Cooperation and Signaling with Uncertain Social Preferences*

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

May 2013

Abstract

This paper investigates behavior in finitely repeated simultaneous and sequential-move prisoner's dilemma games when there is one-sided incomplete information and signaling about players' concerns for fairness, specifically, their preferences regarding "inequity aversion." In this environment, we show that only a pooling equilibrium can be sustained, in which a player type who is unconcerned about fairness initially cooperates in order to disguise himself as a player type who is concerned about fairness. This disguising strategy induces the uninformed player to cooperate in all periods of the repeated game, including the final period, at which point the player type who is unconcerned about fairness takes the opportunity to defect, i.e., he "backstabs" the uninformed player. Despite such last-minute defection, our results show that the introduction of incomplete information can actually result in a Pareto improvement under certain conditions. We connect the predictions of this "backstabbing" equilibrium with the frequently observed decline in cooperative behavior in the final period of finitely-repeated experimental games.

KEYWORDS: Prisoner's Dilemma; Social Preferences; Inequity aversion; Incomplete Information; Signaling; Information Transmission.

JEL CLASSIFICATION: C72, C73, D82.

*We thank the editor and six referees for helpful comments and suggestions on an earlier draft of this paper.

1 Introduction

A large body of experimental evidence suggests that many individuals exhibit concerns for fairness in the income distribution, a characteristic that is also referred to as "social" preferences. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), among others, have provided models where the incorporation of such social preferences can help to explain experimental findings, in particular, greater-than-predicted "cooperative" behavior, that would be difficult to rationalize under standard, "selfish" preferences. This paper contributes to the literature on the role of social preferences in fostering cooperative behavior by studying settings where players have incomplete information about other players' social preferences, specifically, incomplete information about their degree of inequity aversion. Surprisingly, most theoretical analyses of social preferences has been developed in complete information settings, where players can perfectly observe one another's concerns for fairness, e.g., the extent of their inequity aversion. While such an assumption may be reasonable in contexts where players have interacted with one another for several previous periods, complete information about other players' social preferences seems less sensible if players are unfamiliar with the strategic environment, or have had no prior interactions with their opponents, a situation that characterizes the initial round(s) of play of many experimental games. We study incomplete information about other's social preferences by considering signaling games, where players interact in either the simultaneous- or sequential-move versions of the Prisoner's Dilemma (PD) game.[1]

We first investigate the role of information transmission in the twice-repeated version of the simultaneous-move PD game. We show that no separating strategy profile can be sustained in equilibrium, thereby limiting the information transmitted about players' social preferences. By contrast there always exists a pooling equilibrium where the uninformed player is unable to distinguish the type of his opponent from the initial period action choice. If priors are sufficiently high, the uninformed player is attracted towards cooperation in the second and final period, at which point an opponent with low concerns for fairness takes the opportunity to defect, i.e., he "stabs the uninformed player in the back." (We refer to this strategy profile as the "backstabbing" equilibrium.) Interestingly, this equilibrium provides an explanation for a relatively common observation in experimental settings, wherein subjects defect in the last period of their interactions, despite a previous history of cooperation as shown, e.g., in Selten and Stoecker (1986) and Andreoni and Miller (1993) for the PD game, McKelvey and Palfrey (1992) for the centipede game, Camerer and Weigelt (1988) and Brandts and Figueras (2003) for the borrower-lender game, and Anderhub, Engelmann and Güth (2002) for the finitely-repeated trust game.[2] Importantly, this informational explanation, does not rely on subjects' inability to understand the rules of the game, or a failure to backward induct, but rests instead on the existence of incomplete information about other players'

[1] Players in PD games face similar strategic incentives as those in public good games and, more generally, games where players' actions are strategic substitutes. Sequential-move versions are also used to characterize firm-worker wage-effort decisions and the notion of "gift-exchange."

[2] For instance, for the finitely-repeated PD game, Andreoni and Miller's (1993) experiment shows that cooperation peaks in the first round of interaction (86% of subjects cooperate), stays above 50% until round 6, and then falls to about zero in the last (tenth) round.

social preferences.

We further demonstrate that in the backstabbing equilibrium of our incomplete information set-up, the payoffs of both the defecting player and the player suffering the defection, are actually greater than in the equilibrium under complete information. Therefore, the introduction of incomplete information entails a Pareto improvement under certain conditions, thus implying that all individuals should prefer interacting in incomplete rather than complete information settings.

Healy (2007) identifies a similar pooling equilibrium in the context of a finitely-repeated gift-exchange game where the firm manager does not observe the worker's type.[3] To facilitate a comparison of our results with those in Healy (2007), we modify the above signaling game in order to make it strategically equivalent to the gift-exchange game.[4] We show that a pooling equilibrium also emerges in the sequential version of that game. We demonstrate, nonetheless, that this "backstabbing" equilibrium can be supported under different parameter conditions in the simultaneous and sequential-move versions of the PD game, depending upon the first-mover's concern for fairness, which provides a set of testable results in controlled experiments of the simultaneous and sequential PD game.

2 Related Literature

Bolle and Ockenfels (1990) have also analyzed the PD game when players do not observe each others' altruistic motives. They mainly focus on the behavior of the second-mover in a sequential-move version of the game. By contrast, our model analyzes equilibrium play under a signaling game set-up both in the simultaneous- and sequential-move versions of the PD game which, as suggested above, can help to rationalize several experimental observations. Our signaling model also relates to that of Fong (2009) who, in the context of a two period, sequential move gift-exchange game with incomplete information about players' degree of altruism, studies how players convey or conceal their type by modifying the size of their gifts. Unlike our paper, no pooling equilibrium can be sustained in Fong's gift-exchange game, thus hindering his ability to rationalize the "last-minute" defections frequently observed in controlled experiments.[5]

The behavioral assumptions in this paper are also related to those in Kreps et al. (1982). In particular, they show that cooperation can be sustained in the finitely-repeated PD game so long as both players believe there is a small probability that his opponent is "irrational," i.e., plays a

[3] In particular, the worker's type is assumed to be either reciprocator or selfish, since Healy's results do not derive from any preference specification.

[4] In particular, we examine a twice-repeated sequential-move PD game where the first mover is uninformed about the second-mover's social preferences. In this game the first-mover cooperates only when he believes that the second-mover will reciprocate afterwards (which occurs when the second-mover is highly concerned about fairness). These strategic incentives coincide with those in the gift-exchange game analyzed by Healy (2007) whereby the firm manager only offers high wages when he believes that the worker is a reciprocating type.

[5] In a study of the hold-up problem under incomplete contracts, von Siemens (2009) considers a signaling game whereby the seller of a good initially invests in the good's quality, and then the seller and a buyer interact in an ultimatum bargaining game where the buyer makes a take-it-or-leave-it offer to the seller. Similar to our results, his paper also shows that the seller's initial investment can serve to conceal his privately observed fairness concerns to the buyer in order to condition the buyer's offers in the ensuing bargaining.

conditionally cooperative tit-for-tat strategy. In particular, Kreps et al. (1982) suppose that *both* players are uncertain about each other's stage payoffs, i.e., every player ignores the benefit that his opponent obtains from mutual cooperation. Importantly, they show that their conclusions would not hold in a context of one-sided uncertainty in which it is common knowledge that defection is a dominant strategy for the uninformed player.[6] In this paper we show that Kreps et al.'s (1982) cooperative results can be extended to environments with one-sided uncertainty. Specifically, this occurs when it is common knowledge that the uninformed player's best-response is to mimic his opponent's actions, i.e., when the uninformed player is a "reciprocator." In that setting, our "backstabbing" equilibrium predicts cooperation during the first period of play by the informed player type who is unconcerned about fairness and who attempts to convince their uninformed opponent that they will cooperate in subsequent periods.

Kreps and Wilson (1982) also show the existence of a pooling equilibrium in a one-sided incomplete information game in which the privately informed incumbent firm conceals its cost type from uninformed potential entrants. While that paper and several other studies analyzing entry-deterring practices by incumbent firms demonstrate the existence of pooling equilibria where type information is concealed, these pooling equilibria typically violate Cho and Kreps' (1987) equilibrium refinement, known as the "Intuitive Criterion", thus implying that only the fully informative (separating) equilibria can be supported as a robust equilibrium. By contrast, we show that the pooling equilibrium of the games we study survives Cho and Kreps' Intuitive Criterion (as well as other equilibrium refinements), so that the "backstabbing" behavior by the informed player is a robust equilibrium phenomenon.

The next section presents the model. Section four compares equilibrium outcomes in the simultaneous- and sequential-move versions of the PD game under complete information. Section five investigates information transmission about privately observed social preferences in the incomplete information, signaling–PD game, both in its simultaneous and sequential move versions. Section six concludes.

3 Model

Consider the two player stage game shown below. To make this game a Prisoner's Dilemma game, both players' payoffs must satisfy the restriction $b > a > d > c$. In that case, defect (D) becomes a strictly dominant strategy and outcome (D,D) is the unique equilibrium of the one-shot stage

[6] In such a context, the uninformed player defects at every stage of the game and, therefore, the informed player cannot affect the uninformed player's actions. This eliminates the possibility of information transmission; see Kreps et al. (1982) page 251. If, in contrast, the model in Kreps et al. (1982) is modified to allow for one-sided uncertainty where it is common knowledge that reciprocation is a dominant strategy for the uninformed player, then both our model and theirs would yield a similar cooperative outcome. Such cooperation, nonetheless, originates from inequity aversion in our model (which has regularly been observed in experiments), while the irrational reciprocation in Kreps et al. (1982), i.e., which arises when their parameter a is lower than 1, would be more difficult to support experimentally.

game.[7]

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

We focus on the case where both players possess Fehr and Schmidt (1999)-type social preferences, a now standard specification:

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

Here x_i is player i's payoff, and x_j is his opponent's (player j's) payoff. Parameter α_i represents the disutility from allocations that are disadvantageously unequal for player i (i.e., due to envy about player j's higher payoff), while β_i captures the disutility from allocations that are advantageously unequal for player i (e.g., due to guilt over earning a higher payoff than player j). Additionally, Fehr and Schmidt (1999) assume that envy dominates guilt concerns, i.e., $\alpha_i \geq \beta_i$ and $1 > \beta_i \geq 0$.[8] We will contrast this case of "social preferences" (specifically, "inequity aversion") with the more standard, self-regarding or "selfish" preferences where $\alpha_i = \beta_i = 0$ for all i. Introducing social preferences, the stage game can be reformulated as follows:

		Player 2 C	*Player 2* 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 that if player i's concerns about guilt (fairness) are relatively low, $\beta_i < \frac{b-a}{b-c}$, defection becomes a strictly dominant strategy for player i. By contrast, if player i's concern for fairness is relatively high, $\beta_i \geq \frac{b-a}{b-c}$, his best response is to behave "reciprocally" by matching player j's action, i.e., cooperate when j cooperates and defect otherwise.[9]

[7] We also consider the usual second condition on the parameters of PD games, $2a > b + c$, to guarantee that, in the iterative version of the game, mutual cooperation provides a larger payoff than that arising from alternating cooperation and defection.

[8] Intuitively, $\alpha_i \geq \beta_i$ implies that players (weakly) suffer more from inequality directed at them than inequality directed at others. Empirically, estimates of α_i have been found to be 2 to 3 times higher than estimates of β_i. 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 more details, see Fehr and Schmidt (1999).

[9] 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). This result also relates with that of Rabin (1993) for psychological games, where he assumes that players are motivated by the kindness they infer from other players' actions. Rabin (1993) assumes, however, that individuals' kindness parameters are common knowledge among the players. In contrast, we extend our study by allowing for incomplete information.

4 Complete information

We begin by briefly analyzing equilibrium predictions for the simultaneous-move Prisoner's Dilemma (PD) game under complete information about social preferences, which we will later use for comparison with equilibria under asymmetrically informed players. As shown in Duffy and Muñoz-García (2012), if at least one player has relatively low concerns about guilt, the unique Nash equilibrium of the game, (D,D), coincides with that in games where players have *no* concerns about the fairness of the payoff distribution (standard preferences)– see Figure 1 below.

