Patience or Fairness?
Analyzing Social Preferences in Repeated Games

John Duffy
Department of Economics
University of Pittsburgh
Pittsburgh, PA 15260
E-mail: jduffy@pitt.edu

Félix Muñoz-García
School of Economic Sciences
Washington State University
Pullman, WA 99164
E-mail: fmunoz@wsu.edu

January 2012

Abstract

This paper investigates how the introduction of social preferences affects players' equilibrium behavior in both the one-shot and the infinitely repeated version of the Prisoner's Dilemma game. We show that fairness concerns operate as a "substitute" for time discounting in the infinitely repeated game, as fairness helps sustain cooperation for lower discount factors. In addition, such cooperation can be supported under larger parameter values if players are informed about each others' social preferences than if they are uninformed. Finally, our results help to identify conditions under which cooperative behavior observed in recent experimental repeated games can be rationalized using time preferences alone (patience) or a combination of time and social preferences (fairness)

KEYWORDS: Prisoner's dilemma; Repeated games; Inequity aversion; Time discounting, Social Preferences.

JEL CLASSIFICATION: C72, C73, H43, D91.

1 Introduction

Inspired by a large volume of experimental evidence, there has been much recent work on *social*, as opposed to *individual*, preferences reflecting individuals' concern for fairness in the income distribution; see for instance, Fehr and Schmidt (1999) and Bolton and Ockenfels (2000). Much of this literature has examined how such social preferences might facilitate cooperation among individuals who interact in sequential move, strategic environments, such as worker-employee, principal-agent or investor-trustee relationships. However, there has been comparatively little application of social preferences to simultaneous-move games, and in particular, to infinitely repeated versions of those games.[1]

This paper contributes to the literature by investigating how social preferences might facilitate cooperation among players interacting in the canonical simultaneous-move game –the Prisoner's Dilemma– which is appropriate for the study of strategic environments with extreme competitive incentives.[2] Surprisingly, we have found no literature exploring the effects of social preferences in infinitely repeated games.[3] We analyze the interaction between time and social preferences, and provide conditions under which observed cooperation in experimental settings can be explained using either social or time preferences alone, or a combination of both types of preferences.

We first show that by introducing social preferences into a one-shot Prisoner's Dilemma game we can rationalize mutual cooperation in equilibrium, but only when both players assign a sufficiently high value to other individuals' payoffs. We then show in the infinitely repeated version of the game, how social preferences work as a "substitute" for time preferences (discounting) since higher concerns about fairness reduce the minimum discount factor necessary to support cooperative outcomes in the repeated game. Our results help to rationalize experimental observations where players cooperate under relatively low discount factors –values for which the "Folk theorem" for repeated games would *not* predict cooperation.

In addition, we investigate how equilibrium play is affected by the introduction of incomplete information about players' social preferences. We begin by analyzing a signaling version of the twice-repeated prisoner's dilemma game in which the player who is informed about his type (concern for fairness) uses first-period actions to convey or conceal his social preferences to the uninformed player. We identify a pooling equilibrium in which the informed player cooperates in the first period of the game regardless of his type. The unconcerned player chooses to cooperate, in order to "mislead" the uninformed player about his true type. If priors about types are sufficiently high,

[1] See however, Fischbacher and Gächter (2010) who use a strategy method to find direct evidence of social preferences in a linear voluntary contribution experiment that involves simultaneous decisions by groups of four players interacting repeatedly for a finite number of periods (no discounting).

[2] Note that the Prisoner's Dilemma game is strategically equivalent to a voluntary contribution or "public good" game with a finite set of actions. Thus, all our analysis is also applicable to this public good game as well. In the last section of the paper, we extend our results to a larger class of games.

[3] Montero (2007) introduces inequity aversion in the Baron and Forejohn (1989) legislative bargaining game, showing that individuals' social preferences might actually lead to *more* inequality. Intuitively, during the bargaining process the responder experiences a greater disutility from being left outside the winning coalition when he is envious than when he is not, which induces him to accept lower offers thereby increasing payoff inequality. In our model there is no such risk, which eliminates the possibility of this kind of result.

this misleading strategy induces the uninformed player to cooperate in the subsequent period, while informed players who are unconcerned about fairness defect in the second and final period.[4] Interestingly, this pooling equilibrium provides an explanation for a relatively common observation in experimental settings wherein subjects defect in the last period of interaction, despite a previous history of cooperation.[5] We then extend our analysis to infinitely repeated games with incomplete information, where we show that cooperation becomes more difficult to support relative to the case of complete information.

Finally, we analyze the set of feasible, individually rational payoffs that can be achieved when playing the infinitely repeated game and we examine how this set is affected by changes in players' social preferences.[6] In particular, we find that the set of feasible, individually rational payoffs *shrinks* as individuals become more concerned about fairness.[7] Interestingly, this implies a potential confusion in the experimental literature on the source of observed cooperation on repeated games. Indeed, it suggests that such cooperation may not be due to players' high discount factors alone, but could instead arise from a combination of individuals' time and social preferences. We suggest a method for inferring whether the mechanism supporting cooperative behavior in the infinitely repeated game can be explained using time preferences alone or when reliance must be placed on both time and social preferences in order to rationalize cooperative play.

Related literature. Our results are related to those in Kreps et al. (1982), who consider the role of informational asymmetries about players' types in the finitely repeated Prisoner's Dilemma game. Specifically, in their model a "rational" player may assign some probability to the possibility that his opponent "irrationally" plays a conditionally cooperative, tit-for-tat strategy, showing that there is a sequential equilibrium of the finitely repeated game in which the "rational" player imitates the "irrational" player by also playing tit-for-tat. Similarly, in this paper, we demonstrate that the existence of social preferences may lead to cooperation among players in situations where cooperation would not exist among self-interested players. However, we develop our result from

[4] Healy (2007) identifies a similar result in the context of finitely-repeated gift-exchange games where the firm manager does not observe the worker's type (either reciprocator or selfish). The equilibrium in which informed players "mislead" uninformed players can, nonetheless, be supported under different parameter conditions in the simultaneous and sequential versions of the game. Duffy and Munoz-Garcia (2011) elaborate on this difference.

[5] See for instance, Selten and Stoecker (1986) and Andreoni and Miller (1993) for the prisoner's dilemma game, McKelvey and Palfrey (1992) for the centipede game, Camerer and Weigelt (1988), Brandts and Figueras (2003) for the borrower-lender game, and Anderhub, Engelmann and Güth (2002) for the finitely-repeated trust game. Importantly, this informational explanation for cooperative behavior is quite distinct from bounded rationality arguments in which some players misunderstand strategic incentives.

[6] In this sense, our paper is also related to Rabin (1997), who analyzes the introduction of concerns about fairness in *finitely* repeated games under complete information. Similarly to Rabin (1997), we find that players' preferences for fairness facilitate their coordination to play equilibrium outcomes with Pareto superior payoffs. However, Rabin's (1997) results can only be supported when the per-period payoffs are negligible, and he does not investigate the substitutability between social and temporal preferences.

[7] This result provides an effect opposite to that shown by Abreu et al. (1990), wherein the set of equilibrium payoffs in the infinitely repeated game *weakly increases* with increases in the discount factor. Chade *et al.* (2008) present a result in line with that in our paper for hyperbolic (present-biased) preferences where, in the case of the Prisoner's Dilemma game, an increase in players' discount factor expands the set of equilibrium payoffs whereas an increase in players' hyperbolic preferences shrinks this set. Similarly, Yamamoto (2010) demonstrates that the set of equilibrium payoffs does not necessarily expand in the discount factor if players cannot observe a public randomization.

a simpler, behavioral primitive —social preferences, specifically inequity aversion— which is supported by strong empirical evidence; see for instance, Fehr and Fischbacher (2002) or Camerer (2003).[8] Further, we also develop our result in the infinitely repeated Prisoner's Dilemma game (Kreps et al. (1982) only study the finitely repeated version), and we relate fairness concerns to time preferences. The study of cooperation in infinitely repeated Prisoner's Dilemma games has recently become the subject of much study by experimentalists (see, among others, Dal Bó (2005), Normann and Wallace (2006), Aoyagi and Fréchette (2009), Duffy and Ochs (2009), Camera and Casari (2009) and Dal Bó and Fréchette (2011), Blonski et al. (2011) and Fudenberg et al. (2012)) and so an understanding of the mechanisms by which cooperation can be sustained in such environments is both important and timely.

The paper is organized as follows. In the next section we introduce the model. Section three then analyzes equilibrium behavior under complete information, while section four extends our results to incomplete information contexts. Similarly, section five extends our analysis to more general simultaneous-move games —including games with asymmetric payoff structures— and to a more general class of social preferences. Section six elaborates on the set of feasible payoffs and the potential confound we might observe between social and time preferences under certain parameter values. Section seven concludes.

2 Model

Consider the stage game shown below. To make this a Prisoner's Dilemma game, both players' payoffs must satisfy the restriction $b > a > d > c$. In that case, both players' best response in the one-shot game is to choose D, "defect," either when the other player chooses C, "cooperate" (given that $b > a$), or when the other player defects as well (since $d > c$). Hence, the strategy profile (D,D) is the unique equilibrium of the one-shot stage game.

