost all *non*-Nash mutual-min outcomes will not be SIFEs for large enough material payoffs.

Lemma 1:

Suppose that (a_1,a_2) is a mutual-min outcome in which neither player plays all his actions with positive probability is yielding payoffs (π_1^m,π_2^m). If player i is not maximizing his own payoffs given a_j, then $\pi_i^m < \pi_i^*$.

Proof:

Suppose player 1 is not maximizing his own payoffs in a mutual-min outcome. Then player 1 has some strategy d_1 such that $\pi_1(d_1,a_2) > \pi_1^m$. Let $\pi_1'' \equiv \pi_1(d_1,a_2)$. Also, because we are considering only generic games, and (a_1,a_2) is a mutual-min outcome, then for every pure strategy d_2 that is not played with positive probability in the strtegy a_2 by player 2, $\pi_1(a_1,d_2) > \pi_1^m$. But $\pi_1' \equiv \min_{d_2 \in A2,\, d_2 \neq a_2} \pi_1(a_1,d_2) > \pi_1^m$.

Let $\pi_1''' \equiv \min_{a \in A2} \pi_1(d_1,a)$. π_1''' may be less than π_1^m. But because player 1 has the option of mixing between strategies a_1 and d_1, for any strategy $a \in S_2$ that puts probability weight $1-\delta$ on the strategy a_2, player 1's best response will yield him payoff

$$\pi_1(\delta) \geq \text{Max}[[(1-\delta)\cdot\pi_1^m + \delta\cdot\pi_1'],[(1-\delta)\cdot\pi_1'' + \delta\cdot\pi_1''']].$$

Then, because the game is finite, $\text{Min}_{\delta \in [0,1]}\ \pi_1(\delta) > \pi_1^m$. By definition, however, $\pi_1^* \equiv \text{Min}_{\delta \in [0,1]}\ \pi_1(\delta)$, so that this proves that $\pi_1^m < \pi_1^*$. Q.E.D.

Lemma 1 establishes that "typically" mutual-min outcomes are not near SIFEs when material payoffs are large. But some mutual-max outcomes are also not near SIFEs when material stakes are high, because nothing guarantees that a mutual-max outcome does not yield players less than their minmax payoffs. Such is the case in *Gift* (see Figure 1C). The outcome (Gift,Gift) is a mutual-max outcome, and is a fairness equilibrium in the one-shot game if the payoffs are small enough. Although it is inefficient in material terms, it is a pretty natural outcome in one-shot games: if players come to believe in the outcome with common knowledge, then it is unambiguously nice to play Gift rather than No Gift, so that for precisely the same reasons that players won't deviate from the (Cooperate,Cooperate) outcome in the Prisoners' Dilemma, they won't do so here.[16]

Proposition B5 states that this outcome is not sustainable in incremental games if the overall payoffs are high. As with (Dare,Dare) in Chicken, a deviation from the (Gift,Gift) outcome cannot be punished by the other player. Thus, when overall material payoffs are high, the material benefits of deviating may outweigh the fairness benefits from not deviating.

In comparing the class of outcomes considered in Proposition B4 to those outcomes considered in Proposition B5, it is clear that not all outcomes are included. In particular, I have not provided a result about whether non-supported outcomes with payoffs above the minmax levels can be near SIFEs. While indeed some non-supported outcomes can be near SIFEs, Proposition B6 establishes that for any *strictly* non-supported outcome of a game, there exist *some* kindness functions meeting Assumptions 1-6 such that the payoffs from this outcome are not approximated by any *strictly positive* fairness equilibrium of the incremental version of the game.

Proposition B6:

Suppose the outcome (a_1,a_2) is strictly not supported. Then there exists kindness functions meeting Assumptions 1 - 7 such that (a_1,a_2) is not a strictly positive near SIFE.

---

16 Pre-game communication would perhaps be likely to lead players to coordinate on the materially efficient outcome. But while the outcome (Gift,Gift) is inefficient in material terms, it may be the players' best outcome in terms of their overall utility; it is unclear what is the most behaviorally realistic assumption in such a situation.