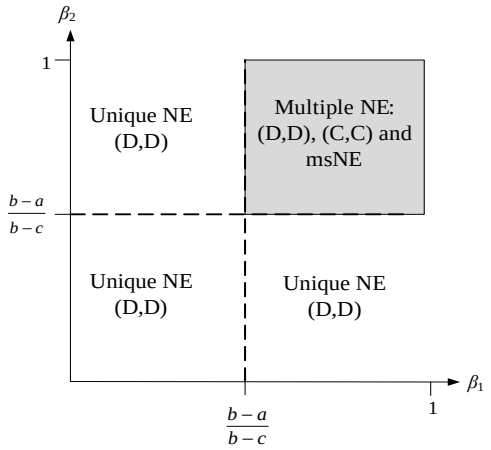

Figure 1. Equilibria in the simultaneous game under complete information.

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 equilibrium where they randomize. The introduction of high concerns for fairness by *both* players transforms the payoff structure of the game from a PD to a Pareto-rankable coordination game, where every player's best response is to select the same action as his opponent, but both players strictly prefer outcome (C,C) to outcome (D,D).

Let us now investigate the one-shot, sequential-move version of the PD game. Under complete information, the second mover (player j) adopts a "reciprocal" strategy (cooperating when the first mover cooperates and defecting otherwise) if his own concern for fairness is sufficiently high, $\beta_j \geq \frac{b-a}{b-c}$, but defects otherwise.[10,11] Given the second mover's response, the first mover (player

[10] This best response function for the second mover resembles that of Falk and Fischbacher (2006). In particular, assuming that individuals are perfectly informed about each others' reciprocal motivations, they show that the second mover might respond by "matching" the first mover's choice if the second mover is sufficiently reciprocally motivated. When the second mover is insufficiently motivate to reciprocate, he responds to any action of the first mover with defection.

[11] Clark and Sefton (1998) provide an experimental test of this best response function. Specifically, they modify the payoff structure in the sequential PD game so that the second mover can obtain a "temptation payoff" if he

i) chooses to cooperate when he anticipates that the second mover is "reciprocal," i.e., $\beta_j \geq \frac{b-a}{b-c}$. The following lemma summarizes this equilibrium result.

Lemma 1. *In the sequential-move PD game where players are informed about each others' social preference parameters, the unique subgame perfect equilibrium of the game prescribes that:*

1. *the first mover cooperates only if the second mover's concerns for fairness are sufficiently high,* $\beta_j \geq \frac{b-a}{b-c}$*, but defects otherwise; and*
2. *the second mover reciprocates if his concerns for fairness are sufficiently high,* $\beta_j \geq \frac{b-a}{b-c}$*, but defects otherwise.*

The equilibrium outcomes associated with different preference parameters are represented in Figure 2. The sequential time structure of the game therefore serves to support the cooperative outcome (C,C) under a larger set of parameter values than in its simultaneous version (compare the shaded areas in Figures 2 and 1, respectively). In particular, cooperation can be sustained in the sequential-move game so long as the second mover is sufficiently concerned about fairness, $\beta_j \geq \frac{b-a}{b-c}$, as Figure 2 illustrates, unlike in the simultaneous game, where such an outcome could only be sustained if *both* players' fairness concerns are sufficiently high, i.e., if $\beta_i, \beta_j \geq \frac{b-a}{b-c}$.

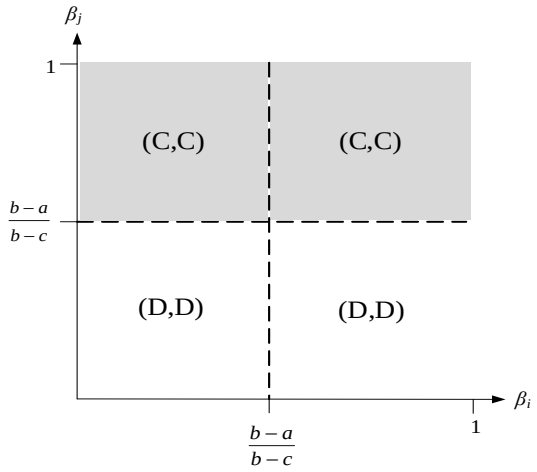

Figure 2. Equilibria in the sequential game under complete information.

is the only player defecting. Note that this payoff structure resembles ours, since payoffs associated to the (C,C) and (D,D) outcomes are unmodified, relative to the standard PD game (with selfish players), but those in which only the second mover defects vary. In particular, they find that the second mover is more likely to respond to cooperation with cooperation as the "temptation payoff" from defecting decreases. This experimental observation is in line with our result, since the second mover has greater incentives to respond to cooperation with cooperation if his concerns for fairness are relatively high (when the "temptation payoff" from defecting decreases), but responds by defecting against any choice by the first mover when he (the second mover) is unconcerned about fairness (when the "temptation payoff" increases).

5 Signaling private concerns about fairness

We now consider the case of asymmetric, incomplete information. Specifically, suppose that nature selects player i's concern for fairness, β_i, and this information is revealed to player i alone. For simplicity we assume that β_i is distributed according to a discrete probability distribution, specifically:

$$\beta_i = \begin{cases} \beta_i^H \text{ with probability } q, \text{ or} \\ \beta_i^L \text{ with probability } 1-q \end{cases}$$

where $\beta_i^H > \frac{b-a}{b-c} > \beta_i^L \geq 0$. We refer to a player i with $\beta_i = \beta_i^H$ as the "concerned" player and a player i with $\beta_i = \beta_i^L$ as the "unconcerned" player. Note that we allow for $\beta_i^L = 0$. By contrast, player j's guilt parameter, β_j, is assumed to be perfectly observable by both players and is commonly known to satisfy the condition: $\beta_j > \frac{b-a}{b-c}$.[12] That is, we here adopt a one-sided incomplete information structure. This set-up contrasts with standard incomplete information games where information is equally distributed among players. While symmetric incomplete information structures may be regarded as a rather special, "knife-edge" case; more general incomplete information games allow for asymmetries in the players' information about the unknown parameters of the game. The one-sided asymmetric information structure we assume is simply an extreme version of such an asymmetric incomplete information game. Furthermore, our information structure is the standard one used in the signaling game literature, allowing for analysis of strategic information transmission from an informed to an uninformed player.

In order to focus on the possibility that player i signals his guilt concern, β_i, to the uninformed player j, we 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, where $\alpha_i \geq \beta_i^H > \frac{b-a}{b-c}$, is common knowledge.[13] Finally, we assume that $\alpha_i^H > \alpha_i^L$, that is, players with strong concerns about inequity aversion exhibit both a larger guilt parameter and a larger envy parameter than do unconcerned individuals. Such consistency in the individuals' comparisons of unequal offers has been empirically documented by Bellemare, Kröger and van Soest (2008), who found, using a representative sample of the Dutch population, that individuals who exhibited strong guilt concerns in play of an ultimatum bargaining game also exhibited strong envy concerns (see p. 833).

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

[13] In order to guarantee that the observation of envy parameter α_i does not allow the uninformed player j to infer the guilt parameter β_i, consider that α_i is distributed according to a continuous distribution function $G(\alpha_i)$, which assigns a positive probability to all $\alpha_i \geq \beta_i^H$. In such setting, observing the precise realization of α_i does not provide player j with any additional information about β_i, other than that β_i must satisfy $\alpha_i \geq \beta_i$, which holds by assumption. Note that if, instead, α_i was distributed according to a discrete probability distribution by which α_i could only take two possible values, the mere observation of the realization of α_i would allow the uninformed player j to infer the value of β_i, thus nullifying the signaling role of player i's actions.

5.1 Twice-repeated simultaneous-move PD game

The timing of the twice-repeated signaling game is as follows. First, before any interaction takes place, player i privately observes his social preferences, but the uninformed player j does not (he only knows the prior probability distribution of β_i). Then players play a simultaneous-move PD game during the first period. After the game is played, payoffs are distributed among players, which allows every player to infer the action that his opponent selected in the first stage game. In particular, the uninformed player j can use that information to update his beliefs about player i's type being high, conditional on player i having cooperated in the previous period, $\mu\left(\beta_i^H|C\right)$, or conditional on player i having defected in the previous period, $\mu\left(\beta_i^H|D\right)$. Given these beliefs, a second and final stage of the simultaneous-move PD game is played and payoffs are accrued.[14] The following proposition states that in this information setting only a pooling equilibrium can be supported in which both types of informed player i cooperate in the first-period stage game.[15]

Proposition 1. *A separating strategy profile cannot be supported as a perfect Bayesian equilibrium (PBE), for any prior q. In contrast, a pooling PBE can be sustained if* $q > \frac{d-c+\alpha_j(b-c)}{a+d-c-b+(\alpha_j+\beta_j)(b-c)} \equiv q^{Sim}(\alpha_j, \beta_j)$ *in which the informed player i cooperates in both the first and the second period when he is concerned about fairness, but cooperates only in the first period when he is unconcerned about fairness. In this strategy profile, the uninformed player j cooperates in the first period, but in the second period he cooperates if and only if player i cooperated in the first period, if beliefs satisfy* $\mu\left(\beta_i^H|C\right) = q \geq q^{Sim}(\alpha_j, \beta_j) > \mu\left(\beta_i^H|D\right)$.

A separating strategy in which player i cooperates (defects) in the first period when his concern for fairness is high (low, respectively), thus revealing his type to the uninformed player j, cannot be sustained in equilibrium. For a separating strategy to be an equilibrium, several conditions would need to be simultaneously satisfied and these conditions violate our initial assumptions. In particular, in a separating equilibrium the highly-concerned player i would cooperate during the first period game in order to reveal his true type while his opponent defects, yielding outcome (C,D). For such a strategy to be part of an equilibrium, player i cannot experience a large disutility from envy, i.e., $\alpha_i^H < \frac{a+c-2d}{b-c}$. In addition, the player i with low concerns for fairness cannot have incentives to mimic the highly-concerned type by choosing to cooperate in the first period. This occurs, specifically, if the disutility from envy that player i bears during the first period (when

[14] Therefore, the uninformed player j (being highly concerned about fairness) does not know whether the game he plays with player i is: (1) a Pareto coordination game, which arises when player i is also highly concerned about fairness; or (2) a game where defection is a strictly dominant strategy for player i, while player j still prefers to mimic the action selected by his opponent, which ensues when player i's concerns for fairness are low. Hence, the latter game can neither be interpreted as a standard PD game or as a Pareto coordination game.

[15] For simplicity, we ignore discounting. However, for completeness at the end of this section we demonstrates that the equilibrium predictions of Proposition 1 are unaffected by allowing for discounting. In addition, note that we use "pooling" equilibrium to refer to strategy profiles in which both types of informed player i cooperate during the first-period game. For robustness, we show that this pooling equilibrium survives Cho and Kreps' (1987) Intuitive Criterion; see Appendix 2. That appendix also provides conditions under which a "non-cooperative" pooling equilibrium —where both types of player i defect in the first period— can be sustained, and under which parameter values it survives Cho and Kreps' (1987) Intuitive Criterion.

the uninformed player j defects), and the disutility from guilt that player i suffers in the second period (when player j cooperates) are sufficiently high, i.e., if $\alpha_i^L + \beta_i^L \geq \frac{c+b-2d}{b-c}$. Combining this condition with the initial assumption of $\beta_i^L < \frac{b-a}{b-c}$, yields that α_i^L must satisfy $\alpha_i^L \geq \frac{a+c-2d}{b-c}$. Intuitively, for a separating equilibrium to exist, the concerned player needs to experience a small disutility from envy if he cooperates in the first-period PD game, i.e., $\alpha_i^H < \frac{a+c-2d}{b-c}$, while the unconcerned player must significantly suffer from envy, $\alpha_i^L \geq \frac{a+c-2d}{b-c}$. These conditions on the envy parameters however, violate our assumption that $\alpha_i^H > \alpha_i^L$, thus implying that such a separating strategy profile cannot be sustained in equilibrium.[16]

By contrast, a pooling equilibrium can be supported in which both types of privately informed player i cooperate in the first period. In particular, the uninformed player j cooperates in both the first period (given his relatively high prior, q) and in the second period, as long as player j observes that player i cooperated in the first stage.[17] Hence, both types of player i cooperate in order to be perceived as being a cooperative type by the uninformed player. The subsequent behavior of the informed player i in the second and final period, however, is strikingly different: while the highly concerned player cooperates once again, the unconcerned player defects. Intuitively, the unconcerned player cooperates during the first period in order to disguise himself as an individual with high fairness concerns thereby inducing the uninformed player to cooperate in period two while the unconcerned player takes the opportunity to defect in period two. The latter "backstabbing" behavior from the player with low concern for fairness is a commonly observed outcome in the final period of experiments involving finitely-repeated interaction, both for the PD and other simultaneous-move games as found in the studies cited in the introduction. In particular, many subjects initially play cooperatively but choose to defect in the last period, even when their opponent has cooperated in all prior periods.[18]

5.1.1 Discounting

In our previous analysis, we assumed for simplicity that players did not discount payoffs over the two periods of their interaction. Importantly, allowing for payoff discounting does not change the

[16] If players' guilt parameters satisfied $\beta_i^H > \beta_i^L$, but envy concerns were lower for those concerned about fairness than for unconcerned players, i.e., if $\alpha_i^H \leq \alpha_i^L$, then such a separating strategy profile could be sustained in equilibrium. However, following the experimental evidence by Bellemare, Kröger and van Soest (2008) as described at the beginning of this section, we think it is more plausible to assume that individuals with high concerns for inequity aversion exhibit both larger guilt and envy concerns than players with low concerns, thus implying that this case will not arise.