		Player 2 C	*Player 2* D
Player 1	C	a,a	c,b
	D	b,c	d,d

In this paper, however, we analyze players who possess Fehr and Schmidt (1999)-type social preferences, a now standard specification. (Section 5 extends our results to other social preferences). For the case of two players, Fehr and Schmidt's (1999) utility function reduces to:

$$U_i(x_i, x_j) = x_i - \alpha_i \max\{x_j - x_i, 0\} - \beta_i \max\{x_i - x_j, 0\},$$

[8] The term "social preferences" encompasses several different formulations, besides inequity aversion including preferences for reciprocity, unconditional kindness (altruism) and spiteful preferences; see for instance, Fehr and Fischbacher (2002). By "social preferences" we will mean inequity aversion as in the formulation of Fehr and Schmidt (1999).

where x_i is player i's payoff, and x_j is the payoff of his opponent (player j). The parameter α_i represents the disutility from allocations that are disadvantageously unequal for player i due to envy about player j's higher payoff, while the parameter β_i captures the disutility from allocations that are advantageously unequal for player i due to guilt over earning a higher payoff than player j. Additionally, Fehr and Schmidt (1999) assume that players' envy is always stronger than their guilt. We capture this by assuming that $\alpha_i \geq \beta_i$ and $1 > \beta_i \geq 0$.[9] We will contrast this case of "social preferences" (which we also refer to throughout as "concerns for fairness") with the more standard, self-regarding preferences where $\alpha_i = \beta_i = 0$ for all i.

Taking social preferences into account, the stage game can be reformulated as follows:

		Player 2	
		C	D
Player 1	C	a, a	$c - \alpha_1(b-c), b - \beta_2(b-c)$
	D	$b - \beta_1(b-c), c - \alpha_2(b-c)$	d,d

Notice in particular, that every player i's utility level decreases when he is either: the player with the highest payoff in the group (due to guilt), e.g., player 1 under outcome (D,C), or when he is the player with the lowest payoff in the group (due to envy), e.g., player 1 under outcome (C,D).

3 Complete information about social preferences

3.1 Stage game

In this section we briefly analyze equilibrium behavior in the one-shot Prisoner's Dilemma game under the assumption of complete information about social preferences. Section 3.2 examines players' equilibrium strategies in the infinitely repeated version of the game, whereas section 4 focuses on the incomplete information game.

Lemma 1. *In the one-shot Prisoner's Dilemma game where players have social preferences ($\alpha_i > \beta_i \geq 0$), the following strategy profiles can be supported as Nash equilibria of the game:*

1. *(D,D), if $\beta_i \leq \frac{b-a}{b-c}$ for any player; and*
2. *(C,C), (D,D) and a mixed strategy Nash equilibrium where every player i randomizes according to probability $\overline{q}(\alpha_j, \beta_j) = \frac{d-c+\alpha_j(b-c)}{a+d-c-b+(\alpha_j+\beta_j)(b-c)}$ if $\beta_i > \frac{b-a}{b-c}$ for both players.*

Hence, if at least one player has relatively low concerns about guilt, the unique Nash equilibrium

[9] Intuitively, $\alpha_i \geq \beta_i$ implies that players (weakly) suffer more from inequality directed at them than inequality directed at others. On the other hand, $\beta_i \geq 0$ means that players dislike being better off than others (this assumption rules out cases in which individuals are status seekers but serves to simplify the analysis). Finally, $\beta_i < 1$ suggests that when player i's payoff is higher than that of player j's by one unit (e.g. a dollar), player i is never willing to give up more than one unit in order to reduce this inequality. For a more detailed explanation of these assumptions, see Fehr and Schmidt (1999).

of the one-shot game, (D,D), coincides with that of the one-shot game where players have no concerns about the fairness of the payoff distribution (standard preferences); see Figure 1.

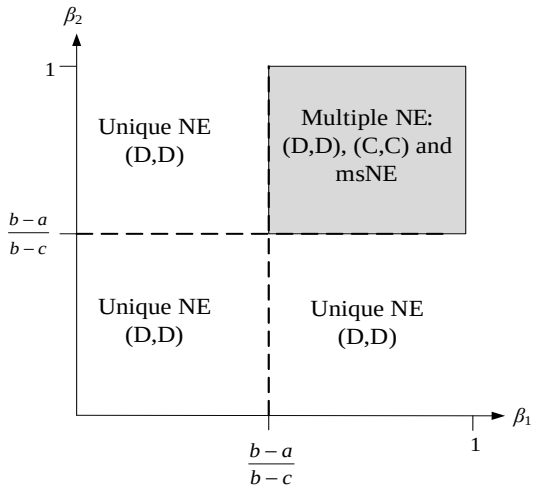

Figure 1. Equilibria in the simultaneous game under complete information.

However, when *both* individuals are sufficiently concerned about fairness —the shaded area of Figure 1— we can identify three different Nash equilibria: one in which both players defect, one in which both players cooperate, and a mixed strategy Nash equilibrium.[10] The introduction of sufficient concerns about fairness by both players thus transforms the payoff structure of the game from a Prisoner's Dilemma to a Pareto-rankable coordination game.[11]

3.2 Infinitely repeated game

Let us now focus on equilibrium strategies in the infinitely repeated version of the Prisoner's Dilemma game under social preferences and complete information.[12] We consider that every player i discounts future payoffs according to a discount factor $0 < \delta_i < 1$.

Proposition 1. *In the infinitely repeated Prisoner's Dilemma game where players have con-*

[10] Note that $\beta_j > \frac{b-a}{b-c}$ is a sufficient condition for probability cutoff $\overline{q}(\alpha_j, \beta_j)$ to satisfy $\overline{q}(\alpha_j, \beta_j) \in (0, 1)$.

[11] In particular, every player's best response is to select the same action as his opponent, but both players strictly prefer (C,C) to (D,D). Note that this best response function is similar to what Cooper et al. (1996) call "best response altruists," namely players for whom cooperate (defect) is their best response to cooperation (defection, respectively), as opposed to what Cooper et al. (1996) refer to as "dominant strategy altruists" for whom cooperation is always a best response, regardless of other players' strategies. Our results in the unrepeated game are also connected with those in Bolton and Ockenfels (2000), who allow for every individual's payoff thresholds to be private information. Unlike our model, however, their paper does not explicitly consider infinitely repeated games, and how equilibrium predictions in such a setting differ when players are symmetrically or asymmetrically informed about each others' social preferences.

[12] For simplicity, we focus on the case in which players' concerns for fairness are not extreme, i.e., $\alpha_i, \beta_i < \frac{b-d}{b-c}$ for all player i. In particular, this guarantees that the utility from reverting to the pure strategy Nash equilibrium of the stage game is still lower than that from playing the mixed strategy Nash equilibrium of the stage game.

cerns for fairness ("F" or $\beta_i > 0$), mutual cooperation can be sustained as the subgame perfect Nash equilibrium (SPNE) of the infinitely repeated game by use of the following grim-trigger strategy by every player i: start cooperating in the first period of the game, and cooperate as long as all players have cooperated in previous periods, but defect otherwise, for any discount factor $1 > \delta_i \geq \delta_i^F(\beta_i)$, where

$$\delta_i^F(\beta_i) \equiv \begin{cases} \delta_i^{NF} - \frac{\beta_i(b-c)(d-a)}{(b-d)[\beta_i(b-c)-b+d]} & \text{for any } \beta_i \leq \frac{b-a}{b-c} \\ 0 & \text{otherwise} \end{cases}$$

and $\delta_i^{NF} \equiv \frac{b-a}{b-d}$ denotes the minimal discount factor supporting cooperation in the infinitely repeated game in the case where individuals are not concerned about fairness ("NF").

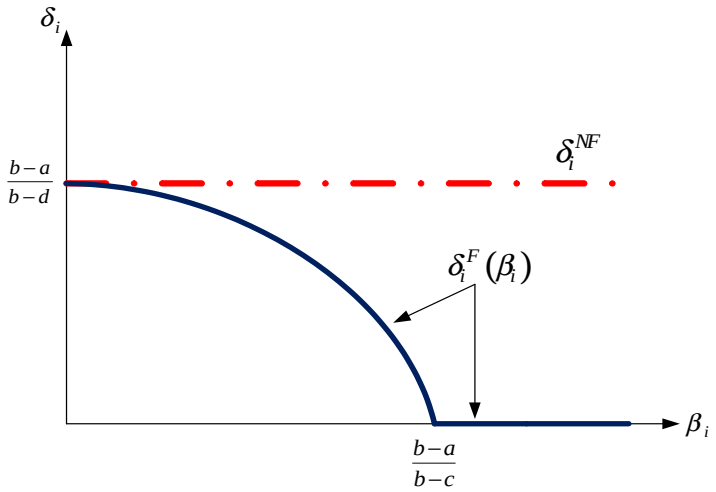

Figure 2. Discount factors $\delta_i^F(\beta_i)$ and δ_i^{NF}.