Proof:

Suppose (without loss of generality) that there exists $\kappa > 0$ such that for all strategies σ_2 of the infinitely repeated games, there exists σ_1 such that the average per-period payoffs from (σ_1,σ_2) yields a payoff of $\pi_i \geq \pi_i(a_1,a_2) + \kappa$ for either $i = 1$ or $i = 2$. Note that if there does not exist any strategy to the infinitely repeated game that guarantees payoffs less than this, then there will not be any such strategy for any finite T either (otherwise player 2 could simply repeat that finite strategy infinitely often).

Simply choose a kindness function f_1 independent of the strategic structure of the game for which, given the payoff scale of payoff set $\Pi(\sigma_2)$ implied by the payoff scale of the game (and thus of all the incremental versions of that game), the following two conditions hold:

a) player 1 would be unwilling to choose a payoff of $\pi_1(a_1,a_2)$ over a payoff of $\pi_1(a_1,a_2) + \kappa$ for any value of f_2, and

b) $f_1 < 0$ for every choice of payoff $\pi_2 < \pi_2(a_1,a_2) + \kappa/2$ for any payoff set Π that includes the payoff $\pi_2(a_1,a_2) + \kappa$.

With such a kindness function, Player 1 would deviate from a_1 if player 2 was playing a punishment strategy σ_2 such that $\Pi(\sigma_2)$ includes a payoff $\pi_1 > \pi_1(a_1,a_2) + \kappa$. If $\Pi(\sigma_2)$ included a payoff $\pi_2 > \pi_2(a_1,a_2) + \kappa$, the $f_1 < 0$. By Proposition A2, it must be that $f_2 \leq 0$, so that these cannot be a strictly positive fairness equilibrium to the overall game. Q.E.D.

Some unsupported outcomes could be positive fairness equilibria in incremental games, of course, depending on the kindness functions. But Proposition B6 establishes that no such outcome is a SIFE or near SIFE for *all* kindness functions meeting the general assumptions. I remind the reader that this contrasts with Proposition B2, where it was shown that for *any* kindness function meeting the general assumptions, supported mutual-max outcomes such as cooperation in the Prisoner's Dilemma are SIFEs.

4. Fairness in Replicated Games

In this section I present some results regarding fairness equilibria in *replicated* games, which are what I call repeated games whose scale *each period* replicates a given one-shot game. Formally, I say that the game $G^T(1)$

constructed from the one-shot game G is a T-*replicated game*.

Proposition B1 of the previous section states that for every fairness-equilibrium outcome in a game, there exists a fairness equilibrium in the incremental version of that game which yields that outcome each period. The same result does *not* hold for replicated games. At the end of this section, I return to formalizing the very strong additional assumptions that guarantee that an analog of Proposition B1 holds for replicated games. But no plausible assumptions will imply that the analogs of Propositions B2 and B3 will hold for replicated games; there may not exist any fairness equilibria in replicated games where a supported mutual-max outcome happens most periods. The cooperative outcome in the Prisoners' Dilemma, for instance, need not be a part of fairness equilibrium in a replicated game.

This section will focus instead on results emphasizing that, as the number of periods becomes arbitrarily large, the set of fairness equilibria in replicated games will look much like the set of Nash equilibria. I begin with the result that, in replicated games, there will be no fairness equilibria in which players get much below their minmax payoffs. This result merely reproduces Proposition B4 of the previous section.