[17] If, instead, priors are sufficiently low, i.e., $q < q^{Sim}(\alpha_j, \beta_j)$, the uninformed player j defects in the first period of the game and, as a consequence, an alternative pooling equilibrium emerges in which the informed player i defects in the first period of interaction, both when he is concerned about fairness and when he is not. Player j's action in the second-period of interaction, however, depends on his off-the-equilibrium beliefs: when they are relatively high, the uninformed player j chooses to cooperate after observing that player i cooperated in the first period, but defects otherwise. In such a setting, player i does not find it profitable to cooperate in the first-period game, thus leading both types of player i to defect. For more details about this pooling equilibrium, see Appendix 1.

[18] If the simultaneous-move PD game is, instead, repeated for $T > 2$ periods, the informed player i becomes more attracted to cooperation, since his cooperation triggers a longer stream of cooperative outcomes, yielding an overall payoff that exceeds that from defecting under larger parameter conditions. Therefore, the pooling equilibrium can be sustained for less restrictive conditions as the number of repetitions increases.

set of equilibrium outcomes identified in Proposition 1. Nonetheless, the introduction of discounting implies that the pooling ("backstabbing") equilibrium can only be sustained under more restrictive conditions, i.e., if player i's guilt parameter, β_i^L, satisfies $\frac{b-a}{b-c} > \beta_i^L \geq \frac{a-b(1-\delta)-d\delta}{(b-c)(\delta-1)}$.[19] Intuitively, with discounting of future payoffs, the unconcerned player i finds the "backstabbing" strategy less attractive since the future benefit he obtains from misleading the uninformed player j and defecting while he cooperates, is now reduced.

5.1.2 Pareto improving effects

One might conjecture that the lack of information in the pooling equilibrium implies that the uninformed player's payoffs are worse under incomplete information than under complete information. The following corollary shows, however, that this conjecture is not necessarily true, by identifying conditions under which the uninformed player actually prefers to play the twice-repeated game under the incomplete information setting (as specified in the "backstabbing" equilibrium of Proposition 1) than under the complete information setting described in section 4.[20]

- **Corollary 1.** *The equilibrium expected utility of the privately informed player i is weakly larger in the incomplete than in the complete information game. However, the uninformed player j prefers to remain uninformed, i.e., his equilibrium expected utility is larger under the incomplete information game, if and only if his disutility from envy satisfies $\alpha_j \leq \frac{a+c-2d}{b-c}$.*

Summarizing, the introduction of incomplete information helps players to coordinate on the cooperative outcome (C,C) until the last period of the game, generating a larger overall utility for both players as long as the envy that uninformed player j suffers in the second-period game (when j cooperates while the unconcerned-type player i defects) is relatively small. Thus, incomplete information about one player's concerns for fairness allows for a Pareto improvement relative to the complete information setting where mutual defection (D,D) is the sole equilibrium possibility. Finally, note that the envy parameter restriction, $\alpha_j \leq \frac{a+c-2d}{b-c}$ is only compatible with the initial assumption on player j's preferences, $\alpha_j \geq \beta_j \geq \frac{b-a}{b-c}$, when the benefit that player j obtains from the first-period cooperative outcome (C,C) offsets the utility he obtains in the last period of play, i.e., if $2a - b > 2d - c$, where outcome (C,C) obtains when he faces a concerned player i but outcome (D,C) obtains when he faces an unconcerned player i who "backstabs" him. Otherwise, player j's equilibrium utility under incomplete information lies below his utility under complete

[19] For a detailed analysis of these conditions, see Appendix 3.

[20] In order to facilitate our utility comparisons, we hereafter assume that, in settings where both outcomes (C,C) and (D,D) can be sustained in equilibrium as in the shaded area of Figure 1 for the simultaneous-move PD game under complete information, players can resort to some coordination mechanism, such as social norms or a stochastic randomization, that enables them to coordinate on the efficient cooperative outcome (C,C). Alternatively, if a pre-play communication stage exists, Demicheli and Weibull (2008) show that, in the context of Pareto coordination games such as that arising when $\beta_i, \beta_j \geq \frac{b-a}{b-c}$ in our model, every evolutionary stable equilibrium induces players to asymptotically coordinate on the cooperative outcome (C,C).

information, for all admissible envy concerns, and only player i would benefit from the introduction of incomplete information.

5.2 Twice-repeated sequential-move PD game

As suggested in the introduction, the above "backstabbing" equilibrium in which the unconcerned player disguises himself as a highly concerned player appears similar to results found in Healy (2007). However, our results differ from Healy (2007) in two respects. First, Healy simply postulates the existence of selfish and reciprocal types whereas we rationalize these different types using empirically plausible Fehr-Schmidt (1999) inequity aversion (social) preferences. Second and more importantly, Healy's results are in the context of a finitely-repeated but *sequential*-move game – the "gift-exchange" game– where he shows that both worker types, "selfish" and "reciprocators," respond to high wage offers from (first-mover) employers by exerting high levels of (costly) work effort and continue to do so until the last periods of interaction, when selfish workers switch from high to low effort.[21] However, as we show below, Healy's pooling equilibrium for the sequential move game cannot be sustained under the same conditions as given in our Proposition 1 for the finitely repeated, simultaneous move game.

To facilitate a comparison of our results with those of Healy (2007), we consider a twice-repeated game of asymmetric incomplete information in which the stage game where players interact is the *sequential*-move PD, rather than the simultaneous-move version studied in the previous section. For now, we shall suppose that it is the *second* mover who is privately informed about his concerns for fairness, either high, $\beta_2^H \geq \frac{b-a}{b-c}$, or low, $\beta_2^L < \frac{b-a}{b-c}$, with probabilities q and $1-q$, respectively. By contrast, the uninformed first mover's high concern for fairness, $\beta_1^H \geq \frac{b-a}{b-c}$, is assumed to be common knowledge among the players. Under these assumptions, the following proposition provides conditions under which a cooperative pooling outcome can be sustained as a PBE.

Proposition 2. *Suppose* $q > q^{Seq}(\alpha_1) \equiv \frac{d-c+\alpha_1(b-c)}{a-c+\alpha_1(b-c)}$. *Then there exists a pooling PBE in the twice-repeated sequential-move PD game under incomplete information where:*

1. *In the first-period game, the uninformed first mover cooperates, while in the second-period game he cooperates only after observing that the second mover cooperated in the first-period game; otherwise the first mover defects in the second-period game, given beliefs* $\mu\left(\beta_2^H|C\right) = q > q^{Seq}(\alpha_1) > \mu\left(\beta_2^H|D\right)$.

2. *The informed second mover cooperates in the first-period game regardless of his type. In the second-period game, the informed second mover cooperates (defects) when he is concerned (unconcerned, respectively) about fairness.*

Furthermore, $q^{Seq}(\alpha_1) > q^{Sim}(\alpha_1, \beta_1)$ *if and only if the uninformed player is highly concerned about fairness,* $\beta_1^H \geq \frac{b-d}{b-c}$.

[21] In particular, this result corresponds to proposition 1 in Healy (2007) where the past actions of all players are observable, but workers' types are not.

The last statement implies that the cooperative pooling equilibrium can be sustained under a larger set of parameter values in the simultaneous-move PD game than in the sequential-move PD game, i.e., $q > q^{Seq}(\alpha_1) > q^{Sim}(\alpha_1, \beta_1)$, when the uninformed first mover is highly concerned about fairness. Intuitively, when the uninformed first mover interacts in a sequential-move game, he can anticipate that his defection will be responded to with defection by the second mover, regardless of the second mover's type, leading to a payoff of d. By contrast, in the simultaneous version of the game, his defection yields an expected utility of $q[b-\beta_1^H(b-c)]+(1-q)d$. Since his guilt feeling is substantial, i.e., $\beta_1^H \geq \frac{b-d}{b-c}$ implies $q[b-\beta_1^H(b-c)]+(1-q)d < d$, he experiences a larger disutility from playing defection in the simultaneous than in the sequential game, making defection more attractive in the sequential version of the game. This result yields the testable implication that defection (cooperation) should be more frequently observed (rarely observed, respectively) during the first periods of play of a finitely-repeated sequential-move PD game, as studied in Healy (2007), than in the simultaneous-move version.

Furthermore, note that a given reduction in the uninformed player's envy, α_1, produces a larger decrease in $q^{Sim}(\alpha_1, \beta_1)$ than in $q^{Seq}(\alpha_1)$, making the cooperative pooling equilibrium sustainable under a larger set of parameter conditions in the simultaneous-move than in the sequential-move version of the PD game. If, by contrast, the uninformed player is not highly concerned about social preferences, i.e., if $\frac{b-a}{b-c} \leq \beta_1^H < \frac{b-d}{b-c}$, then the cutoffs satisfy $q^{Sim}(\alpha_1, \beta_1) > q^{Seq}(\alpha_1)$, and the cooperative equilibrium can be sustained under a larger set of parameter conditions in the sequential-move game than in the simultaneous-move game.

Similarly to the simultaneous-move PD game, the following corollary shows that the introduction of incomplete information in the sequential-move PD game can actually help players coordinate on the cooperative outcome until the last period of their interaction, resulting in a Pareto improvement relative to the complete information setting.

Corollary 2. *The equilibrium expected utility of the informed second mover is weakly larger in the incomplete than in the complete information game. However, the uninformed first mover prefers to remain uninformed, i.e., his equilibrium expected utility is larger under the incomplete information game, if and only if his disutility from envy is sufficiently low, i.e., if* $\alpha_1 \leq \frac{a+c-2d}{b-c}$.

5.3 Changing the order of moves

We next consider equilibrium play in the sequential PD game when the player possessing private information about his social preferences is switched, from the second mover to the first mover in the sequential-move PD game.

Corollary 3. *If players interact in a twice-repeated, sequential-move PD game where the first mover is privately informed about his concerns for fairness while the second mover is uninformed, only the following pooling strategy profile can be supported as a PBE of the signaling game: In the first-period game, the first mover cooperates regardless of his type while the second mover responds*

by mimicking the first mover's action, i.e., cooperating (defecting) after observing cooperation (defection, respectively). A similar argument applies to the second-period game, for all possible second mover beliefs $\mu\left(\beta_i^H|C\right) = q \in (0,1)$ *and* $\mu\left(\beta_i^H|D\right) \in (0,1)$.

When the uninformed player is the second mover, he can react to any deviation toward defection by the informed first mover in both the first and second periods of the game, since he observes the first-mover's action before choosing his own action. This time structure thus protects the uninformed second mover from any defection by the informed first mover. This differs from the sequential PD game analyzed in Proposition 2, where the uninformed player was the first mover and he could still suffer a "backstabbing" defection by the second mover in the last period of the game.[22]

6 Summary and Directions for Further Research

Incomplete information. There is large body of evidence from experimental economics studies suggesting that individuals exhibit other-regarding or "social" preferences as opposed to the more standard self-regarding preferences. Social preferences have been formally modeled in an effort to explain why experimental data often departs from equilibrium predictions. To date, most models of social preferences have assumed that players can perfectly observe each others' social preferences. This might be a reasonable assumption in strategic environments where players have been interacting with one another for long periods of time, allowing their social preferences to be revealed through their prior action choices. Nonetheless, in contexts where such a long history of play is not available, incomplete information regarding social preferences would seem to be the more reasonable assumption.