Figure 2 provides an illustration of how the minimum discount factor supporting mutual cooperation in the infinitely repeated game when players have concerns for fairness, $\delta_i^F(\beta_i)$, varies with β_i. To facilitate comparison, Figure 2 also includes the discount factor sustaining cooperation in the infinitely repeated game in the case where players are *not* concerned about fairness, δ_i^{NF}. Notice that when players do not assign any value to fairness (when $\beta_i = 0$, at the vertical intercept of Figure 2), the minimal discount factors necessary to support mutual cooperation, $\delta_i^F(\beta_i)$ and δ_i^{NF}coincide. For strictly positive values of β_i, Figure 2 can be divided into two regions. First, for relatively low concerns about fairness, $\beta_i \leq \frac{b-a}{b-c}$, every player has incentives to unilaterally deviate from the cooperative outcome since $b - \beta_i(b-c) \geq a$, as in the standard Prisoner's Dilemma game. Each player's incentive to deviate is, however, reduced by the guilt he experiences from obtaining a higher payoff than other players, $\beta_i(b-c)$. This result is illustrated by the fact that $\delta_i^F(\beta_i) \leq \delta_i^{NF}$ and that $\delta_i^F(\beta_i)$ decreases in β_i. Mutual cooperation can then be sustained under a broader set of parameter values when players possess social preferences than when they do not. For relatively high concerns about fairness, $\beta_i > \frac{b-a}{b-c}$, players have no incentives to unilaterally deviate from the

cooperative outcome since $a > b - \beta_i(b-c)$, which implies that cooperation can now be sustained for all discount factors, as illustrated in Figure 2 for $\beta_i > \frac{b-a}{b-c}$.

Hence, players' concerns about fairness make the Folk theorem for repeated games with discounting (see, e.g., Fudenberg and Maskin (1986)) applicable under a broader range of parameter values (discount factors).[13] This result can help to explain experimental findings such as those reported by Murnighan and Roth (1983, Table 4), Dal Bó (2005, Table 5) and Dal Bó and Fréchette (2011, Tables 3-4) where a small but significant fraction of experimental subjects playing an indefinitely repeated Prisoner's Dilemma game are observed to cooperate even when continuation probabilities (induced discount factors) do not support such cooperation as an equilibrium of the repeated game under standard, self-interested preferences. Our result showing that fairness concerns may "substitute" for patience can be used to rationalize such experimental observations.[14] Section six discusses how to disentangle time and social preferences as mechanisms for sustaining cooperation in experimental settings.

4 Incomplete information about social preferences

4.1 Signaling private concerns about fairness

In this section we relax the assumption of complete information about social preferences. We use a standard signaling game information structure in which one player's social preferences are commonly known while the other "informed" player's social preferences are his private information, i.e., we study the case of one-sided asymmetric information. Specifically, suppose that nature selects player i's concern for fairness, β_i, either β_i^H with probability q, or β_i^L with probability $1-q$, where $\beta_i^H > \frac{b-a}{b-c} > \beta_i^L \geq 0$, and such realization is the private information of player i only. If an informed player i has $\beta_i = \beta_i^H$ ($\beta_i = \beta_i^L$) we refer to him as the "concerned" ("unconcerned," respectively) player. Note that we allow for $\beta_i^L = 0$. By contrast, player j's guilt parameter, β_j, is common knowledge for both players and $\beta_j > \frac{b-a}{b-c}$.[15] In order to focus on the possibility that player i signals his guilt concern, β_i, to the uninformed player j, let us assume that both individuals' envy concerns, α_i and α_j, are also common knowledge. Hence, player i holds private information about his guilt parameter β_i alone, since the precise value of α_i, either high $\alpha_i = \alpha_i^H$ or low $\alpha_i = \alpha_i^L$, where $\alpha_i^H > \alpha_i^L \geq \beta_i^H > \beta_i^L$, is common knowledge.

Duffy and Munoz-Garcia (2011) describe equilibrium behavior in the twice-repeated game with one-sided asymmetric information among players. Specifically, separating strategy profiles cannot be supported as Perfect Bayesian Equilibrium (PBE) of the signaling game for any prior q, whereas

[13] Mutual cooperation can also be supported as the SPNE of the infinitely repeated game by the use of other type of strategies, such as those in which defection is punished only during a limited number of time periods, or other reciprocal strategies like "tit-for-tat." In this section, we focus for simplicity in one type of strategy in order to analyze how social preferences can work as substitute for temporal preferences.

[14] Other potential explanation of this observed behavior could include incomplete information among the players or learning.

[15] Otherwise, player j would find defection to be a dominant strategy in the second period simultaneous-move game, and the first-period player i's actions would not affect his opponent's future play.

a pooling strategy profile can be sustained as a PBE in which both types of informed player i cooperate in the first period if priors are sufficiently high $q \geq \overline{q}(\alpha_j, \beta_j)$.[16] The uninformed player j cooperates both in the first period (given the relatively high prior) and in the second period, conditional on observing that player i cooperated in the first stage. Hence, by cooperating in the first period, the highly concerned player i guarantees outcome (C,C) in the second period of the game. In contrast, the unconcerned player i "disguises" himself as a player with high concerns for fairness who will cooperate in the following period. This misleading strategy induces player j to cooperate in the second period, where the unconcerned player i takes the opportunity to defect, yielding outcome (D,C). This "backstabbing" result might account for observed behavior in experimental settings, as in the literature suggested in the introduction, where subjects initially cooperate but choose to defect in the final period of the repeated game.

4.2 Repeated game under incomplete information

We next analyze the infinitely repeated version of the incomplete information game described in the previous section *with* discounting of future payoffs, i.e., player i's discount factor is $\delta_i \in (0,1)$. Unlike in the previous section, we now allow for two-sided asymmetric information about guilt parameters, i.e., every player i's guilt is either $\beta_i = \beta_i^H > \frac{b-a}{b-c}$ with probability $q_i \in (0,1)$ or $\beta_i = \beta_i^L < \frac{b-a}{b-c}$ with probability $1 - q_i$. We further assume that it is common knowledge that $\alpha_i = \alpha_j = \overline{\alpha}$, and that player i's and j's discount factors are δ_i and δ_j, entailing that the only element of uncertainty that a given player faces is his opponent's guilt parameter. Let us consider the case in which first-period actions transmit valuable information about a player's concern for fairness by examining the strategy profile where players cooperate during the first period of the game if and only if their discount factor is sufficiently high.[17] Following the first period of the game, players observe payoffs which allow them to perfectly infer the true type of their opponent. Incomplete information thus plays a role during the first period of the game alone, since in all subsequent periods, players' concerns for fairness are perfectly inferred from their first-period actions.[18] Furthermore, this information allows every player to predict whether his opponent will

[16] Since players interact during only two periods, we consider no discounting. In addition, note that we use "pooling" equilibrium to refer to strategy profiles in which both types of player i cooperate during the first-period game. In particular, Duffy and Munoz-Garcia (2011) show that cutoff $\overline{q}(\alpha_j, \beta_j)$ is $\overline{q}(\alpha_j, \beta_j) = \frac{d-c+\alpha_j(b-c)}{a+d-c-b+(\alpha_j+\beta_j)(b-c)}$. For robustness, the authors also show that this pooling equilibrium survives the Cho and Kreps' (1987) Intuitive Criterion.

[17] If, in contrast, both players select cooperate or both select defect during the first period of the game, then first-period actions do not communicate information about player i's type. In such cases, every player's period payoffs would not depend on the other player's type, and the introduction of incomplete information would not substantially modify our complete information analysis.

[18] We assume that every player compares his stage game payoff at every time period with that of his opponent, in order to evaluate the disutility from envy or guilt. Such information is typically available to subjects in experimental studies of indefinitely repeate games. Oechssler (2011) has recently proposed an alternative approach to incorporating social preferences into finitely repeated games where players compare their own discounted aggregate payoff at the end of the game with that of their opponent under complete informaiton. By contrast, under incomplete information our period-by period approach allows for dynamic strategies to arise due to the inequality that individuals experience over the course of the game. The alternative approach where player compare payoff at the end of the repeated game only allows for the emergence of dynamic strategies across supergames. We therefore prefer our approach of comparing invidual payoffs period-by-period.

cooperate in the continuation game. In particular, if player i's discount factor δ_i is sufficiently high, i.e., if $\delta_i \geq \delta^F(\beta_i^L) \geq \delta^F(\beta_i^H)$, then player i cooperates in the continuation game both when $\beta_i = \beta_i^L$ and when $\beta_i = \beta_i^H$. Similarly, if player i's discount factor is sufficiently low, he defects regardless of his type, i.e., when $\delta^F(\beta_i^L) \geq \delta^F(\beta_i^H) > \delta_i$. If, instead, player i's discount factor is intermediate, he cooperates when he has high concerns for fairness, $\delta_i \geq \delta^F(\beta_i^H)$, but defects otherwise, $\delta^F(\beta_i^L) > \delta_i$, i.e., $\delta^F(\beta_i^L) > \delta_i \geq \delta^F(\beta_i^H)$. In the first two cases there is no information transmission from first-period actions since all player types either cooperate or defect in the continuation game. By contrast, in the third case, first-period actions can communicate information about the players' type. Therefore, we focus on the latter case where $\delta^F(\beta_i^L) > \delta_i \geq \delta^F(\beta_i^H)$.