To obtain this result, I need to assume that fairness payoffs do not increase as much as material payoffs as we replicate a game. Like Assumption 3, this captures the idea that for very large marginal material stakes, players' behavior is dominated by material self-interest. Assumption 3 is not sufficient here because replication changes the strategic structure of the game. With some overkill, I strengthen Assumption 3 by assuming a universal bound for all games on the values of the kindness functions as well as the relative slope of kindness functions along the Pareto-frontier of players' payoff opportunity sets.[17]

---

[17] This last component, as formalized by part 2 of Assumption 7, limits the proportion of the range of kindness values consistent with Pareto efficiency that can come from a small change sacrifice by player i in his material well-being. As such, it is similar in spirit to Assumption 1, Part 2b. Assumption 7 says roughly that changes in fairness payoffs arising from changes along the Pareto-frontier must be approximately proportional to the material sacrifice made; Assumption 6 specifically insists, by contrast, that if the Pareto-frontier is negligible relative to the entire payoff-opportunity set, changes in the fairness payoffs can be much greater for material sacrifices along the Pareto-frontier than for comparable material sacrifices in choices among inefficient outcomes.

Assumption 7:

There exists $N_7 > 0$ and positive-valued function $k(c)$ such that for all games and for each $i = 1, 2$,

1) For all a_i, b_j, $f_i(a_i, b_j) \in [-N_7, N_7]$; and

2) For all a, a', α, $\alpha' \in S_i$ and $b \in S_j$ such that

a) $\pi_j(a_i, b_j)$, $\pi_j(a_i', b_j)$, $\pi_j(\alpha_i, b_j)$, $\pi_j(\alpha_i', b_j) \in [\pi_j^l(b_j), \pi_j^h(b_j)]$, and

b) $\pi_j(a, b) - \pi_j(a', b) = \pi_j(\alpha, b) - \pi_j(\alpha', b)$,

$$[f_i(a, b) - f_i(a', b)]/[f_i(\alpha, b) - f_i(\alpha', b)] \leq N_7.$$

Proposition C1 shows that, with this bound on fairness payoffs, in replicated games with an arbitrarily large number of periods, self interest will guarantee that neither player gets a per-period payoff (much) less than his minmax payoff:

Proposition C1:

Suppose the kindness functions meets Assumptions 1-7. Then for all $\varepsilon > 0$, there exists T^* such that for all $T > T^*$, every FE of the game $G^T(1)$ is such that each player i is getting average per-period payoff greater than $\pi_i^* - \varepsilon$.

Proof:

Given the bound on the fairness payoffs, this result is a trivial extension of the proof of Propositions B4 and B5, because such an equilibrium must involve a player doing arbitrarily worse materially than his minmax payoff. Q.E.D.

Proposition C1 indicates that many outcomes that are possible in one-shot games will be ruled out in replicated games. Recalling Lemma 1, it implies, for instance, that most non-Nash mutual-min outcome cannot occur nearly all of the time in replicated games.

The converse of Proposition C1 does not hold true--there may be payoffs above the minmax payoffs that do not result from any fairness equilibrium of a replicated game. But Proposition C2 establishes that a slightly modified version of the Nash folk theorem for finitely repeated games holds for fairness equilibrium in replicated games.[18] In particular, so long as there

---

[18] See Benoit and Krishna (1987) for the statement of the Nash folk theorem for finitely repeated games.

exists a strict and strictly negative Nash equilibrium that yields all players above their minmax payoffs, all outcomes that are above the minmax payoffs can be part of a fairness equilibrium in replicated games:

Proposition C2:

Suppose that there exists a strict and strictly negative Nash equilibrium (d_1,d_2) with payoffs that exceeds the minmax payoffs for each player. Then for any kindness functions meeting Assumptions 1 - 7, for every $\varepsilon > 0$, and for every outcome (a_1,a_2) yielding per-period payoffs $\pi_i(a_1,a_2) > \pi_i^*$ for each player, there exists T^* such that for all $T > T^*$, there exists a fairness equilibrium of the game $G^T(1)$ where $|\pi_i(\sigma_1,\sigma_2) - \pi_i(a_1,a_2)| < \varepsilon$ for each player i.

Proof:

Consider the game $G^{K^2}(1)$, and the strategies (σ_1,σ_2) as defined as follows. Player i plays a_i in each period t from 1 to K^2-K iff the other player j has played a_j in periods 1 to t-1. If player j does not play a_j in any period t from 1 to K^2-K, then player i minmaxes j from periods t+1 to K^2. If player j plays a_j in all periods from 1 to K^2-K, then player i plays d_i in the last K periods.