Backstabbing behavior in both simultaneous and sequential games. In this paper we examine how equilibrium play in the simultaneous and sequential-move PD game is affected by the introduction of incomplete information about players' social preferences. Our results identify a pooling ("backstabbing") equilibrium in which a player unconcerned about fairness (who is not inequity averse) initially cooperates in order to mislead an uninformed player. This pooling equilibrium may help to explain subjects' end-game behavior when they suddenly act non-cooperatively in the final period of finitely repeated experimental games. Despite the presence of such end-game defection, our analysis also indicates that, when envy concerns are not very high, the incomplete information environment can actually result in a Pareto improvement relative to complete information environments in which the cooperative outcome is never sustainable for any length of time. We also demonstrate that this same type of pooling equilibrium can be supported in repeated games where the stage game is a sequential-move version of the PD, although under different parameter conditions. The latter set-up allows us to compare our results with existing results found in

[22] Note that, in the context of the simultaneous-move PD game, our equilibrium results in section 5.1 would not be affected if we modified which player holds private information about his social preferences, either player 1 (the row player) or 2 (the column player), since our results in those propositions are valid for any player $i = \{1,2\}$ and $j \neq i$.

the literature and to predict which version of the game (simultaneous or sequential) can sustain cooperation under a larger set of parameter values in experimental settings.

Further research. In this paper we have focused on a single game –the Prisoner's Dilemma– in which the conflict between individual and social preferences is potentially large. Our analysis can nonetheless be directly applied to other strategic environments where social preferences have been extensively studied, for instance in public good and bargaining games. Such extensions would provide a better understanding of whether incomplete information about concerns for fairness results in greater cooperation and how players might use their actions to reveal or conceal their social preferences to other players. Finally, our equilibrium results offer several predictions that could be tested in a laboratory experiment. In particular, subjects' social preferences could first be estimated based on their choices in simple, complete-information games, such as the "dictator" game. Afterwards, subjects with high concerns for fairness could be paired either with subjects with high or low concerns for fairness to play the twice-repeated signaling game where the stage game is either the simultaneous-move or sequential-move version of the PD game. The subjects' choices could then be examined against our equilibrium predictions. We leave such a test to future research.

Appendix

Appendix 1 - Noncooperative pooling equilibria

Proposition A. *A pooling PBE can be sustained in which player i defects in the first period both when he is concerned about fairness and when he is not, and:*

a. *Player i defects in the second period, both when he is concerned about fairness and when he is not. The uninformed player j defects in the first and second period, regardless of player i's choices during the first stage, given beliefs $\mu\left(\beta_i^H|D\right) = q < q^{Sim}(\alpha_j, \beta_j)$ and $\mu\left(\beta_i^H|C\right) < q^{Sim}(\alpha_j, \beta_j)$; and*

b. *Player i defects in the second period when he is unconcerned about fairness, but cooperates in equilibrium otherwise given $\alpha_i^H \geq \frac{a+c-2d}{b-c}$ and $\alpha_i^L + \beta_i^L \geq \frac{c+b-2b}{b-c}$. The uninformed player j defects in the first period. In the second period player j defects after observing that player i defects in the first period but cooperates otherwise, given beliefs $\mu\left(\beta_i^H|D\right) = q < q^{Sim}(\alpha_j, \beta_j) < \mu\left(\beta_i^H|C\right)$.*

c. *Player i defects in the second period when he is concerned about fairness, but cooperates in equilibrium otherwise given $\frac{b-a}{b-c} \leq \beta_i^H < \frac{b-d}{b-c}$. The uninformed player j cooperates in the first period. In the second period, player j cooperates after observing that player i defects in the first period but defects otherwise, given beliefs $\mu\left(\beta_i^H|D\right) = q \geq q^{Sim}(\alpha_j, \beta_j) > \mu\left(\beta_i^H|C\right)$.*

Proof. Let us investigate the pooling equilibrium in which both types of informed player i defect in the first period of the game. First, note that after observing an action from player i in the first period, player j's beliefs in this pooling equilibrium are $\mu(\beta_i^H|C) \equiv \mu \in [0,1]$ and $\mu(\beta_i^H|D) = q$. Given these beliefs, let us now analyze player j's best response in the second period of the game. In particular, after observing a D choice in the first period (in equilibrium), player j cannot infer player i's social preferences and must therefore make her second period choice according to an expected utility comparison. In particular, player j cooperates in the second period if

$$qa + (1-q)[c - \alpha_j(b-c)] \geq q[b - \beta_j(b-c)] + (1-q)d.$$

That is, if $q \geq \frac{d-c+\alpha_j(b-c)}{a+d-c-b+(\alpha_j+\beta_j)(b-c)} \equiv q^{Sim}(\alpha_j, \beta_j)$. Note that this cutoff strategy coincides with the one he uses when choosing between C and D at the beginning of the first period. After observing C being played in the first-period stage game (which occurs off-the-equilibrium) player j cannot infer player i's social preferences either, and must therefore choose C or D in the second period according to an expected utility comparison. Specifically, player j cooperates in the second period if and only if

$$\mu a + (1-\mu)[c - \alpha_j(b-c)] \geq \mu[b - \beta_j(b-c)] + (1-\mu)d.$$

That is, if $\mu \geq q^{Sim}(\alpha_j, \beta_j)$. Let us now investigate the informed player i's actions during the first period:

1. If $q, \mu \geq q^{Sim}(\alpha_j, \beta_j)$ player j cooperates in the first period of the game, as well as in the second period, both after observing that player i selects C and D. The highly-concerned player i cooperates given that $a + a \geq b - \beta_i^H(b-c) + a$, which holds since $\beta_i^H \geq \frac{b-a}{b-c}$ by definition. Hence, the prescribed strategy profile cannot be supported as a pooling PBE if $q, \mu \geq q^{Sim}(\alpha_j, \beta_j)$.

2. If $q, \mu < q^{Sim}(\alpha_j, \beta_j)$ player j defects in the first period of the game, as well as in the second period, both after observing that player i selects C and D. On the one hand, the highly-concerned player i defects if $c - \alpha_i^H(b-c) + d \leq d + d$, which implies $\alpha_i^H \geq 0 > \frac{c-d}{b-c}$, which holds by definition. On the other hand, the unconcerned player i defects if $c - \alpha_i^L(b-c) + d \leq d + d$, which implies $\alpha_i^L \geq 0 > \frac{c-d}{b-c}$, which also holds by definition. Therefore, this pooling strategy profile can be sustained as a PBE if $q, \mu < q^{Sim}(\alpha_j, \beta_j)$; as described in Proposition A(a).

3. If $q \geq q^{Sim}(\alpha_j, \beta_j) > \mu$ player j cooperates in the first period of the game, as well as in the second period but only if he observes that player i chose D in the first period. On the one hand, the highly-concerned player i defects if $a + d \leq b - \beta_i^H(b-c) + a$, which is satisfied if $\beta_i^H \leq \frac{b-d}{b-c}$, where $\frac{b-a}{b-c} \leq \beta_i^H \leq \frac{b-d}{b-c}$. On the other hand, the unconcerned player i defects if $a + d \leq b - \beta_i^L(b-c) + b - \beta_i^L(b-c)$, which holds if $\beta_i^L < \frac{2b-a-d}{2(b-c)}$, which is satisfied since $\beta_i^L < \frac{b-a}{b-c} < \frac{2b-a-d}{2(b-c)}$. Hence, this pooling strategy profile can be supported as PBE if $\frac{b-a}{b-c} \leq \beta_i^H \leq \frac{b-d}{b-c}$ and $q \geq q^{Sim}(\alpha_j, \beta_j) > \mu$; as described in Proposition A(c).

4. If $q < q^{Sim}(\alpha_j, \beta_j) \leq \mu$ player j defects in the first period of the game, as well as in the second period but only if he observes that player i chose C in the first period. On the one hand, the informed highly-concerned player i defects if $c - \alpha_i^H(b-c) + a \leq d + d$, or $\alpha_i^H \geq \frac{a+c-2d}{b-c}$, where $\frac{a+c-2d}{b-c} > \frac{b-a}{b-c}$ if $(b-c) < 2(a-d)$. On the other hand, the unconcerned player i defects if $c - \alpha_i^L(b-c) + b - \beta_i^L(b-c) \leq d + d$, or $\alpha_i^L + \beta_i^L \geq \frac{c+b-2d}{b-c}$. Therefore, the prescribed strategy profile can be sustained as a pooling PBE if $q < q^{Sim}(\alpha_j, \beta_j) \leq \mu$, the social preference parameters of the highly concerned player i satisfy $\alpha_i^H \geq \frac{a+c-2d}{b-c}$ and those of the relatively unconcerned player i satisfy $\alpha_i^L + \beta_i^L \geq \frac{c+b-2d}{b-c}$; as described in Proposition A(b). ■

In the pooling equilibrium in which both types of player i defect (part a of Proposition A), the uninformed player j's beliefs are so "pessimistic" that he chooses to defect in the second stage of the game, regardless of player i's choices in the first period. Consequently, player i defects, both when he is concerned and when he is unconcerned about fairness. The strategy profile in part (b) describes a similar pooling equilibrium to that in part (a), but in which player j's off-the-equilibrium-path beliefs are sufficiently high to induce him to cooperate after observing cooperation. Thus, player i's choice induces player j to cooperate after observing cooperation but to defect otherwise. Consequently, deviating towards cooperation becomes a more attractive option than in part (a), where all actions are responded to with defection in the second period. In order to support the pooling equilibrium where both types of player i defect, the gain that player i

obtains from second period cooperation cannot offset the disutility from envy experienced from cooperating in the first period, yielding outcome (C,D). In particular, we require $\alpha_i^H \geq \frac{a+c-2d}{b-c}$ and $\alpha_i^L + \beta_i^L \geq \frac{c+b-2b}{b-c}$. Finally, in the pooling equilibrium described in part (c), the uninformed player j interprets a deviation towards cooperation as most likely originating from an unconcerned player i, i.e., $\mu\left(\beta_i^H|C\right) < q^{Sim}(\alpha_j, \beta_j)$, which are rather "insensible" off-the-equilibrium beliefs. Appendix 2 below confirms this suspicion by showing that the pooling equilibrium in part (c) does not survive Cho and Kreps' (1987) Intuitive Criterion. By contrast, we demonstrate that all other equilibria survive this refinement criterion.

6.1 Appendix 2 - Equilibrium refinement

Lemma A. *The pooling PBE described in Proposition 1 survives Cho and Kreps' (1987) Intuitive Criterion under all parameter values. All PBEs in Proposition A (Appendix 1) also survive the Intuitive Criterion, except for the pooling equilibrium described in part (c).*

Proof. *Proposition 1*. Let us first check whether the pooling PBE where both types of player i cooperate in the first period under $q \geq q^{Sim}(\alpha_j, \beta_j)$ survives the Intuitive Criterion. Regarding the unconcerned player i, if he deviates towards defection, the highest utility he can obtain is $b - \beta_i^L(b-c) + b - \beta_i^L(b-c)$, which exceeds his equilibrium utility of $a + b - \beta_i^L(b-c)$. Regarding the highly-concerned player i, if he deviates to defection, the highest utility he can obtain is $a + a$, which coincides with his equilibrium utility from cooperating. Hence only the unconcerned player i has incentives to deviate towards defection, allowing the uninformed player j to restrict his posterior beliefs to $\mu(\beta_i^H|D) = 0$. Hence, a defection can only originate from an unconcerned player i, inducing player j to defect in the second-period stage game, yielding a total utility for the unconcerned player i of $b - \beta_i^L(b-c) + d$, which does not exceed his equilibrium utility given that $a + b - \beta_i^L(b-c) \geq b - \beta_i^L(b-c) + d$ since $a \geq d$. Hence no type of player i wants to deviate from the pooling PBE where both types cooperate, and therefore this pooling equilibrium survives the Intuitive Criterion.