Proposition 2. *In the infinitely repeated Prisoner's Dilemma game with two-sided uncertainty, the following strategy profile can be supported as a PBE of the game:*

1. *In the first period play of the stage game, every player i cooperates when his guilt parameter is high, $\beta_i = \beta_i^H$, but defects when his guilt parameter is low, $\beta_i = \beta_i^L$, if and only if his discount factor, δ_i, satisfies $\delta_i^{UF}(\overline{\alpha}, \beta_i^L) > \delta_i \geq \delta_i^{UF}(\overline{\alpha}, \beta_i^H)$, where*

$$\delta_i^{UF}(\overline{\alpha}, \beta_i) \equiv 1 + \frac{(d-a)q_j}{d - c + q_j(b + c - 2d) - (b-c)[q_j\beta_i + (q_j - 1)\overline{\alpha}]} \quad \text{for any } \beta_i = \{\beta_i^L, \beta_i^H\}$$

2. *In subsequent period plays of the stage game, every player i cooperates if and only if all players cooperated in all prior periods for any discount factor $\delta_i \geq \delta_i^F(\beta_i^H)$ when his guilt parameter is high, $\beta_i = \beta_i^H$, and for any discount factor $\delta_i \geq \delta_i^F(\beta_i^L)$ when his guilt parameter is low, $\beta_i = \beta_i^L$.*

In addition, $\delta_i^{UF}(\overline{\alpha}, \beta_i)$ is decreasing in guilt aversion, β_i, but is increasing in envy aversion, $\overline{\alpha}$. Furthermore, $\delta_i^{UF}(\overline{\alpha}, \beta_i) \geq \delta_i^F(\beta_i)$ for any $\beta_i = \{\beta_i^L, \beta_i^H\}$; $\delta_i^{UF}(\overline{\alpha}, \beta_i) = \delta_i^F(\beta_i)$ as $q_j \to 1$, and $\delta_i^{UF}(\overline{\alpha}, \beta_i) = 1$ as $q_j \to 0$.

As suggested above, this strategy profile prescribes that every player i starts cooperating if his discount factor is sufficiently high, $\delta_i \geq \delta_i^{UF}(\overline{\alpha}, \beta_i)$, and continues cooperating as long as his opponent has cooperated in the past. Otherwise, players revert to defection thereafter. Two points are noteworthy.

First, the minimum discount factor supporting cooperation $\delta_i^{UF}(\overline{\alpha}, \beta_i)$ is decreasing in β_i, confirming the "substitutability" between time preferences and guilt found earlier for the complete information version of the infinitely repeated game. However, the minimum discount factor, $\delta_i^{UF}(\overline{\alpha}, \beta_i)$, is now *increasing* in $\overline{\alpha}$. Intuitively, larger concerns about envy raise the minimum discount factor needed to sustain cooperation during the first stage of the game. Specifically, under incomplete information about fairness concerns, players face the possibility that his opponent does not cooperate during that first stage, reducing the equilibrium payoff of the former from a to $c - \alpha_i(b - c)$.

Second, the minimum discount factor inducing an uninformed player to cooperate in the first period, $\delta_i^{UF}(\overline{\alpha}, \beta_i)$, is higher (more demanding) than under complete information, $\delta_i^F(\beta_i)$. As the probability of facing a cooperative player j tends to zero, $q_j \to 0$, player i's minimum discount factor supporting cooperation approaches one, indicating that cooperation is very unlikely to occur. This minimum discount factor is decreasing in q_j, and approaches the minimum discount factor for the complete information environment, $\delta_i^F(\beta_i)$, when the probability of facing a cooperative opponent approaches one, i.e., $q_j \to 1$. Finally, note that the results in Proposition 2 about two-sided asymmetric information embody one-sided asymmetric information as a special case, where $q_j = 1$ while $q_i \in (0, 1)$. In this context, player j is uninformed about player i's type but player i observes $\beta_j = \beta_j^H$, entailing that player j cooperates in the first period of the infinitely repeated game if his discount factor δ_j satisfies $\delta_j^{UF}(\overline{\alpha}, \beta_j^L) > \delta_j \geq \delta_j^{UF}(\overline{\alpha}, \beta_j^H)$ since he is still uninformed about his opponent's type, while player i cooperates if δ_i satisfies $\delta_i^F\left(\beta_i^L\right) > \delta_i \geq \delta_i^F(\beta_i^H)$ given that he is informed.

5 Extension to more general games and preferences

In this section we analyze equilibrium strategies in a more general class of infinitely repeated games. In particular, we consider simultaneous-move games with complete information and a finite number of players and actions. For simplicity, we restrict attention to games where players can choose cooperative action choices that improve their per-period payoffs relative to those in the Nash equilibrium of the stage game which we denote by $\widetilde{x_i}$.[19] That is, we consider games where there exists an action profile $a = (a_i, a_{-i})$ with payoff $U_i(a_i, a_{-i}) = x_i$, where $x_i > \widetilde{x_i}$, for every player i.[20] When players do not assign a value to fairness, i.e., $\beta_i = \alpha_i = 0$, mutual cooperation can be sustained as a SPNE of the infinitely repeated game for any discount factor δ_i such that $\delta_i \geq \delta_i^{NF}$ for all i; see, e.g., Friedman (1971). Similarly, as we have shown, when players with social preferences assign a value to fairness, mutual cooperation can be supported for any discount factor δ_i such that $\delta_i \geq \delta_i^F(\beta_i)$. This section examines the conditions for which $\delta_i^F(\beta_i) \leq \delta_i^{NF}$, i.e., that cooperation can be supported under a larger set of parameter values when players are concerned about fairness than when they are not.

We begin by defining a "weak symmetry" condition that we will make use of in Proposition 3 below. Specifically, a game satisfies "weak symmetry" if and only if all players' payoffs coincide when they play the Nash equilibrium of the stage game, i.e., $\widetilde{x_i} = \widetilde{x_j}$, as well as when they play the cooperative outcome in the repeated game, $x_i = x_j$. For instance, in the Prisoner's Dilemma

[19] Similar to previous sections, we focus on the case in which players are not extremely concerned about social preferences, so that the utility from the pure strategy Nash equilibrium of the stage game is higher than that from the mixed strategy equilibrium of the stage game. As a consequence, minmax payoffs when players are concerned about fairness coincide with those when players are not concerned. For simplicity, we consider the existence of a unique Nash equilibrium in the stage game. Our results can be extended to stage games with multiple Nash equilibria, and use $\widetilde{x_i}$ to denote the payoff that individuals obtain in the Nash equilibrium providing the highest payoff.

[20] For simplicity, we assume that payoff $U_i(a_i, a_{-i}) = x_i$ can be achieved using pure strategies. Otherwise, one can suppose that any randomization producing payoff x_i is publicly observed by all players, thus allowing deviations to be detected by every player.

game, this weak symmetry assumption implies that players 1 and 2 earn the same payoff when they both choose to defect and when they both choose to cooperate. Hence, if both individuals' payoffs coincide under these two strategy profiles, then utility levels when players are concerned about fairness will not be diminished (since there is no inequality in the payoff distribution). Note that weak symmetry is not as restrictive as stronger forms of symmetry, whereby player i's payoff coincides with that of player j at *every* strategy profile.

Proposition 3. *If the stage game's payoff structure satisfies the weak symmetry condition, then* $\delta_i^F(\beta_i) \leq \delta_i^{NF}$ *holds for all parameter values,* α_i *and* β_i*. Otherwise,* $\delta_i^F(\beta_i) \leq \delta_i^{NF}$ *if and only if*

$$\frac{\max_{a_i} U_i(a) - x_i}{\max_{a_i} U_i(a) - \widetilde{x_i}} \geq \frac{\max_{a_i} U_i^F(a) - x_i^F}{\max_{a_i} U_i^F(a) - \widetilde{x_i}^F}.$$

This result confirms our previous intuition from the Prisoner's Dilemma game: cooperation can be supported for weakly broader conditions when players are concerned about fairness than when they are not. In particular, when the game is weakly symmetric in players' payoffs, and players are concerned about fairness, their utility from deviating is reduced by the guilt they experience from being the player with the higher payoff. Importantly, this result can be applied to many simultaneous-move games besides the Prisoner's Dilemma game, including voluntary contribution (public good) games, and coordination games. By contrast, when the game does not satisfy the weak symmetry condition, cooperative outcomes can be supported with a weakly lower discount factor only if the above condition is satisfied. Intuitively, this condition holds if a player's incentives to deviate from the cooperative outcome are relatively stronger when he is unconcerned about fairness than otherwise. The following corollary shows that the result of Proposition 3 can be extended to players with social preferences different from those in Fehr and Schmidt (1999).

Corollary 1. *The result from Proposition 3,* $\delta_i^F(\beta_i) \leq \delta_i^{NF}$*, holds both under linear and non-linear social preferences.*

Proposition 3 therefore holds not only for the linear, Fehr and Schmidt (1999) specification of social preferences, where every unit of payoff inequality induces the same disutility (either in the form of envy or guilt), but also for more general (possibly *non-linear*) social preferences, such as those suggested by Neilson (2006):

$$U_i(x) = x_i - \alpha_i \sum_{i \neq j} u(x_i - x_j),$$

where u is any continuous function of the level of payoff inequality, $x_i - x_j$, and α_i is player i's sensitivity to such payoff inequality. Specifically, we can assume that u is increasing in $x_i - x_j$ whenever $x_i > x_j$, i.e., individuals experience a disutility (guilt) from receiving a higher payoff relative to other players in the population. Further, the disutility from guilt can be either increasing

in payoff inequality (if u is a convex function), or decreasing in payoff inequality (if u is a concave function). Our results are thus applicable to players with relatively general social preferences.