Claim 1: For K large enough, this constitutes a Nash equilibrium. This is because, since (d_1,d_2) is a Nash equilibrium, clearly neither player gains from deviating in periods K^2-K+1 to K^2. A deviation before period K^2-K from (a_1,a_2) may lead to a one-shot gain. But for K large enough, and given the fact that $\pi_i(d_1,d_2) > \pi_i^*$, this is clearly a Nash equilibrium.

Claim 2: For i = 1,2, $\text{Lim}_{K\to\infty}$ $\text{Min}[\pi_i^h(\sigma_j)-\pi_i^l(\sigma_j),\pi_j^h(\sigma_j)-\pi_j^l(\sigma_j)] = \infty$. That is, the Pareto frontiers of $\Pi(\sigma_2)$ and $\Pi(\sigma_1)$ become arbitrarily large as K becomes arbitrarily large. This is so because, given that (d_1,d_2) is a strictly negative Nash equilibrium, there must be a deviation by each player i that helps the other player. Because the Nash equilibrium is strict, this deviation must hurt player i. Therefore, there exists a deviation by player i in each of the last K periods that would give him a lower payoff and player j a higher payoff. As K becomes large, the Pareto-frontier must therefore become infinite.

Claim 3: $\text{Lim}_{K\to\infty}\text{Max}_{\sigma_i'}$ $|f_i(\sigma_i',\sigma_j)|/\pi_i(\sigma_i',\sigma_j) = 0$. This is because as the Pareto frontier becomes larger and larger, Assumption 7 says that the maximal fairness gains of a deviation are bounded, and, because the slopes for

these Pareto-frontiers are bounded for all K, Assumption 7 says that the player cannot get a significant percentage of that fairness benefit without sacrificing a significant percentage of the available material costs.

Claim 3 implies that there exists $\hat{T}$ such that for all $T > \hat{T}$, $\text{Max}_{\sigma_i'} |f_i(\sigma_i', \sigma_j)| / \pi_i(\sigma_i', \sigma_j) < 1/N_7$, where N_7 is the bound on f_2 from Assumption 7. For all such T, (σ_1, σ_2) is therefore a fairness equilibrium.

Q.E.D.

The proof of Proposition C2 is much like the proof of the Nash folk theorem for finitely repeated games. Players support any feasible outcome for most of the game by threatening minmax punishments for deviation. At the end they play a Nash equilibrium; to make sure that they don't deviate earlier, it must be that the payoffs from this Nash equilibrium exceed the minmax payoffs. The only additional step in applying this to fairness equilibrium is to show that neither player's fairness payoffs are great enough to induce a deviation. But since the Pareto-frontiers are becoming very large, Assumption 7 establishes that this is the case.

None of the results above guarantee that, if the cooperative outcome is a fairness equilibrium in the one-shot Prisoners' Dilemma of a given scale, it is a fairness equilibrium in the replicated Prisoners' Dilemma. This turns out not to be true except for a restricted class of mutual-max outcomes, and then only if one makes substantially stronger assumptions about the kindness functions. I provide conditions for such a result in the Appendix.

5. Conclusion

Some experiments on repeated games exhibit a pattern of decline in levels of cooperation from the first to the last period, with levels of cooperation in the final period being low but not zero.[19] Among the interpretations for the observed decay in cooperation in the repeated Prisoners' Dilemma or repeated public-goods games are two that are manifestly inconsistent with the

---

[19] Confusing the interpretation of these experiments somewhat is the fact that the patterns do not as much as we might expect differ from experiments with *non*-strategic repeated play, where players should not logically anticipate that their behavior in early periods will much affect their environment in later periods. For a good survey of some of the literature on one-shot and repeated-game cooperation, see Dawes and Thaler (1988).

assumptions of my model. First, I have ignored strategic uncertainty: It may be that players are uncertain whether they are playing the "nice" or the "mean" equilibrium. While my discussion of the repeated Prisoners' Dilemma concentrated on the issue of whether full cooperation was possible, the fully non-cooperative outcome will always also be a fairness equilibrium, and players may be uncertain which equilibrium they are playing. By focusing only on equilibrium rather than non-equilibrium analysis, my model misses out on the dynamics of strategic uncertainty and learning over time.