Proposition A, part a. Let us now check whether the pooling equilibrium in which both types of player i defect under $q, \mu < q^{Sim}(\alpha_j, \beta_j)$ survives the Intuitive Criterion. Regarding the highly-concerned player i, if he deviates towards cooperation the highest utility he can obtain is $c - \alpha_i^H(b-c) + a$, which exceeds his equilibrium utility of $d + d$ if $\alpha_i^H < \frac{a+c-2d}{b-c}$. [Recall that condition $\alpha_i^H < \frac{a+c-2d}{b-c}$ is compatible with the initial assumption of $\beta_i^H \geq \frac{b-a}{b-c}$ if $b - c < 2(a-d)$]. Regarding the unconcerned player i, if he deviates towards cooperation the highest utility he can obtain is $c - \alpha_i^L(b-c) + b - \beta_i^L(b-c)$, which exceeds his equilibrium utility of $d + d$ only if $\alpha_i^L + \beta_i^L < \frac{c+b-2d}{b-c}$. Using the conditions we found for the concerned and unconcerned player i, let us examine under which cases this pooling equilibrium survives the Intuitive Criterion:

1. When both $\alpha_i^H < \frac{a+c-2d}{b-c}$ and $\alpha_i^L + \beta_i^L < \frac{c+b-2d}{b-c}$ hold, both types of player i have incentives to deviate, and the uninformed player j cannot restrict his beliefs upon observing a deviation to cooperation (off-the-equilibrium). As a consequence, no type of player i has incentives

to modify his equilibrium action, and hence this pooling equilibrium survives the Intuitive Criterion.

2. When condition $\alpha_i^H < \frac{a+c-2d}{b-c}$ holds but $\alpha_i^L + \beta_i^L < \frac{c+b-2d}{b-c}$ does not, the concerned player i has incentives to deviate towards cooperation but the unconcerned type does not. Conditions $\alpha_i^H < \frac{a+c-2d}{b-c}$ and $\alpha_i^L + \beta_i^L > \frac{c+b-2d}{b-c}$, however, are incompatible with the initial assumption of $\beta_i^L < \frac{b-a}{b-c}$. (In particular, combining $\alpha_i^L + \beta_i^L \geq \frac{c+b-2d}{b-c}$ with the assumption of $\beta_i^L < \frac{b-a}{b-c}$, we obtain that α_i^L must satisfy $\alpha_i^L \geq \frac{a+c-2d}{b-c}$. This result is, however, incompatible with condition $\alpha_i^H < \frac{a+c-2d}{b-c}$ that we obtained for the concerned player i since, by definition, envy concerns must satisfy $\alpha_i^H > \alpha_i^L$.) Hence, these parameter combinations are not feasible.

3. When neither condition $\alpha_i^H < \frac{a+c-2d}{b-c}$ nor $\alpha_i^L + \beta_i^L < \frac{c+b-2d}{b-c}$ hold, then neither type of player i has incentives to deviate. Consequently, the uninformed player j cannot restrict his off-the-equilibrium beliefs, and no type of player i has incentives to change his equilibrium behavior. Therefore, this pooling equilibrium survives the Intuitive Criterion.

4. When condition $\alpha_i^H < \frac{a+c-2d}{b-c}$ does not hold but $\alpha_i^L + \beta_i^L < \frac{c+b-2d}{b-c}$ does, the concerned player i does not have incentives to deviate towards cooperation but the unconcerned player i does. Therefore, the uninformed player j restricts his off-the-equilibrium beliefs to $\mu(\beta_i^H|C) = 0$ and $\mu(\beta_i^H|D) = 1$, since cooperation can only originate from the unconcerned player i. Consequently, player j cooperates in the second period after observing D but defects after observing C. Conditional on this response by player j, the highly-concerned player i does not have incentives to deviate towards cooperation since he would obtain $c - \alpha_i^H(b-c) + d$, which does not exceed his equilibrium utility of $d + d$. Similarly, the unconcerned player i does not have incentives to deviate towards cooperation, since he would obtain $c - \alpha_i^L(b-c) + d$, which is lower than his equilibrium utility of $d + d$. Hence the pooling PBE in which both types of player i defect when $q, \mu < q^{Sim}(\alpha_j, \beta_j)$ survives the Intuitive Criterion if $\alpha_i^H \geq \frac{a+c-2d}{b-c}$ and $\alpha_i^L + \beta_i^L < \frac{c+b-2d}{b-c}$.

Proposition A, part b. Let us now examine the pooling equilibrium in which both types of player i defect under $\mu \geq q^{Sim}(\alpha_j, \beta_j) > q$. Regarding the highly-concerned player i, if he deviates towards cooperation the highest utility he can obtain is $c + \alpha_i^H(b-c) + a$, which exceeds his equilibrium utility of $d + d$ only if $\alpha_i^H < \frac{a+c-2d}{b-c}$, which violates the equilibrium conditions. Hence, the highly concerned player i does not have incentives to deviate. Regarding the unconcerned player i, if he deviates towards cooperation the highest utility he can obtain is $c + \alpha_i^L(b-c) + b - \beta_i^L(b-c)$, which exceeds his equilibrium utility of $d + d$ only if $\alpha_i^L + \beta_i^L < \frac{c+b-2d}{b-c}$, which violates the equilibrium conditions. Therefore, the unconcerned player i does not deviate towards cooperation either. Since no type of player i deviates from his equilibrium action, player j's posterior beliefs are unmodified, and this pooling PBE survives the Intuitive Criterion if $\mu > q^{Sim}(\alpha_j, \beta_j) > q$.

Proposition A, part c. Let us finally examine the pooling equilibrium in which both types of player i defect under $q \geq q^{Sim}(\alpha_j, \beta_j) > \mu$. Regarding the highly-concerned player i, if he deviates

towards cooperation, the highest utility he can obtain is $a + a$, which exceeds his equilibrium utility of $b - \beta_i^H(b-c) + a$ since $\beta_i^H \geq \frac{b-a}{b-c}$ by definition. Hence, the highly-concerned player has incentives to deviate. Regarding the unconcerned player i, if he deviates towards cooperation, the highest utility he can achieve is $a + b - \beta_i^L(b-c)$, which does not exceed his equilibrium utility of $b - \beta_i^L(b-c) + b - \beta_i^L(b-c)$ since $\beta_i^L < \frac{b-a}{b-c}$. Therefore, only the concerned player i has incentives to deviate towards cooperation. Hence, the uninformed player j can restrict his off-the-equilibrium path beliefs after observing cooperation to $\mu(\beta_i^H|C) = 1$, inducing him to cooperate in the second period game as a consequence. Thus, the concerned player i obtains a higher utility by deviating from his equilibrium action of defection towards cooperation, and consequently, the pooling equilibrium in which both types of player i defect under $q \geq q^{Sim}(\alpha_j, \beta_j) > \mu$ violates the Intuitive Criterion. Finally, note that since we consider just two types of privately informed player i, who are either "concerned" or "unconcerned" about fairness, the application of Cho and Kreps' Intuitive Criterion and Banks and Sobel's (1987) Universal Divinity criterion (also referred to as the D1-Criterion) would lead to the same equilibrium predictions. ■

6.2 Appendix 3 - Role of discounting

Separating equilibrium. Let us first analyze the separating strategy profile in which the highly-concerned player i cooperates but the unconcerned player i defects. First, note that the uninformed player j's beliefs and best response in the second-period game coincide with that given in Proposition 1, since he does not need to consider discounting in this case. In the first period of the game, however, player j selects C or D based on his expected utility, thus cooperating if and only if

$$qa + (1-q)[c - \alpha_j(b-c)] \geq q[b - \beta_j(b-c)] + (1-q)d.$$

or $q \geq q^{Sim}(\alpha_j, \beta_j)$. As in the proof of Proposition 1, let us now investigate the informed player i's action during the first-period stage game:

1. If $q > q^{Sim}(\alpha_j, \beta_j)$, the uninformed player j cooperates during the first period. On the one hand, the highly-concerned player i thus cooperates if $a + \delta a \geq b - \beta_i^H(b-c) + \delta d$, since cooperation (defection) is responded to with cooperation (defection, respectively) in the subsequent period. Solving for β_i^H, we obtain that player i cooperates if $\beta_i^H \geq \frac{b-\delta d-a(1+\delta)}{b-c}$. And since $\beta_i^H \geq \frac{b-a}{b-c} > \frac{b-d-2a}{b-c} > \frac{b-\delta d-a(1+\delta)}{b-c}$ by definition, the above condition is satisfied, implying that the highly-concerned player i cooperates in the first period, as prescribed in this separating strategy profile. On the other hand, the unconcerned player i cooperates if $a + \delta[b - \beta_i^L(b-c)] \geq [b - \beta_i^L(b-c)] + \delta d$, which simplifies to $\delta \geq \frac{a-\left[b-\beta_i^L(b-c)\right]}{d-\left[b-\beta_i^L(b-c)\right]}$. However, since $a - \left[b - \beta_i^L(b-c)\right] < 0$ because $\beta_i^L < \frac{b-a}{b-c}$, and $d - \left[b - \beta_i^L(b-c)\right] > 0$ since $\beta_i^L < \frac{b-a}{b-c} < \frac{b-d}{b-c}$, then ratio $\frac{a-\left[b-\beta_i^L(b-c)\right]}{d-\left[b-\beta_i^L(b-c)\right]} < 0$, implying that the condition on the discount factor $\delta \geq \frac{a-\left[b-\beta_i^L(b-c)\right]}{d-\left[b-\beta_i^L(b-c)\right]}$ holds for all $\delta \in [0,1]$. Hence, the unconcerned player i also cooperates in the first period, which violates the above separating strategy profile.

2. If $q < q^{Sim}(\alpha_j, \beta_j)$, the uninformed player j defects during the first period. On the one hand, the highly-concerned player i cooperates if $c - \alpha_i^H(b-c) + \delta a \geq d + \delta d$, or $\alpha_i^H \leq \frac{\delta a + c - d(1+\delta)}{b-c}$; note that this cutoff is higher than $\frac{b-a}{b-c}$ if and only if $a > \frac{d(1+\delta)+b-c}{1+\delta}$. On the other hand, the unconcerned player i defects if $c - \alpha_i^L(b-c) + \delta\left[b - \beta_i^L(b-c)\right] \leq d + \delta d$, which implies $\alpha_i^L \geq \frac{c+b(1-\beta_i^L)\delta + c\delta\beta_i^L - d(1+\delta)}{b-c}$. Therefore, for a separating equilibrium to exist in which player i cooperates if and only if he is highly concerned about fairness, prior probability q must satisfy $q < q^{Sim}(\alpha_j, \beta_j)$, envy must satisfy $\alpha_i^H \leq \frac{\delta a + c - d(1+\delta)}{b-c}$, and $\alpha_i^L + \beta_i^L \geq \frac{c+b(1-\beta_i^L)\delta + c\delta\beta_i^L - d(1+\delta)}{b-c}$. The initial assumption $\beta_i^L < \frac{b-a}{b-c}$ and condition $\alpha_i^L \geq \frac{c+b(1-\beta_i^L)\delta + c\delta\beta_i^L - d(1+\delta)}{b-c}$, however, are incompatible with $\alpha_i^H \leq \frac{\delta a + c - d(1+\delta)}{b-c}$. (In particular, combining $\alpha_i^L \geq \frac{c+b(1-\beta_i^L)\delta + c\delta\beta_i^L - d(1+\delta)}{b-c}$ with the assumption of $\beta_i^L < \frac{b-a}{b-c}$, we obtain that α_i^L must satisfy $\alpha_i^L \geq \frac{\delta a + c - d(1+\delta)}{b-c}$. This result is, however, incompatible with condition $\alpha_i^H \leq \frac{\delta a + c - d(1+\delta)}{b-c}$ that we obtained for the concerned player i since, by definition, envy concerns must satisfy $\alpha_i^H > \alpha_i^L$.) As a consequence, this separating strategy profile cannot be sustained as a PBE of the game.

Pooling equilibrium. Let us analyze the pooling strategy profile where both types of informed player i cooperate in the first period of the game. Beliefs coincide with those in the pooling equilibrium of Proposition 1, i.e., $\mu(\beta_i^H|C) = q$ (in equilibrium) and $\mu(\beta_i^H|D) \equiv \gamma \in [0,1]$ (off-the-equilibrium path). Given these beliefs, the uninformed player j cannot infer anything about the type of player i after observing that player i chose C in the first period (in equilibrium) and must therefore select C or D in the second period according to an expected utility comparison similar to that given in the proof of Proposition 1. Specifically, player j cooperates in the second-period PD game if and only if

$$qa + (1-q)[c - \alpha_j(b-c)] \geq q[b - \beta_j(b-c)] + (1-q)d.$$

That is, if $q \geq q^{Sim}(\alpha_j, \beta_j)$. Similarly, in the first period stage game, player j is uninformed about player i's concern for fairness, and hence chooses to cooperate in the first-period PD game according to the same cutoff strategy that conditions on the prior probability that player i's type is high. After observing that player i selected D in the first-period stage game (off-the-equilibrium path behavior) player j cannot infer player i's social preferences either, and must therefore select C or D in the second period of the game according to an expected utility comparison. Specifically, player j cooperates in the second period if and only if

$$\gamma a + (1-\gamma)[c - \alpha_j(b-c)] \geq \gamma[b - \beta_j(b-c)] + (1-\gamma)d.$$

That is, if $\gamma \geq q^{Sim}(\alpha_j, \beta_j)$. Let us now investigate the informed player i's actions during the first-period stage game:

1. If $q, \gamma \geq q^{Sim}(\alpha_j, \beta_j)$ player j cooperates in both the first and second periods of the game, both after observing that player i selects C and D. On the one hand, the highly-concerned

player i cooperates if $a+\delta a \geq b-\beta_i^H(b-c)+\delta a$, which holds since $\beta_i^H \geq \frac{b-a}{b-c}$ by definition. On the other hand, the unconcerned player i defects since $a+\delta\left[b-\beta_i^L(b-c)\right] \leq b-\beta_i^L(b-c)+\delta\left[b-\beta_i^L(b-c)\right]$, which holds given that $\beta_i^L < \frac{b-a}{b-c}$. Therefore, this pooling strategy profile *cannot* be supported if $q,\gamma \geq q^{Sim}(\alpha_j,\beta_j)$.