6 Feasible and individually rational payoffs

Let us finally examine how our previous results translate into the set of feasible payoffs for the infinitely repeated game. For simplicity, we focus on the two-player Prisoner's Dilemma game. Figure 3(a) below represents the set of feasible payoffs for the case where individuals do not assign any value to fairness considerations, $\beta_i = 0$, i.e., payoff pairs within the quadrilateral with vertices (a,a), (b,c), (d,d) and (c,b).[21] Let us denote this set of feasible payoffs by FP_{NF}, where, as before, the subscript NF denotes that players are not concerned about fairness. Formally, the set of feasible payoffs is defined as the convex hull of all payoffs $x \in \mathbb{R}^2_+$ feasible under the set of available actions $a \in A$, i.e., FP =convex hull $\{x \mid$there exists $a \in A$ such that $U(a) = x\}$.

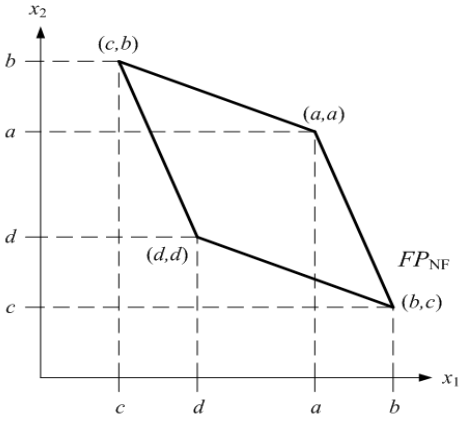

Figure 3(a). Set of feasible payoffs.

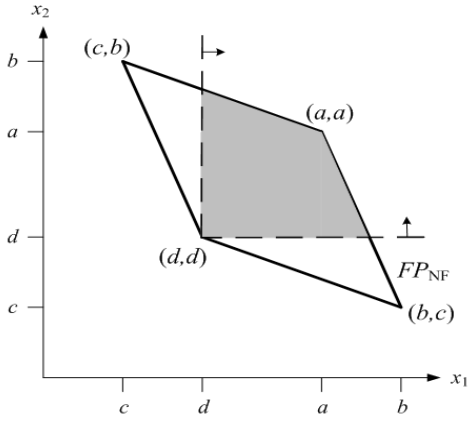

Figure 3(b). Individually rational payoffs.

Figure 3(b), additionally, shades the set of feasible payoffs that are *individually rational*, i.e., all those payoffs such that $x_i > \widehat{x_i}$ for both players, where $\widehat{x_i}$ is the reservation utility (or minmax value) $\widehat{x_i} = \min_{a_j} \left[\max_{a_i} U_i(a_i, a_j) \right]$. Hence, the shaded area in Figure 3(b) depicts the set of feasible, individually rational payoffs, where the minmax payoff pair is (d, d).

Let us now analyze how the feasible set is affected as players' concerns about fairness increase. In particular, Figure 4 below illustrates sets of feasible payoffs for players with positive concerns about fairness, $\beta_i > 0$, and compares those with the set of feasible payoffs for a player who assigns no value to fairness, FP_{NF}.

[21] For simplicity, Figures 3ab consider the case where $d < \frac{b+c}{2} < a$. We would obtain similar figures under alternative parametric restrictions.

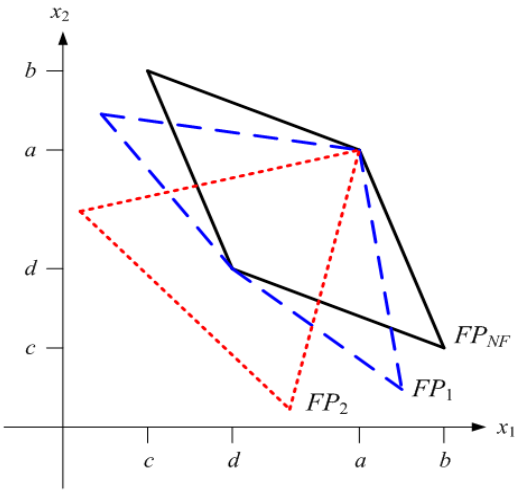

Figure 4. Set of feasible payoffs for $\beta_i > 0$.

Figure 4 reveals that as fairness concerns increase, the set of feasible payoffs shrinks like a fan closing its arms along the main diagonal, with its end at the pair of payoffs resulting from mutual cooperation, (a, a). Specifically, the blue set (long-dash) of feasible payoffs, FP_1, illustrates players with low concerns about fairness, i.e., $\beta_1, \beta_2 \in \left(0, \frac{b-a}{b-c}\right]$. Further increases in fairness concerns are represented by the red set (short-dash) of feasible payoffs, FP_2, where $\beta_1, \beta_2 \in \left(\frac{b-a}{b-c}, \frac{b-d}{b-c}\right]$. Note that at FP_1 defection is still a best response to cooperation. At FP_2, however, cooperation becomes a best response to cooperation.

Next, for the FP_2 set of feasible payoffs illustrated in Figure 4, Figure 5 below shades the portion of that set representing feasible and individually rational –FIR– payoffs. (Other FIR sets given FP sets are constructed similarly). Notice first that the FIR payoffs when players are concerned about fairness are not simply the payoffs to the northeast of the payoff pair (d, d), as in the case where individuals are not concerned about fairness (possess standard, self-interested preferences). Instead, when players are concerned about fairness, they now experience a disutility from all payoffs that lie away from the main diagonal (unequal payoff vectors), which results in the set of FIR payoffs becoming more compressed around egalitarian payoffs. We identify the set of FIR payoffs in the case where players are concerned with fairness, in particular the constraining payoff vectors $\overline{x_i}$ as illustrated in Figure 5, in the following corollary.

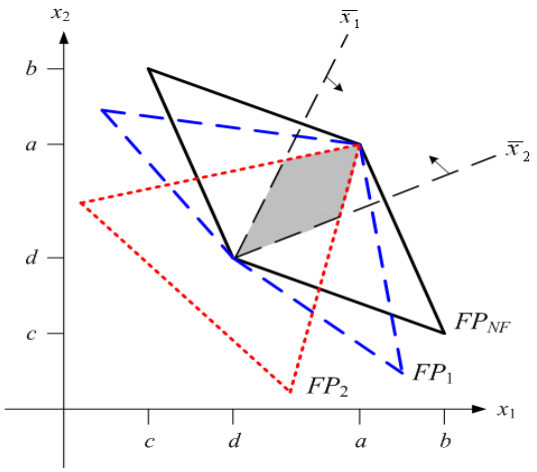

Figure 5. Effects of higher α_i on the FIR set.

Corollary 2. *In the infinitely repeated Prisoner's Dilemma game where individuals have social preferences, every player i's individually rational payoffs (within the set of feasible payoffs) must satisfy $x_i > \overline{x_i}$, where $\overline{x_i} \equiv \frac{d}{1+\alpha_i} + \frac{\alpha_i}{1+\alpha_i} x_j$ for every player i. Additionally, $\overline{x_i}$ is increasing in the envy parameter, α_i.*

Hence, as individuals become more envious (as α_i increases), the lower bound of the set of FIR payoffs, $\overline{x_i}$, shifts (downwards for player 1, $\overline{x_1}$, and upwards for player 2, $\overline{x_2}$) shrinking this set from above and below, respectively —resulting in the shaded area, as illustrated in Figure 5. Furthermore, increases in players' guilt aversion β_i must satisfy the preference assumption that $\alpha_i \geq \beta_i$. Thus, higher β_i will serve to shrink the size of the FIR set, as illustrated in Figure 5. If, in contrast $\alpha_i, \beta_i \to 0$, Corollary 2 reveals that the FIR set coincides with that illustrated in Figure 3(b) for selfish players, $\overline{x_i} = \overline{x_j} = d$.

Note the different roles played by envy and guilt concerns in the repeated game. On the one hand, Proposition 1 indicates that the minimal discount factor necessary to support cooperation in the infinitely repeated game decreases with players' guilt concerns (β_i). On the other hand, increases in individual's envy concerns (α_i) affect how egalitarian the payoff distribution in the repeated game must be, provided that players choose to cooperate in the repeated game, i.e., as Corollary 2 describes, an increase in α_i shrinks the set of FIR payoffs. Therefore, guilt serves as a "tool" for supporting cooperation under larger parameter values, whereas envy allows players to reach more equitable payoffs, provided that cooperative behavior can be sustained.