A second possibility omitted from my model is incomplete information. I have assumed that the existence and precise form of fairness preferences are common knowledge to the players. One interpretation of the experimental evidence is that each player may be nervous that she is not playing with other fair players. This nervousness is well justified by the experiments themselves, in the sense that many players in one-shot Prisoners' Dilemmas do not behave cooperatively, and some of these seem to be doing so despite believing others will cooperate.

Uncertainty over players' true preferences is reminiscent of the model in Kreps *et al* (1982), whose results seem to match the experimentally observed decay in cooperation. They show that substantial cooperation is possible in the repeated Prisoners' Dilemma if one assumes that players perceive that there is a small probability that one of them is a "crazy" type. My model is analogous to theirs insofar as it shows that a small departure from standard assumptions can lead to cooperation. The models differ in the end play: because Kreps et al posit only a small likelihood that players aren't purely self-interested, play in the final periods of the game will be predominantly non-cooperative.[20]

Obviously, the two models are complementary. Indeed, the naturalness of positing that the most likely "crazy" types are tit-for-tatters seems to derive in part from the same assumptions incorporated into my model--that people tend to be "reciprocal altruists" who reward good behavior and punish

---

[20] A second potential difference is that my results only hold with small per-period payoffs. The incomplete-information explanation does not per se depend on the per-period payoffs. Of course, if the "crazy types" have preferences similar to those I assume, there will be scale effects in incomplete-information models as well. Moreover, if one interprets the existence of crazy types as boundedly rational, the hypothesis that people are more rational in decisions that have larger stakes will also indicate greater cooperation under the incomplete-information model of repeated games.

bad behavior by others. Combining the models would seem to explain the experimental evidence best--the decay of cooperation over time seems to suggest an incomplete-information, reputations model, but the relatively high levels of cooperation (in both one-shot and repeated-game settings) seems to require the existence of fairness-oriented types with significantly higher than zero probability.

Outside experimental settings, I suspect that my model's assumption of near certainty about the preferences and behavior of other people is more realistic. In real-world relationships, such as interaction among co-workers, behavior may well settle down on (either friendly or unfriendly) outcomes in which players have clear expectations about, and attitudes towards, those around them.

Appendix

In this Appendix, I provide a further restriction on the kindness function, and a sufficient condition on the strategic structure of the game, such that a fairness-equilibrium outcome in a one-shot game can be supported each period in a fairness equilibrium to the replicated game. In order to establish this result, we need to first assume that the measure of how nice or mean a player is being depends only on the size and shape of the Pareto frontier of his payoff opportunity set. Assumption 8 formalizes this:

Assumption 8:

Suppose that the Pareto frontier of $\Pi(b_j)$ in game G is an arithmetic transformation $(\pi_i, \pi_j) = (\pi_i'+m, \pi_j'+m)$ of the Pareto frontier of $\Pi(b_j')$ in the game G'. Then for all a', b_j' such that $\pi_j(a', b_j') \in [\pi_j^l(b_j'), \pi_j^h(b_j')]$, $f_i(a, b_j) = f_i(a', b_j')$ iff $\pi_j(a, b_j) = \pi_j(a', b_j') + m$.

Assumption 8 is somewhat restrictive because it requires that the measure of player's kindness on the Pareto-frontier be independent of the rest of the player's payoff opportunity set.[21] Even with this assumption, however, we

---

[21] This means, for instance, that players cannot be more forgiving of "merely selfish" behavior in situations where players have the opportunity to be very nasty than in situations where the meanest a player can do is to be selfish.

cannot reproduce Proposition B1 for replicated games. We need additionally to assume that the outcome is supported in a stronger way than we had in the previous section. In particular, it must be that each player has not simply a repeated-game strategy but a *stage-game* strategy that supports the outcome:

Definition 9:

An outcome (a_1,a_2) in G is *one-shot supported* if there exists an outcome (b_1,b_2) such that, for i = 1,2, $\pi_i(b_1,b_2) < \pi_i(a_1,a_2)$.