2. If $q,\gamma < q^{Sim}(\alpha_j,\beta_j)$ player j defects both in the first and second periods of the game, both after observing that player i selects C and D. The highly concerned player i defects, however, since $c-\alpha_i^H(b-c)+\delta d \leq d+\delta d$, which implies $\alpha_i^H \geq 0 \geq \frac{c-d}{b-c}$, which holds by definition. Hence, this pooling strategy profile *cannot* be sustained if $q,\gamma < q^{Sim}(\alpha_j,\beta_j)$.

3. If $q \geq q^{Sim}(\alpha_j,\beta_j) > \gamma$ player j cooperates in the first period of the game and in the second period he cooperates only after observing that player i chose C in the first-period stage game. On the one hand, the highly-concerned player i cooperates since $a+\delta a \geq b-\beta_i^H(b-c)+\delta d$, or $\beta_i^H \geq \frac{b+\delta d-a(1+\delta)}{b-c}$. Given that $\frac{b-a}{b-c} > \frac{b+d-2a}{b-c} > \frac{b+\delta d-a(1+\delta)}{b-c}$ for all $\delta \in [0,1]$, then condition $\beta_i^H \geq \frac{b+\delta d-a(1+\delta)}{b-c}$ is satisfied from the assumption $\beta_i^H \geq \frac{b-a}{b-c}$. On the other hand, the unconcerned player i cooperates if $a+\delta\left[b-\beta_i^L(b-c)\right] \geq b-\beta_i^L(b-c)+\delta d$, or $\beta_i^L \geq \frac{a-b(1-\delta)-d\delta}{(b-c)(\delta-1)}$. Moreover, since $\frac{b-a}{b-c} > \frac{a-b(1-\delta)-d\delta}{(b-c)(\delta-1)}$, then the initial condition on this unconcerned player i, $\beta_i^L < \frac{b-a}{b-c}$, implies that for all $\beta_i^L \in \left[\frac{a-b(1-\delta)-d\delta}{(b-c)(\delta-1)}, \frac{b-a}{b-c}\right)$ this type of player i also cooperates, and this pooling strategy profile *can* be supported as a PBE for $q \geq q^{Sim}(\alpha_j,\beta_j) > \gamma$. (As a remark, notice that the length of the interval of values for β_i^L for which the pooling equilibrium can be sustained, $\frac{b-a}{b-c} - \frac{a-b(1-\delta)-d\delta}{(b-c)(\delta-1)} = \frac{(a-d)\delta}{(b-c)(1-\delta)}$, collapses to zero when $\delta \to 0$, but becomes infinite (i.e., the pooling equilibrium can be supported for all values of β_i^L) when $\delta \to 1$.

4. If $q < q^{Sim}(\alpha_j,\beta_j) \leq \gamma$ player j defects in the first period of the game and in the second period he defects only after observing that player i selected C in the first-period stage game. The highly concerned player i defects, however, since $c-\alpha_i^H(b-c)+\delta a \leq d+\delta a$, which implies $\alpha_i^H \geq 0 \geq \frac{c-d}{b-c}$, which is satisfied by definition. Thus, this pooling strategy profile *cannot* be sustained if $q < q^{Sim}(\alpha_j,\beta_j) \leq \gamma$. ■

6.3 *Proof of Lemma 1*

From the text we know that the best response of the second mover (player j) is to select the same action as the first mover when $\beta_j \geq \frac{b-a}{b-c}$, but to defect when $\beta_j < \frac{b-a}{b-c}$, regardless of the action selected by the first mover. Formally, for any action a_i that the first mover selects, the second mover's best response function is:

$$a_j(a_i) = \begin{cases} C \text{ if } a_i = C \text{ and } \beta_j \geq \frac{b-a}{b-c}; \text{ and} \\ D \quad \text{otherwise} \end{cases}$$

When the second mover's fairness concerns satisfy $\beta_j \geq \frac{b-a}{b-c}$, the second mover responds by selecting the action chosen by the first mover. Therefore, the first mover's payoff from cooperating

(which is responded to with cooperation) is a, while that from defecting (which is responded to with defection) is d. Since $a > d$ by definition, the first mover prefers to cooperate when the second mover's concerns about fairness satisfy $\beta_j \geq \frac{b-a}{b-c}$, for any preference parameters of the first mover.

When, instead, the second mover's fairness concerns satisfy $\beta_j < \frac{b-a}{b-c}$, the second mover responds by defecting, regardless of the action previously selected by the first mover. In this case, if the first mover selects cooperation his utility is $c - \alpha_i(b-c)$, while if he defects his payoff is d. Since $c - \alpha_i(b-c) < d$ for any $\alpha_i \geq 0$, the first mover defects when the second mover's fairness concerns satisfy $\beta_j < \frac{b-a}{b-c}$. ■

6.4 *Proof of Proposition 1*

Separating equilibrium. Let us first analyze the separating strategy profile in which the highly-concerned player i cooperates but the unconcerned player i defects. First, note that after observing an action from player i in the first period of the game, player j's beliefs in this separating strategy profile are updated to $\mu(\beta_i^H|C) = 1$ and $\mu(\beta_i^H|D) = 0$. Given these beliefs, let us now analyze player j's best response in the second period of the game. In particular, after observing C in the first period, the uninformed player j believes that his opponent, i, is a highly concerned type (so that i will continue to select C in the second-period stage game). Since player j is highly concerned about fairness, $\beta_j \geq \frac{b-a}{b-c}$, it follows that $a > b - \beta_j(b-c)$ and so the uninformed player j will choose to cooperate in the second period of game. However, after observing a D in the first period, the uninformed player j believes that his opponent is an unconcerned type who will choose D. Given that $d > c - \alpha_j(b-c)$ by definition, the uninformed player j will choose to defect in the second-period stage game. Let us now examine the first period of the game. In the first period, the uninformed player j must select C or D based upon an expected utility comparison; specifically, j cooperates in the first period of the twice repeated PD game if and only if:

$$qa + (1-q)[c - \alpha_j(b-c)] \geq q[b - \beta_j(b-c)] + (1-q)d.$$

That is, if $q \geq q^{Sim}(\alpha_j, \beta_j)$. Let us now investigate the informed player i's action during the first-period stage game:

1. If $q > q^{Sim}(\alpha_j, \beta_j)$, the uninformed player j cooperates during the first period. On the one hand, the highly-concerned player i thus cooperates if $a + a \geq b - \beta_i^H(b-c) + d$, since cooperation (defection) is responded to with cooperation (defection, respectively) in the subsequent period. Solving for β_i^H, we obtain that player i cooperates if $\beta_i^H \geq \frac{b-d-2a}{b-c}$. Further, since $\beta_i^H \geq \frac{b-a}{b-c} > \frac{b-d-2a}{b-c}$ by definition, the above condition is satisfied, implying that the highly-concerned player i cooperates in the first period, as prescribed in this separating strategy profile. On the other hand, the unconcerned player i cooperates since $a + [b - \beta_i^L(b-c)] \geq [b - \beta_i^L(b-c)] + d$, which holds given that $a \geq d$. Hence, the unconcerned player i also cooperates in the first period, which violates the above separating strategy profile.

2. If $q < q^{Sim}(\alpha_j, \beta_j)$, the uninformed player j defects during the first period. On the one hand, the highly-concerned player i cooperates if $c-\alpha_i^H(b-c)+a \geq d+d$, or $\alpha_i^H \leq \frac{a+c-2d}{b-c}$; note that this cutoff is higher than $\frac{b-a}{b-c}$ if and only if $b-c < 2(a-d)$. On the other hand, the unconcerned player i defects if $c-\alpha_i^L(b-c)+b-\beta_i^L(b-c) \leq d+d$, which implies that $\alpha_i^L+\beta_i^L \geq \frac{c+b-2d}{b-c}$. Therefore, for a separating equilibrium to exist in which player i cooperates if and only if he is highly concerned about fairness, the prior probability q must satisfy $q < q^{Sim}(\alpha_j, \beta_j)$, envy concerns must satisfy $\alpha_i^H \leq \frac{a+c-2d}{b-c}$, and $\alpha_i^L+\beta_i^L \geq \frac{c+b-2d}{b-c}$. The initial assumption that $\beta_i^L < \frac{b-a}{b-c}$ and the condition $\alpha_i^L+\beta_i^L \geq \frac{c+b-2d}{b-c}$, however, are incompatible with $\alpha_i^H \leq \frac{a+c-2d}{b-c}$. (In particular, combining $\alpha_i^L+\beta_i^L \geq \frac{c+b-2d}{b-c}$ with the assumption of $\beta_i^L < \frac{b-a}{b-c}$, we obtain that α_i^L must satisfy $\alpha_i^L \geq \frac{a+c-2d}{b-c}$. This result is, however, incompatible with condition $\alpha_i^H < \frac{a+c-2d}{b-c}$ that we obtained for the concerned player i since, by definition, envy concerns must satisfy $\alpha_i^H > \alpha_i^L$.) As a consequence, this separating strategy profile cannot be sustained as a PBE of the game.

Let us now analyze the opposite separating strategy profile in which the highly-concerned player i defects but the unconcerned player i cooperates. First, note that after observing an action from player i in the first-period stage game, player j's beliefs in this separating strategy profile are updated to $\mu(\beta_i^H|C) = 0$ and $\mu(\beta_i^H|D) = 1$. Given these beliefs, let us now analyze player j's best response during the second period game. In particular, after observing C in the first-period stage game, he believes that his opponent is unconcerned about fairness, thus implying that his opponent will not cooperate in the second-period game. Player j defects as a consequence in the second-period PD game since $d > c-\alpha_j(b-c)$. By contrast, after observing D in the first period, player j believes that his opponent is highly concerned, and therefore that his opponent will cooperate in the second-period stage game. It follows that player j cooperates in the second period since $a \geq b-\beta_j(b-c)$ given that $\beta_j \geq \frac{b-a}{b-c}$ by definition. Let us now examine the first-period game. Regarding the uninformed player j, he must choose C or D according to an expected utility calculation. In particular, player j cooperates during the first-period PD game if and only if

$$q[c-\alpha_j(b-c)]+(1-q)a \geq qd+(1-q)[b-\beta_j(b-c)].$$

That is, if $q \geq q^{Sim}(\alpha_j, \beta_j)$. Let us now investigate the informed player i's actions during the first-period stage game:

1. If $q \geq q^{Sim}(\alpha_j, \beta_j)$, the uninformed player j cooperates during the first-period stage game. On the one hand, the highly-concerned player i defects (as prescribed) if $a+d \leq b-\beta_i^H(b-c)+a$, or $\beta_i^H \leq \frac{b-d}{b-c}$, where $\frac{b-d}{b-c} > \frac{b-a}{b-c}$, and hence player i defects when being highly concerned if $\frac{b-a}{b-c} \leq \beta_i^H \leq \frac{b-d}{b-c}$. On the other hand, the unconcerned player i cooperates (as prescribed) if $a+d \geq b-\beta_i^L(b-c)+b-\beta_i^L(b-c)$, which implies that $\beta_i^L \geq \frac{2b-a-d}{2(b-c)}$, which cannot hold since $\frac{2b-a-d}{2(b-c)} > \frac{b-a}{b-c}$ and $\beta_i^L < \frac{b-a}{b-c}$. Hence, this separating strategy profile cannot be sustained if $q \geq q^{Sim}(\alpha_j, \beta_j)$.