Hence, we can conclude that the set of FIR payoffs in the infinitely repeated game weakly shrinks as players become more concerned with fairness. This finding may be contrasted with that of Abreu et al. (1990), who show, in the context of infinitely repeated games in which players are not concerned about fairness, that the set of FIR payoffs weakly *expands* with increases in players' discount factor (i.e., as players assign a higher value to future payoffs, the set of FIR payoffs that

can be supported as equilibria of the repeated game expands). Thus our result complements that of Abreu et al. (1990) by suggesting the existence of an *opposing* force affecting the size of the set of FIR payoffs: higher discount factors weakly expand this set, while higher concerns about fairness serve to shrink the same set. In other words, our results show that the introduction of considerations about fairness work as a tool to reduce the multiplicity of strategy profiles that can be sustained as perfect equilibria in the infinitely repeated game.[22]

6.1 Patience or fairness? Experimental evidence

The above results suggest that a certain pair of payoffs can be sustained with a continuum of discount factors and concerns about fairness (different combinations of δ_i and β_i). Importantly, this implies that observed cooperation between players in infinitely repeated games could be due to a mix of these two factors.[23] Hence, our results suggest the possibility of some confusion as to which concern it is that leads players to sustain cooperation over time in repeated game experiments: is it patience alone (high δ_i values), is it fairness alone (high β_i values), or is it a combination of the two?

Our previous results provide a partial answer to this question. In the area in which no overlap occurs (the unshaded area of FP_{NF} in Figure 5 which does not overlap with FP_2, for instance, and within boundaries $\overline{x_1}$ and $\overline{x_2}$), players' cooperation is sustained because of individuals' time preferences alone. However, in the overlapping regions (the shaded area of FP_{NF} coinciding with that of FP_2 in Figure 5), players' cooperation in repeated games could be supported by combinations of discount factors and/or concerns about fairness.

Let us relate this theoretical prediction to some experimental data from an indefinitely repeated Prisoner's Dilemma game experiment reported in Duffy and Ochs (2009). In Figure 6 we show FIR payoffs for the parameterization of the indefinitely repeated Prisoner's Dilemma game that Duffy and Ochs implemented in the laboratory.[24] We then add average payoff data (from Duffy and Ochs's fixed pairings, indefinitely repeated game treatment) so as to compare these realized payoffs with our equilibrium predictions. In this figure we use black dots to represent the average payoffs accruing to fixed pairs of subjects over all rounds played in an indefinitely repeated Prisoner's Dilemma game.

[22] Similarly to the standard literature on repeated games, however, our results still predict multiple strategy profiles being supported in the subgame perfect equilibrium of the game. Nonetheless, our results help eliminate equilibria where per-period payoffs are relatively asymmetric if players are highly concerned about social preferences.

[23] We suspect that a confound between patience and fairness concerns also exists in finitely repeated games, which are more frequently studied in the experimental literature.

[24] Duffy and Ochs (2009) parameterize the stage game using $a = 20$, $b = 30$, $c = 0$ and $d = 10$. These numbers correspond to payoffs in US$ cents per round played. They use a continuation probability, $\delta = .90$, to test whether repeated interaction and learning lead to further cooperation. For details see Duffy and Ochs (2009).

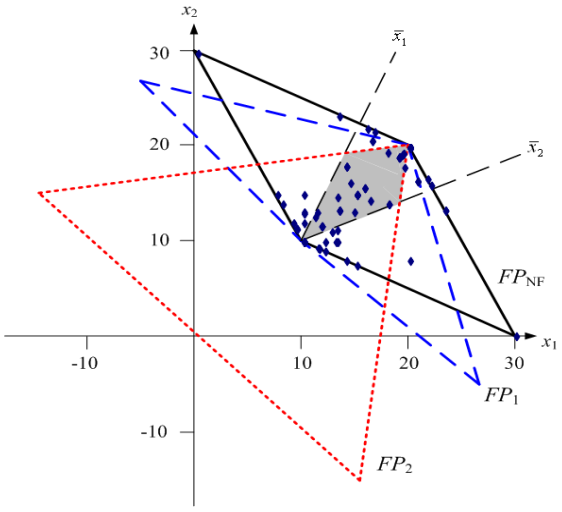

Figure 6

In particular, Figure 6 compares our predictions with respect to observed behavior for the case where players' social preferences are moderate, $\beta_1 = \beta_2 = \frac{1}{3}$, labeled FP_2. As Figure 6 reveals, individual subject behavior in the experiment can be explained: (1) by relying on individuals' time preferences alone (see the average payoff values lying outside the set of shaded FIR payoffs in Figure 6 but within the FP_{NF} set); or (2) relying both on individuals' time *and* social preferences (payoffs lying within the shaded set of FIR payoffs in Figure 6). In particular, we observe that most of the experimental observations on payoffs (64%) lie within this FIR set. Such payoff observations can be supported using either social or time preferences, or some combination of both. The rest of the experimental observations, lying outside the FIR set, cannot be sustained using social preferences alone (or a combination of social and time preferences) but can be supported based on time preferences alone. As concerns about fairness become more extreme, however, the set of FIR payoffs shrinks. As a consequence, more payoff pairs start to lie outside the FIR set of players sustaining social preferences, but still lie within the FIR set for players who are not concerned about fairness as represented in Figure 3(b). Hence, such payoff outcomes can be rationalized on the basis of time preferences alone.[25]

7 Conclusions

In this paper we have investigated how the introduction of social preferences and fairness concerns may affect players' equilibrium behavior in both one-shot and infinitely repeated versions of the Prisoner's Dilemma game. In particular, we analyze how fairness concerns modify players' incentives

[25] Note that, as in other experimental tests of infinitely repeated games, the game is repeated a finite number of times and hence players' observed average payoffs can differ from the predicted expected payoffs in the infinitely repeated game.

to cooperate in both versions of the game. In the one-shot stage game, we show that introducing players who are concerned about fairness might lead to cooperative outcomes in equilibrium, but only if both players assign a sufficiently high value to guilt. We then show that, in the infinitely repeated version of the game, the cooperative outcome can be sustained in equilibrium for lower discount factors when players are concerned about fairness than when they are not. This finding is consistent with some experimental evidence from indefinitely repeated games where the induced discount factor is too low to support cooperation under the assumption of rational, self-interested players.

We then investigate information transmission when players interact in a twice-repeated simultaneous prisoner's dilemma game with one-sided asymmetric information about one player's concerns for fairness. A pooling equilibrium can be supported in which an informed player unconcerned about fairness initially cooperates in order to mislead his uninformed opponent. Specifically, this misleading strategy induces the uninformed player to cooperate in the subsequent game, when the unconcerned player takes the opportunity to defect. This pooling equilibrium might explain incidences of end-game or "last-minute" defections in experimental settings. We also examine the infinitely repeated version of the two-sided incomplete information game, showing that cooperation becomes more difficult to sustain under incomplete than under complete information.

Finally, our findings suggest a potential confound in the interpretation of experimental results showing high levels of cooperative behavior in infinitely (indefinitely) repeated games, which has recently become the subject of much study in the experimental literature. First, our findings can be used to rationalize observed cooperative behavior in experimental settings with low induced discount factors where the Folk theorem for repeated games with discounting (under standard preferences) would predict an absence of cooperative behavior. Second, even in settings where this Folk theorem applies, we have shown how observed cooperation frequencies may be explained by time preferences alone or by a combination of time and social preferences. As a first step toward disentangling these two effects, we provide payoff vectors for which cooperation in the repeated game may only be rationalized using time discounting. Nonetheless, more experimental research on this topic is clearly needed, in order to clarify this potential confound.

Appendix

Proof of Lemma 1

Let us first analyze player i's best response function. If $\beta_i \leq \frac{b-a}{b-c}$ then defect is a strictly dominant strategy for player i. Indeed if player j cooperates, player i prefers to defect since $a \leq b - \beta_i(b-c)$ given that $\beta_i \leq \frac{b-a}{b-c}$, and if player j defects player i prefers to defect because $c - \alpha_i(b-c) < d$ given that $\frac{c-d}{b-c} < 0 \leq \alpha_i$ by definition. If instead $\beta_i > \frac{b-a}{b-c}$, then player i's best response to cooperation is to cooperate since $a > b - \beta_i(b-c)$ for all $\beta_i > \frac{b-a}{b-c}$, but his best response to defection is to defect given that $c - \alpha_i(b-c) < d$ for all $\frac{c-d}{b-c} < 0 \leq \alpha_i$. Thus in this case where $\beta_i > \frac{b-a}{b-c}$, player j may

cooperate with probability q so as to make player i indifferent between cooperating and defecting:

$$qa + (1-q)[c - \alpha_i(b-c)] = q[b - \beta_i(b-c)] + (1-q)d,$$

which yields $q = \frac{d-c+\alpha_i(b-c)}{a+d-c-b+(\alpha_i+\beta_i)(b-c)} \equiv \overline{q}(\alpha_i, \beta_i)$. In addition, note that the probability cutoff $\overline{q}(\alpha_i, \beta_i)$ is positive and smaller than one since $\beta_i > \frac{b-a}{b-c}$. Given players' best responses, if either $\beta_i \leq \frac{b-a}{b-c}$ or $\beta_j \leq \frac{b-a}{b-c}$, then the unique Nash equilibrium of the game is (D,D). Otherwise (if both $\beta_i > \frac{b-a}{b-c}$ and $\beta_j > \frac{b-a}{b-c}$), then both players' best response to C is C, and both players' best response to D is D. Hence, when $\beta_i, \beta_j > \frac{b-a}{b-c}$ (C,C) and (D,D) are Nash equilibria of the game in pure strategies. We now must check for the existence of mixed strategy equilibria. We know that if $\beta_i > \frac{b-a}{b-c}$, then player i is indifferent between selecting C and D if player j randomizes with probability $q = \overline{q}(\alpha_i, \beta_i)$, as described above. By symmetry, there is a mixed strategy Nash equilibrium where player i cooperates with probability $\overline{q}(\alpha_j, \beta_j)$ and player j cooperates with probability $\overline{q}(\alpha_i, \beta_i)$. ■

Proof of Proposition 1

Consider a representative period and suppose that both players have cooperated in all prior periods. If player i deviates to D (the best response to player j choosing C when $\beta_i < \frac{b-a}{b-c}$), then player j's trigger strategy specifies the play of D for all future periods following the deviation. Thus in this case, the deviation by player i in that period yields him the discounted payoff of $[b - \beta_i(b-c)] + \frac{\delta_i}{1-\delta_i}d$.