With Assumption 8 and Definition 9, we can state a final result:

Proposition C3:

Suppose that the kindness functions meet Assumptions 1-8. If the mutual-max outcome (a_1,a_2) is a fairness equilibrium in the game G, and is one-shot supported, then there exists a fairness equilibrium in the T-replicated game in which (a_1,a_2) is played every period.

Proof:

If the strategies b_1 and b_2 support the outcome (a_1,a_2), then let the repeated-game strategy σ_i be such that player i starts out playing a_i, and continues playing it each period so long as player j plays a_j. If player j deviates, player i plays b_i from then on.

Under Assumptions 1-8, the strategies (σ_1,σ_2) constitute a fairness equilibrium in $G^T(1)$ for any T, because, given the b_1 and b_2 one-shot support (a_1,a_2), the sets $\Pi(\sigma_2)$ and $\Pi(\sigma_1)$ have the same Pareto frontiers as do the sets $\Pi(a_2)$ and $\Pi(a_1)$ in G. Q.E.D.

Proposition C3 does not even assure that Cooperation is possible in the replicated Prisoners' Dilemma if it is a fairness equilibrium in the one-shot Prisoners' Dilemma, because (Cooperate,Cooperate) is *not* one-shot supported. However, Proposition C3 does imply that if cooperating each period is a fairness equilibrium in the *thrice*-replicated Prisoners' Dilemma, then it is a fairness equilibrium in any replicated Prisoners' Dilemma of four or more periods, because there does exist a two-period strategy by the players that supports cooperation.[22]

---

22 The difference in the supportability of cooperation between the two- and

Even when it holds in some form, Proposition C3 addresses only the issue of whether a cooperative outcome in a one-shot situation can be replicated in a multi-period play of the game; it does not indicate that cooperation can be *more* easily supported in replicated games than in one-shot. While Proposition C2 shows that repetition can help achieve cooperation in most of the cases where it helps when players do not care about fairness, replication does not seem to greatly assist fairness-oriented players to cooperate. Also note that Proposition C1 implies that replication *destroys* many of the fairness equilibria in one-shot games. While mostly replication seems to rule out negative fairness equilibria, it can also render cooperative-but-inefficient outcomes such as (Gift,Gift) in *Gift* impossible even when they are fairness equilibria in the one-shot game.

References

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

Andreoni, James (1993), "Cooperation in Public Goods Experiments: Kindness or Confusion?", University of Wisconsin Social Systems Research Institute Working Paper 93-09, March 11.

Benoit, J.P. and Krishna, V. (1987), "Nash Equilibria of Finitely Repeated Games," International Journal of Game Theory 16.

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

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

Goetze, David and Orbell, John (1988), "Understanding and Cooperation," Public Choice 57, 275-279.

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

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

---

three-shot games is not due to lack of correlating devices, etc.. One period punishment phases cannot achieve what two-period punishment phases can because they do not allow for one player punishing the other player for not behaving appropriately during that other player's punishment phase.

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

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

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

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

Kolpin, Van, "Equilibrium Refinements in Psychological Games," Games and Economic Behavior, forthcoming.

Kreps, D., Milgrom, P., Roberts, J., and Wilson, R. (1982), "Rational Cooperation in the Finitely Repeated Prisoners' Dilemma," Journal of Economic Theory 27, 245-252, 486-502.

Prasnikar, V. and Roth, A.E. (1992), "Considerations of Fairness and Strategy: Experimental Data from Sequential Games," Quarterly Journal of Economics.

Rabin, Matthew (1993), "Incorporating Fairness into Game Theory and Economics," American Economic Review 83(5), December, 1281-1302.

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