2. If $q < q^{Sim}(\alpha_j, \beta_j)$, the uninformed player j defects during the first-period stage game. On the one hand, the highly-concerned player i cooperates if $c - \alpha_i^H(b-c) + d \leq d + a$, or $\frac{c-a}{b-c} < 0 \leq \alpha_i^H$, which is true by definition. On the other hand, the unconcerned player i defects since $c - \alpha_i^L(b-c) + d \leq d + b - \beta_i^L(b-c)$, or $(b-c)(\beta_i^L - \alpha_i^L) \geq (b-c)$, which can only hold if $\beta_i^L \geq \alpha_i^L$, which is false by definition. Hence, this separating strategy profile cannot be supported if $q < q^{Sim}(\alpha_j, \beta_j)$.

Pooling equilibrium. Let us analyze the pooling strategy profile where both types of informed player i cooperate in the first period of the game. First, note that after observing an action from player i in the first-period stage game, player j's beliefs in this pooling strategy profile are $\mu(\beta_i^H|C) = q$ (in equilibrium) and $\mu(\beta_i^H|D) \equiv \gamma \in [0,1]$ (off-the-equilibrium path). Given these beliefs, let us now analyze player j's best response during the second period of the game. In particular, after observing C in the first period (in equilibrium) player j cannot infer player i's social preferences and must therefore select C or D in the second period according to an expected utility comparison. In particular, player j cooperates in the second-period PD game if and only if

$$qa + (1-q)[c - \alpha_j(b-c)] \geq q[b - \beta_j(b-c)] + (1-q)d.$$

That is, if $q \geq q^{Sim}(\alpha_j, \beta_j)$. Similarly, in the first period stage game, player j is uninformed about player i's concern for fairness, and hence chooses to cooperate in the first-period PD game according to the same cutoff strategy on the prior probability that player i's type is high. After observing that player i selected D in the first-period stage game (off-the-equilibrium) player j cannot infer player i's social preferences either, and must therefore select C or D in the second period of the game according to an expected utility comparison. Specifically, player j cooperates in the second period if and only if

$$\gamma a + (1-\gamma)[c - \alpha_j(b-c)] \geq \gamma[b - \beta_j(b-c)] + (1-\gamma)d.$$

That is, if $\gamma \geq q^{Sim}(\alpha_j, \beta_j)$. Let us now investigate the informed player i's actions during the first-period stage game:

1. If $q, \gamma \geq q^{Sim}(\alpha_j, \beta_j)$ player j cooperates in both the first and second periods of the game, both after observing that player i selects C and D. On the one hand, the highly-concerned player i cooperates if $a + a \geq b - \beta_i^H(b-c) + a$, which holds since $\beta_i^H \geq \frac{b-a}{b-c}$ by definition. On the other hand, the unconcerned player i defects since $a + b - \beta_i^L(b-c) \leq b - \beta_i^L(b-c) + b - \beta_i^L(b-c)$, which holds given that $\beta_i^L < \frac{b-a}{b-c}$. Therefore, this pooling strategy profile cannot be supported if $q, \gamma \geq q^{Sim}(\alpha_j, \beta_j)$.

2. If $q, \gamma < q^{Sim}(\alpha_j, \beta_j)$ player j defects both in the first and second periods of the game, both after observing that player i selects C and D. The highly concerned player i defects, however,

since $c - \alpha_i^H(b-c) + d \leq d + d$, which implies $\alpha_i^H \geq 0 \geq \frac{c-d}{b-c}$, which holds by definition. Hence, this pooling strategy profile cannot be sustained if $q, \gamma < q^{Sim}(\alpha_j, \beta_j)$.

3. If $q \geq q^{Sim}(\alpha_j, \beta_j) > \gamma$ player j cooperates in the first period of the game and in the second period he cooperates only after observing that player i chose C in the first-period stage game. On the one hand, the highly-concerned player i cooperates since $a + a \geq b - \beta_i^H(b-c) + d$, or $\beta_i^H \geq \frac{b+d-2a}{b-c}$. Given that $\frac{b-a}{b-c} > \frac{b+d-2a}{b-c}$, then condition $\beta_i^H \geq \frac{b+d-2a}{b-c}$ is satisfied from $\beta_i^H \geq \frac{b-a}{b-c}$. On the other hand, the unconcerned player i cooperates since $a + b - \beta_i^L(b-c) \geq b - \beta_i^L(b-c) + d$, or $a \geq d$. Hence, this pooling strategy profile can be supported as a PBE for $q \geq q^{Sim}(\alpha_j, \beta_j) > \gamma$.

4. If $q < q^{Sim}(\alpha_j, \beta_j) \leq \gamma$ player j defects in the first period of the game and in the second period he defects only after observing that player i selected C in the first-period stage game. The highly concerned player i defects, however, since $c - \alpha_i^H(b-c) + a \leq d + a$, which implies $\alpha_i^H \geq 0 \geq \frac{c-d}{b-c}$, which is satisfied by definition. Thus, this pooling strategy profile cannot be sustained if $q < q^{Sim}(\alpha_j, \beta_j) \leq \gamma$. ■

6.5 *Proof of Corollary 1*

Informed player i. Under complete information, when player i is highly concerned, i.e., when $\beta_i \geq \frac{b-a}{b-c}$, then players' preferences satisfy $\beta_i, \beta_j \geq \frac{b-a}{b-c}$, implying that outcome (C,C) can be sustained in pure strategies under complete information. (As discussed in section 5.1, we assume that, for simplicity, players can resort to some kind of coordination mechanism, such as social norms or a common randomization device by which players are able to coordinate on the efficient, cooperative outcome (C,C). Thus, if this simultaneous-move game is repeated twice, (C,C) would arise during both periods when both players' guilt aversion is sufficiently high.) In this context, player j's equilibrium utility is $a + a$. When player i is privately informed about his high concern for fairness, $\beta_i \geq \frac{b-a}{b-c}$, then outcome (C,C) is similarly played during both stages of the repeated game, thus yielding the same equilibrium utility for player i. If, by contrast, player i has low concerns for fairness, i.e., $\beta_i < \frac{b-a}{b-c}$, then players' preferences satisfy $\beta_j \geq \frac{b-a}{b-c} > \beta_i$, implying that outcome (D,D) can be supported under complete information. Hence, if this simultaneous-move game is repeated twice, (D,D) would arise during both periods when at least one player's guilt aversion is sufficiently high. Player i's equilibrium utility is thus $d + d$. When player i is privately informed about his low concern for fairness, $\beta_i < \frac{b-a}{b-c}$, then Proposition 1 shows that outcome (C,C) can be sustained in the first-period stage game, while (D,C) emerges in the second-period stage game, thus entailing an equilibrium utility of $a + b - \beta_i(b-c)$ for player i. Therefore, player i's equilibrium utility under incomplete information $a + b - \beta_i(b-c)$, is larger than that under complete information, $d + d$, if $\beta_i \leq \frac{a+b-2d}{b-c}$. Note that this condition holds for all $\beta_i < \frac{b-a}{b-c}$, given that $\frac{b-a}{b-c} < \frac{a+b-2d}{b-c}$, which is satisfied since $a > d$ by assumption. Therefore, player i's equilibrium utility is weakly larger under incomplete than complete information for all parameter values.

Uninformed player j. Let us investigate whether player j obtains a larger expected utility in

the complete or incomplete information version of the game. In particular, if we first evaluate player j's expected utility in the complete information game, but before being informed about which is player i's type (concerned or unconcerned), we obtain $q[a+a]+(1-q)[d+d]$. Specifically, the first component denotes the case in which player j is informed about player i being concerned for fairness, which yields the (C,C) outcome in both periods. The second component, however, reflects the case in which player j is informed about player i being unconcerned, whereby only the (D,D) outcome can be sustained under complete information; as shown in Section 4. By contrast, in the incomplete information game, the uninformed player j's expected utility is $q[a+a]+(1-q)[a+c-\alpha_j(b-c)]$, where the first component represents the payoffs he obtains when player i's type is high and (C,C) is played in both periods, while the second component describes his utility when player i's type is low, whereby (C,C) arises in the first period of interaction, but (D,C) emerges in the second period. Hence, player j's ex-ante expected utility is larger in the incomplete than the complete information version of the game if and only if

$$q[a+a]+(1-q)[a+c-\alpha_j(b-c)] > q[a+a]+(1-q)[d+d],$$

which simplifies to $a+c-\alpha_j(b-c) > 2d$, or $\alpha_j \leq \frac{a+c-2d}{b-c}$. Note that this condition on α_j is compatible with the initial assumption of $\alpha_j \geq \beta_j \geq \frac{b-a}{b-c}$ if $\frac{b-a}{b-c} < \frac{a+c-2d}{b-c}$, i.e., if $2a-b > 2d-c$, which holds when the payoff from promoting the cooperative outcome (C,C) is sufficiently high. ■

6.6 *Proof of Proposition 2*

Separating PBE. Let us first analyze the separating strategy profile where the (privately informed) second mover cooperates when being concerned about fairness but defects otherwise. First, note that after observing an action from the second mover in the first-period sequential-move game, the first mover's beliefs about β_2^H are updated according to Bayes' rule and become $\mu(\beta_2^H|C)=1$ and $\mu(\beta_2^H|D)=0$. Given these beliefs, the first mover cooperates in the second period after observing that the second mover cooperated in the first-period stage game, but defects otherwise. After these choices, the second mover reciprocates the first mover in the second-period stage game if the second mover's concerns are high, but defects otherwise.

During the first-period stage game, the second mover cooperates when being concerned but defects otherwise, as prescribed. Hence, the uninformed first mover cooperates if the expected utility from cooperation exceeds that from defection. That is,

$$qa+(1-q)[c-\alpha_1(b-c)] \geq qd+(1-q)d$$

or $q \geq q^{Seq}(\alpha_1)$. Let us finally investigate the second mover's behavior during the first-period game:

1. If $q \geq q^{Seq}(\alpha_1)$, the first mover cooperates in the first-period stage game. On one hand, the concerned second mover cooperates if $a+a \geq b-\beta_2^H(b-c)+d$, since defection is responded with defection in the subsequent stage game. Solving for β_2^H, we obtain that the highly-concerned

second mover cooperates (as prescribed) if $\beta_2^H \geq \frac{b-d-2a}{b-c}$. In addition, since $\frac{b-a}{b-c} > \frac{b-d-2a}{b-c}$ and $\beta_2^H \geq \frac{b-a}{b-c}$ by definition, condition $\beta_2^H \geq \frac{b-d-2a}{b-c}$ holds. Therefore, the concerned second mover cooperates in the first period, as prescribed in this separating strategy profile. On the other hand, the unconcerned second mover defects if $a + [b - \beta_2^L(b-c)] < [b - \beta_2^L(b-c)] + d$, i.e., $a < d$, which violates our initial assumptions. Hence, the unconcerned second mover also cooperates in the first period, implying that this separating strategy profile cannot be sustained as a PBE if priors satisfy $q \geq q^{Seq}(\alpha_1)$.

2. If $q < q^{Seq}(\alpha_1)$, the first mover defects in the first-period stage game. On one hand, the concerned second mover cooperates if $c - \alpha_2^H(b-c) + a \geq d + d$, or $\alpha_2^H \leq \frac{a+c-2d}{b-c}$. On the other hand, the unconcerned second mover defects if $c - \alpha_2^L(b-c) + b - \beta_2^L(b-c) \leq d + d$, which implies $\frac{c+b-2d}{b-c} \leq \alpha_2^L + \beta_2^L$. Thus, for a separating equilibrium to exist, in which the second mover cooperates only when he is concerned about fairness, the first-mover's beliefs must satisfy $q < q^{Seq}(\alpha_1)$ and parameter values must satisfy $\alpha_2^H \leq \frac{a+c-2d}{b-c}$ and $\frac{c+b-2d}{b-c} \leq \alpha_2^L + \beta_2^L$. The initial condition $\beta_2^L < \frac{b-a}{b-c}$ and $\alpha_2^L + \beta_2^L \geq \frac{c+b-2d}{b-c}$, however, are incompatible with $\alpha_2^H \leq \frac{a+c-2d}{b-c}$. [See the discussion in the proof of Proposition 1, specifically when checking for the existence of a separating equilibrium in the case that priors satisfy $q < q^{Sim}(\alpha_j, \beta_j)$.] As a consequence, this separating strategy profile cannot be sustained as a PBE of the game.