By contrast, if player i does not deviate in that period, so that both individuals continue cooperating, player i obtains a discounted payoff of $\frac{1}{1-\delta_i}a$. Comparing these two payoffs, we find that the deviation by player i is unprofitable if and only if $\delta_i \geq \frac{(b-a)-\beta_i(b-c)}{(b-d)-\beta_i(b-c)} \equiv \delta_i^F(\beta_i)$ for every player i. Note that, in the case that players do not assign any value to guilt, $\beta_i = 0$, we have $\delta_i^F(0) = \frac{b-a}{b-d} \equiv \delta_i^{NF}$. In addition, $\delta_i^F(\beta_i)$ is decreasing in β_i given that $\frac{\partial \delta_i^F(\beta_i)}{\partial \beta_i} = -\frac{(b-c)(a-d)}{[d-c\beta+b(\beta-1)]^2}$ is negative since $b > c$ and $a > d$. Furthermore, $\delta_i^F(\beta_i) > 0$ for all $\beta_i < \frac{b-a}{b-c}$. We can then express $\delta_i^F(\beta_i)$ as a function of δ_i^{NF}, as follows:

$$\frac{(b-a)-\beta_i(b-c)}{(b-d)-\beta_i(b-c)} = \frac{b-a}{b-d} - \frac{\beta_i(b-c)(d-a)}{(b-d)\left[\beta_i(b-c)-b+d\right]}.$$

Notice further that the difference $\delta_i^{NF} - \delta_i^F(\beta_i) = \frac{\beta_i(b-c)(d-a)}{(b-d)[\beta_i(b-c)-b+d]}$ is positive for all $\beta_i < \frac{b-d}{b-c}$, and that $\delta_i^F(\beta_i)$ becomes zero for $\beta_i \geq \frac{b-a}{b-c}$.

Finally, we need to show that a player would choose D forever, once either individual deviated in an earlier period. In order to prove this, note that if player j deviates, then he would be required to play D in all future periods. Further, player i's best response to individual j's playing D is to play D himself (we showed that in lemma 1). Therefore, the trigger strategies defined above comprise a subgame perfect equilibrium of this infinitely repeated Prisoner's Dilemma game. ■

Proof of Proposition 2

First note that, if players' concerns about fairness can be perfectly inferred from their choices during the first period of the game, then beginning with the second period, the game is one of complete information, resembling the one addressed in Proposition 1. In particular, after the first period of the game player i either: (1) cooperates regardless of his type if and only if his discount factor δ_i is sufficiently high, i.e., if $\delta_i \geq \delta^F(\beta_i^L) \geq \delta^F(\beta_i^H)$; (2) defects regardless of his type if and only if his discount factor is sufficiently low, i.e., if $\delta^F(\beta_i^L) \geq \delta^F(\beta_i^H) > \delta_i$; or (3) cooperates if his concern for fairness is high, $\delta_i \geq \delta^F(\beta_i^H)$, but defects if his concern for fairness is low, $\delta_i < \delta^F(\beta_i^L)$, which occurs when his discount factor is intermediate, i.e., when $\delta^F(\beta_i^L) > \delta_i \geq \delta^F(\beta_i^H)$. In the first two cases there is no information transmission from player i's first-period actions to his opponent (player j), since all types of player i either cooperate or defect in the continuation game. By contrast, in the third case, first period actions may communicate information about the player i's type. We focus on this case next. (Recall that the envy parameter $\overline{\alpha}$ and players' discount factors are common knowledge among players). Every player i cooperates during the first period of the game when his preferences for fairness are high $\beta_i = \beta_i^H > \frac{b-a}{b-c}$, if

$$q_j\left[a+\frac{\delta_i}{1-\delta_i}a\right]+(1-q_j)\left[c-\overline{\alpha}(b-c)+\frac{\delta_i}{1-\delta_i}d\right] \geq q_j\left[b-\beta_i^H(b-c)+\frac{\delta_i}{1-\delta_i}d\right]+(1-q_j)\left[d+\frac{\delta_i}{1-\delta_i}d\right]$$

where q_j denotes the probability that $\beta_j = \beta_j^H > \frac{b-a}{b-c}$, while $1-q_j$ is the probability of $\beta_j = \beta_j^L < \frac{b-a}{b-c}$. Solving for δ_i, we obtain that cooperation can be supported if and only if

$$\delta_i \geq 1+\frac{(d-a)q_j}{d-c+q_j(b+c-2d)-(b-c)[q_j\beta_i^H+(q_j-1)\overline{\alpha}]} \equiv \delta_i^{UF}(\overline{\alpha},\beta_i^H)$$

If player i's preferences for fairness are low, $\beta_i = \beta_i^L < \frac{b-a}{b-c}$, defection during the first stage of the game can be supported if

$$q_j\left[a+\frac{\delta_i}{1-\delta_i}a\right]+(1-q_j)\left[c-\alpha_i(b-c)+\frac{\delta_i}{1-\delta_i}d\right] < q_j\left[b-\beta_i^L(b-c)+\frac{\delta_i}{1-\delta_i}d\right]+(1-q_j)\left[d+\frac{\delta_i}{1-\delta_i}d\right]$$

which simplifies into $\delta_i < \delta_i^{UF}(\overline{\alpha},\beta_i^L)$, where $\delta_i^{UF}(\overline{\alpha},\beta_i^L) \geq \delta_i^{UF}(\overline{\alpha},\beta_i^H)$. Therefore, player i's discount factor δ_i must satisfy $\delta_i^{UF}(\overline{\alpha},\beta_i^L) > \delta_i \geq \delta_i^{UF}(\overline{\alpha},\beta_i^H)$ for this equilibrium to be sustained. Finally, note that

$$\frac{\partial\delta_i^{UF}(\overline{\alpha},\beta_i)}{\partial\beta_i} = -\frac{(b-c)(a-d)q_j}{[c-d-q_j(b+c-2d)+(b-c)(q_j\beta_i+(q_j-1)\overline{\alpha})]^2} < 0$$

and

$$\frac{\partial\delta_i^{UF}(\overline{\alpha},\beta_i)}{\partial\overline{\alpha}} = -\frac{(b-c)(a-d)(q_j-1)q_j}{[c-d-q_j(b+c-2d)+(b-c)(q_j\beta_i+(q_j-1)\overline{\alpha})]^2} > 0$$

Proof of Proposition 3

Let us first consider the case in which players do not exhibit social preferences. Assume that there is an action profile $a = (a_i, a_{-i})$ with payoff $U(a) = x$, where $x \in X$ and $x_i > \widetilde{x}_i$ for every player i, and consider the following strategy profile: in period zero each player i plays a_i. Each player i continues to play a_i so long as a was played in all previous periods. If at least one player did not play according to a, then every player i reverts to the minmax action for the rest of the game, with associated payoff $\widetilde{x}_i$. This strategy profile is a Nash equilibrium of the infinitely repeated game for discount factors, δ_i, such that

$$\frac{1}{1-\delta_i} x_i \geq \max_{a_i} U_i(a) + \frac{\delta_i}{1-\delta_i} \widetilde{x}_i \iff \delta_i \geq \frac{\max_{a_i} U_i(a) - x_i}{\max_{a_i} U_i(a) - \widetilde{x}_i} = \delta_i^{NF}$$

This strategy profile is subgame perfect, given that, in every subgame off-the-equilibrium path, the strategies are to play $\widetilde{x}_i$ forever, the Nash equilibrium of the stage game. Finally, note that when player i is concerned about fairness, his maximal benefit to a deviation from cooperation, $\max_{a_i} U_i^F(a)$, is weakly lower than that when he is not concerned about fairness, $\max_{a_i} U_i(a)$, because of the guilt he experiences from being the player with the highest payoff, i.e., $\max_{a_i} U_i^F(a) \leq \max_{a_i} U_i(a)$, Hence, $\delta_i^{NF} \geq \delta_i^F(\beta_i)$ is satisfied if and only if

$$\frac{\max_{a_i} U_i(a) - x_i}{\max_{a_i} U_i(a) - \widetilde{x}_i} \geq \frac{\max_{a_i} U_i^F(a) - x_i^F}{\max_{a_i} U_i^F(a) - \widetilde{x}_i^F},$$

where we do not impose any assumption on the symmetry of payoffs, i.e., allowing for $x_i \neq x_i^F$ and $\widetilde{x}_i \neq \widetilde{x}_i^F$ for any $\alpha_i, \beta_i > 0$. Otherwise, when the payoff structure satisfies weak symmetry, so that both $x_i = x_i^F$ and $\widetilde{x}_i = \widetilde{x}_i^F$ hold, this implies that the above inequality becomes

$$\frac{\max_{a_i} U_i(a) - x_i}{\max_{a_i} U_i(a) - \widetilde{x}_i} \geq \frac{\max_{a_i} U_i^F(a) - x_i}{\max_{a_i} U_i^F(a) - \widetilde{x}_i}$$

which can be simplified to $\max_{a_i} U_i(a)(x_i - \widetilde{x}_i) \geq \max_{a_i} U_i^F(a)(x_i - \widetilde{x}_i)$, which is satisfied for any parameter values, since $\max_{a_i} U_i(a) \geq \max_{a_i} U_i^F(a)$ and $x_i > \widetilde{x}_i$. Therefore, $\delta_i^{NF} \geq \delta_i^F(\beta_i)$. ■