Pooling PBE. Let us now analyze the pooling strategy profile where both types of second mover cooperate in the first-period stage game. First, note that after observing an action from the second mover during the first-period sequential-move PD game, the first mover's beliefs about β_2^H in this pooling strategy profile cannot updated using Bayes' rule and hence are $\mu(\beta_2^H|C) = q$ (in equilibrium) and $\mu(\beta_2^H|D) \equiv \gamma \in [0,1]$ (off-the-equilibrium path). Given these beliefs, the first mover cooperates in the second-period stage game after observing that the second mover chose C (in equilibrium) in the first-period stage game if $q \geq q^{Seq}(\alpha_1)$. If the first mover observes the second mover selecting D in the first-period stage game (off-the-equilibrium path), then he cooperates in the second-period stage game if and only if

$$\gamma a + (1-\gamma)[c - \alpha_1(b-c)] \geq \gamma d + (1-\gamma)d.$$

That is, if $\gamma \geq q^{Seq}(\alpha_1)$. Let us now investigate the informed player (the second mover) during the first-period sequential-move game:

1. If $q, \gamma \geq q^{Seq}(\alpha_1)$ the first mover cooperates in both the first- and second-period stage game after observing any action from the second mover in the first-period stage game. On the one hand, the concerned second mover cooperates (as prescribed) if $a + a \geq b - \beta_2^H(b-c) + a$, which holds since $\beta_2^H \geq \frac{b-a}{b-c}$ by definition. On the other hand, the unconcerned second mover defects since $a + b - \beta_2^L(b-c) \leq b - \beta_2^L(b-c) + b - \beta_2^L(b-c)$, which holds since $\beta_2^L < \frac{b-a}{b-c}$ by definition. Therefore, this pooling strategy profile cannot be supported as a PBE if $q, \gamma \geq q^{Seq}(\alpha_1)$.

2. If $q, \gamma < q^{Seq}(\alpha_1)$ the first mover defects both in the first- and second-period stage game, both after observing that the second mover selects C and D in the first-period stage game. The concerned second mover defects, however, since $c - \alpha_2^H(b-c) + d \leq d + d$, which implies $\alpha_2^H \geq 0 \geq \frac{c-d}{b-c}$, which holds by definition. Hence, this pooling strategy profile cannot be sustained as a PBE if $q, \gamma < q^{Seq}(\alpha_1)$.

3. If $q \geq q^{Seq}(\alpha_1) > \gamma$ the first mover cooperates in the first-period stage game, while in the second-period stage game he cooperates only after observing that the second mover selected C (in equilibrium) in the first-period stage game. On one hand, the concerned second mover cooperates since $a + a \geq b - \beta_2^H(b-c) + d$, or $\beta_2^H \geq \frac{b+d-2a}{b-c}$. Given that $\frac{b-a}{b-c} > \frac{b+d-2a}{b-c}$, then condition $\beta_2^H \geq \frac{b+d-2a}{b-c}$ is satisfied from $\beta_2^H \geq \frac{b-a}{b-c}$. On the other hand, the unconcerned second mover cooperates since $a + b - \beta_2^L(b-c) \geq b - \beta_2^L(b-c) + d$, or $a \geq d$. Hence this pooling strategy profile can be supported as a PBE for $q \geq q^{Seq}(\alpha_1) > \gamma$.

4. If $q < q^{Seq}(\alpha_1) \leq \gamma$ the first mover defects in the first-period stage game, while in the second-period stage game he defects only after observing that the second mover chose C (in equilibrium) in the first-period stage game. The concerned second mover defects, however, since $c - \alpha_2^H(b-c) + d \leq d + a$, which implies $\alpha_2^H \geq 0 \geq \frac{c-a}{b-c}$, which is satisfied by definition. Thus, this pooling strategy profile cannot be sustained as a PBE if $q < q^{Seq}(\alpha_1) \leq \gamma$. ■

6.7 *Proof of Corollary 2*

Second mover. When he is highly concerned about fairness, i.e., when $\beta_2 \geq \frac{b-a}{b-c}$, then players' preferences satisfy $\beta_1, \beta_2 \geq \frac{b-a}{b-c}$, implying that outcome (C,C) can be sustained in pure strategies under complete information. (Thus, if this sequential-move game is repeated twice, (C,C) would arise during both periods when both players' guilt aversion is sufficiently high.) In this context, the second mover's equilibrium utility is $a + a$ under complete information. When the second mover is privately informed about his own high concern for fairness, $\beta_2 \geq \frac{b-a}{b-c}$, then outcome (C,C) is similarly played during both stages of the repeated game (as described in Proposition 2), thus yielding the same equilibrium utility for the (informed) first mover. When, instead, the first mover is informed that his concern for fairness is low, i.e., $\beta_2 < \frac{b-a}{b-c}$, players' preferences satisfy $\beta_1 \geq \frac{b-a}{b-c} > \beta_2$, implying that the outcome (D,D) can be supported under complete information, as described in Lemma 1. Hence, if this sequential-move game is repeated twice, (D,D) would arise during both periods, entailing that an equilibrium utility of $d + d$ for the second mover. When the second mover is privately informed about his low concern for fairness, $\beta_2 < \frac{b-a}{b-c}$, then Proposition 2 shows that in the first-period sequential-move game both players cooperate, while in the second-period game the first mover cooperates and the second mover responds by defecting. Therefore, the second-mover's equilibrium utility under incomplete information, $a+b-\beta_2(b-c)$, is larger than that under complete information, $d+d$, if $\beta_2 \leq \frac{a+b-2d}{b-c}$. Note that this condition holds for all $\beta_2 < \frac{b-a}{b-c}$, given that $\frac{b-a}{b-c} < \frac{a+b-2d}{b-c}$, which is satisfied since $a > d$ by assumption. Therefore, the second-mover's equilibrium utility is weakly larger under incomplete than under complete information for

all parameter values.

First mover. Let us investigate whether the uninformed first mover obtains a larger expected utility in the complete or incomplete information game. First, under complete information, the first mover obtains an expected utility of $q[a+a]+(1-q)[d+d]$. The first component reflects the fact that the first mover is informed about the second mover's concerns being high and, since his own concerns are also high, outcome (C,C) can be supported in both periods; as shown in Lemma 1. However, the second component represents the case in which the first mover is informed that the second mover is unconcerned. In this case, the first mover can anticipate that any of his actions will be responded to with defection and, as shown in Lemma 1, the only outcome that can be sustained in equilibrium is (D,D) during both periods of interaction. In contrast, when players compete in the incomplete information game, the expected utility of the uninformed first mover is $q[a+a]+(1-q)[a+c-\alpha_1(b-c)]$. Indeed, the first component corresponds to the case where the second mover has high fairness concerns, which implies the cooperative outcome (C,C) obtains in both time periods, while the second component corresponds to the case where the second mover's fairness concerns are low. In that setting, as described in the backstabbing equilibrium identified of Proposition 2, both players cooperate in their first period of interaction, but in the second period the (uninformed) first mover cooperates while the (informed) second mover responds by defecting. Comparing the first mover's expected utility in both settings, we obtain that he prefers to remain uninformed if

$$q[a+a]+(1-q)[a+c-\alpha_1(b-c)] > q[a+a]+(1-q)[d+d],$$

which reduces to $a+c-\alpha_1(b-c)-2d>0$, or $\alpha_1 \leq \frac{a+c-2d}{b-c}$. Note that this condition on α_1 is compatible with the initial assumption of $\alpha_1 \geq \beta_1 \geq \frac{b-a}{b-c}$ if $\frac{b-a}{b-c} < \frac{a+c-2d}{b-c}$, i.e., if $2a-b>2d-c$, which holds when the payoff from promoting the cooperative outcome (C,C) is sufficiently high. ■

6.8 *Proof of Corollary 3*

Separating PBE. Let us first analyze the separating strategy profile where the first mover cooperates when he has high concerns for fairness but defects otherwise. First, note that after observing an action from the first mover in the first-period sequential-move PD game, the second mover's beliefs about β_1^H are updated according to Bayes' rule, becoming $\mu(\beta_1^H|C)=1$ and $\mu(\beta_1^H|D)=0$. Recall that the second mover is concerned about fairness by definition, $\beta_2^H \geq \frac{b-a}{b-c}$, and that this information is common knowledge. Hence, given the above beliefs, the second mover's best response is to "mimic" the action selected by the first mover (as described in Lemma 1), both in the first and second-period sequential PD games. Regarding the first mover, when he has high concerns for fairness he cooperates since $a+a \geq d+d$. If unconcerned, the first mover defects (as prescribed) if $a+d<d+d$, which cannot hold given that $a>d$ by definition. Hence, the separating strategy profile cannot be supported as a PBE of the game.

Pooling PBE. Let us now analyze the pooling strategy profile where both types of first mover

cooperate in the first-period game. First, note that after observing an action from the first mover during the first-period sequential-move PD game, the second mover's beliefs about β_1^H become $\mu(\beta_1^H|C) = q$ (in equilibrium) and $\mu(\beta_1^H|D) \equiv \gamma \in [0,1]$ (off-the-equilibrium path). Because the second mover's best response is to "mimic" the action selected by the first mover, we do not need to analyze equilibrium played under different beliefs as we did in the proof of Proposition 2. Regarding the first mover, when he has high concerns for fairness he cooperates since $a+a \geq d+d$. If unconcerned, the first mover also cooperates (as prescribed) given that $a+a \geq d+d$. Hence, the pooling strategy profile can be supported as a PBE of the game, for all q and $\gamma \in [0,1]$. ■

References

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

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

[3] Banks, Jeffrey S. and Joel Sobel (1987) "Equilibrium selection in signaling games," Econometrica, 55, pp. 647-661.

[4] Bellemare, Charles, Sabine Kröger and Arthur van Soest (2008) "Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities," Econometrica, 76, pp. 815-839.

[5] Bolle, Friedel and Peter Ockenfels (1990) "Prisoner's dilemma as a game with incomplete information," Journal of Economic Psychology, 11, pp. 69-84.

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

[7] Brandts, Jordi and Neus Figueras (2003) "An exploration of reputation formation in experimental games," Journal of Economic Behavior and Organization, 50, pp. 89-115.

[8] Camerer, Colin and Keith Weigelt (1988) "Experimental tests of the sequential equilibrium reputation model," Econometrica, 56, pp. 1.36.

[9] Clark, Kenneth and Martin Sefton (2001) "The sequential prisoner's dilemma: Evidence on reciprocation," The Economic Journal, 111, pp. 51-68.

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

[11] 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.

[12] Demicheli, Stephano and Jörgen W. Weibull (2008) "Language, meaning, and games: A model of communication, coordination, and evolution," American Economic Review, 98(4), pp. 1292-1311.

[13] Duffy, John and Félix Muñoz-García (2012) "Patience or fairness? Analyzing social preferences in repeated games," Games, 3(1), pp. 56-77.

[14] Dufwenberg, Martin and Georg Kirchteiger (2004) "A theory of sequential reciprocity," Games and Economic Behavior, 47, pp. 268-298.

[15] Falk, Armin and Urs Fischbacher (2006) "A theory of reciprocity," Games and Economic Behavior, pp. 293-315.

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

[17] Fong, Yuk-Fai (2009) "Private information of nonpaternalistic altruism: Exaggeration and reciprocation of generosity," The B.E. Journal of Theoretical Economics, vol. 9, article 1.

[18] Healy, P.J. (2007) "Group reputations, stereotypes, and cooperation in a repeated labor market" American Economic Review, 97(5), pp. 1751-1773.

[19] Kreps, David, Paul Milgrom, John Roberts and Robert Wilson (1982) "Rational cooperation in the finitely repeated prisoners' dilemma," Journal of Economic Theory 27, pp. 245-52.

[20] Kreps, David and Robert Wilson (1982) "Reputation and imperfect information," Journal of Economic Theory 27, 253-279.

[21] McKelvey, Richard D. and Thomas R. Palfrey (1992) "An experimental study of the centipede game," Econometrica, 60, pp. 803-836.

[22] Rabin, Matthew (1993) "Incorporating fairness into game theory and economics," American Economic Review, 83(5), pp. 1281-1302.

[23] 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.

[24] von Siemens, Ferdinand (2009) "Bargaining under incomplete information, fairness, and the hold-up problem," Journal of Economic Behavior and Organization, 71, pp. 486-94.