Proof of Corollary 1

Note that social preferences are introduced in the proof of Proposition 3 by considering that a player's maximal benefit to a deviation from cooperation when he is concerned about fairness, $\max_{a_i} U_i^F(a)$, is weakly lower than that when he is not concerned about fairness, $\max_{a_i} U_i(a)$, i.e., $\max_{a_i} U_i^F(a) \leq \max_{a_i} U_i(a)$. No conditions are assumed about the players' payoffs x_i and x_i^F, or about $\widetilde{x}_i$ and $\widetilde{x}_i^F$. These assumptions embody both linear and non-linear social preferences. ■

Proof of Corollary 2

First, note that the payoff from the pure strategy Nash equilibrium of the stage game, d, exceeds that from the mixed strategy Nash equilibrium of the stage game if and only if $\beta_i < \frac{b-d}{b-c}$, which holds in this section. Payoffs pairs (x_i, x_j) above the reservation utility for player i imply that $x_i - \beta_i(x_i - x_j) > d$ when payoffs satisfy $x_i \geq x_j$, imply that $x_i - \alpha_i(x_j - x_i) \geq d$ when instead payoffs satisfy $x_i < x_j$. More compactly we have

$$\begin{aligned} x_i &\geq \frac{d}{1-\beta_i} - \frac{\beta_i}{1-\beta_i} x_j \text{ for all } i \text{ and } j, \text{ if } x_i > x_j; \text{ and} \\ x_i &\geq \frac{d}{1+\alpha_i} + \frac{\alpha_i}{1+\alpha_i} x_j \text{ for all } i \text{ and } j, \text{ if } x_i < x_j, \end{aligned}$$

respectively. These two lower bounds cross at payoff $x_j = d$; the first is below the second for all $x_j > d$, and similarly for player j. Consider the lower bounds for player i. For all $x_i > d$, the first bound is below the second, and hence only the second inequality is binding for every player i. Therefore, the set of individually rational payoffs can be defined by $x_i \geq \frac{d}{1+\alpha_i} + \frac{\alpha_i}{1+\alpha_i} x_j$. Differentiating with respect to α_i we obtain $\frac{\partial \overline{x}_i}{\partial \alpha_i} = \frac{x_j - d}{(1+\alpha_i)^2}$, which is positive for all $x_j > d$ in the relevant region of the set of FIR payoffs. ■

References

[1] Abreu, Dilip, David G. Pearce and Ennio Stachetti (1990) "Toward a theory of discounted repeated games with imperfect monitoring," 58(5), pp. 1041-1063.

[2] Anderhub, Vital, Dirk Engelmann and Werner Güth (2002) "An experimental study of the repeated trust game with incomplete information," 48(2), pp. 197-216.

[3] Andreoni, James and J.H. Miller (1993) "Rational cooperation in the finitely repeated prisoner's dilemma: experimental evidence," The Economic Journal, 103, pp. 570-585.

[4] Aoyagi, Masaki and Guillaume Fréchette (2009) "Collusion as public monitoring becomes noisy: Experimental evidence," Journal of Economic Theory, 144, pp. 1135-1165.

[5] Baron, D. P. and J. A. Ferejohn (1989) "Bargaining in Legislatures," American Political Science Review, 94(1), pp. 73-88.

[6] Bolton, Gary E. and Axel Ockenfels (2000) "ERC: A theory of equity, reciprocity, and competition," American Economic Review, 90, pp. 166-93.

[7] Blonski, Matthias, Peter Ockenfels and Giancarlo Spagnolo (2011) "Equilibrium Selection in the Repeated Prisoner's Dilemma: Axiomatic Approach and Experimental Evidence," American Economic Journal: Microeconomics, 3 (August), pp. 164–192.

[8] Brandts, Jordi and Neus Figueras (2003) "An Exploration of Reputation Formation in Experimental Games," Journal of Economic Behavior and Organization, 50, pp. 89-115.

[9] Camera, Gabriele and Marco Casari (2009) "Cooperation among Strangers under the Shadow of the Future," American Economic Review, 99(3), pp. 979–1005.

[10] Camerer, Colin and Keith Weigelt (1988) "Experimental Tests of the Sequential Equilibrium Reputation Model," Econometrica, 56, pp. 1-36.

[11] Camerer, Colin F. (2003) *Behavioral game theory*, Princeton: Princeton University Press.

[12] Chade, Hector, Pavlo Prokopovych and Lones Smith (2008) "Repeated games with present-biased preferences," Journal of Economic Theory, 139, pp. 157-175.

[13] Cho, In-Koo and David Kreps (1987) "Signaling games and stable equilibrium," Quarterly Journal of Economics, vol. 102, 179-222.

[14] Cooper, Russell, Douglas V. DeJong, Robert Forsythe and Thomas W. Ross (1996) "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games," Games and Economic Behavior, 12, pp. 187-218.

[15] Dal Bó, Pedro (2005) "Cooperation under the Shadow of the Future: Experimental Evidence from Infinitely Repeated Games," American Economic Review, 95(5), pp. 1591-1604.

[16] Dal Bó, Pedro and Guillaume R. Fréchette (2011) "The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence," American Economic Review 101(1), pp. 411–29.

[17] Duffy, John and Jack Ochs (2009) "Cooperative behavior and the frequency of social interaction," Games and Economic Behavior, 66, pp. 785-812.

[18] Duffy, John and Felix Munoz-Garcia (2011) "Signaling concerns about fairness: Cooperation under uncertain social preferences," working paper.

[19] Fehr, Ernst and Urs Fischbacher (2002) "Why social preferences matter - The impact of non-selfish motives on competition, cooperation and incentives," Economic Journal, 112, pp C1-C33.

[20] Fehr, Ernst and Klaus Schmidt (1999) "A theory of fairness, competition and cooperation," Quarterly Journal of Economics, 114, pp. 817-68.

[21] Fischbacher, Urs and Simon Gächter (2010) "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments," American Economic Review, 100(1), pp. 541–56

[22] Friedman, James (1971) "A noncooperative equilibrium for supergames," Review of Economic Studies, 38, pp. 1-12.

[23] Fudenberg, Drew and Eric Maskin (1986) "The folk theorem in repeated games with discounting or with incomplete information," Econometrica, 54, pp. 533-556.

[24] Fudenberg, Drew, David G. Rand and Anna Dreber (2011), "Slow to Anger and Fast to Forgive: Cooperation in an Uncertain World," forthcoming in American Economic Review.

[25] Hauk, Esther (2003) "Multiple prisoner's dilemma games with(out) an outside option: an experimental study," Theory and Decision, 54, pp. 207-229.

[26] Healy, P.J. (2007) "Group Reputations, Stereotypes, and Cooperation in a Repeated Labor Market" American Economic Review, 97(5), pp. 1751–1773.

[27] Kreps, David, Paul Milgrom, John Roberts and Robert Wilson (1982) "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma," Journal of Economic Theory 27, pp. 245-52.

[28] McKelvey, Richard D. and Thomas R. Palfrey (1992) "An Experimental Study of the Centipede Game," Econometrica, 60, pp. 803-836.

[29] Montero, Maria (2007) "Inequity Aversion may Increase Inequity," Economic Journal, 117, pp. 192-204.

[30] Murnighan, J. Keith and Alvin E. Roth (1983) "Expecting continued play in Prisoner's Dilemma games: a test of several models," Journal of Conflict Resolution, 27(2), pp. 279-300.

[31] Neilson, William S. (2006) "Axiomatic Reference-Dependence in Behavior towards Others and towards Risk," Economic Theory, 28(3), pp. 681-92.

[32] Normann, Hans-Theo and Brian Wallace (2006) "The impact of the termination rule on cooperation in a Prisoner's Dilemma experiment," working paper.

[33] Oechssler, Jörg (2011) "Finitely Repeated Games with Social Preferences," Working paper 515, University of Heidelberg.

[34] Rabin, Matthew (1997) "Fairness in Repeated Games," University of California at Berkeley working paper, no 97-252.

[35] Selten, Reinhard and Rolf Stoecker (1986) "End behavior in sequences of finite Prisoner's Dilemma supergames. A learning theory approach," Journal of Economic Behavior and Organization, 7(1), pp. 47-70.

[36] Yamamoto, Yuichi (2010) "The use of public randomization in discounted repeated games," International Journal of Game Theory, 39, pp. 431-443.