Signaling Concerns about Fairness: Cooperation under Uncertain Social Preferences

By

John Duffy and Felix Munoz-Garcia

December 2010
Abstract

This paper investigates incomplete information and signaling about players’ inequity aversion in the simultaneous and sequential-move prisoner’s dilemma game. We first evaluate the role of incomplete information according to: (1) whether uncertainty helps select the efficient equilibrium outcome, and (2) whether more cooperation can be sustained under incomplete than under complete information. We then examine the possibility of information transmission among individuals in a signaling game. A separating equilibrium can be supported in which players with high concerns about fairness bear the cost of cooperating in order to reveal their type to opponents, thus promoting cooperation in subsequent periods. We also find a pooling equilibrium in which a player unconcerned about inequity aversion initially cooperates in order to mislead the uninformed player. This misleading strategy induces cooperation from the uninformed player in the subsequent stage of the game, moment at which the unconcerned player takes the opportunity to defect. This “backstabbing” equilibrium helps explain frequently observed behavior in finitely-repeated experiments.

Keywords: Prisoner’s Dilemma; Inequity aversion; Incomplete Information; Signaling.
JEL classification: C72, C73, D82.
1 Introduction

A large body of experimental evidence suggests that many individuals appear to exhibit concerns for fairness in the income distribution, also referred as “social” preferences. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) 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. Most theoretical analysis of the role of social preferences has been developed in complete information settings, wherein players are able to perfectly observe the extent of one another’s concerns for fairness, i.e., the parameters one another’s social preference functions. This assumption may be reasonable in certain contexts, for instance, when players have interacted with one another for several previous periods. However, complete information regarding the extent to which players have social preferences seems less sensible if players are unfamiliar with the strategic environment or have had no interactions with their opponents, a situation that characterizes the initial round(s) of play of many experimental games.

This paper contributes to the literature on the role of social preferences in fostering cooperative behavior by studying the case where players have incomplete information regarding the extent of other player’s social preferences, specifically, incomplete information about other player’s concerns for fairness in the distribution of payoffs. We study this issue using simultaneous- and sequential-move versions of the canonical Prisoner’s Dilemma (PD) game played either once or a finite number of times.¹

As our benchmark, we first investigate the one-shot, simultaneous-move PD game under complete information. We show that when players are unconcerned about fairness, defection remains a strictly dominant strategy, as in the standard prisoner's dilemma. In contrast, when players are concerned about fairness, players' best response function is to “mimic” the other player’s action: cooperate when the other player cooperates and defect otherwise. Therefore, under complete information about social preferences, the prisoner's dilemma becomes strategically equivalent to a Pareto coordination game.²

We then introduce incomplete information about players social preferences into the one-shot simultaneous move PD game evaluating the effects of such uncertainty on equilibrium play according to two criteria: (1) whether cooperation can be sustained more generally (i.e., for a wider range of social preference parameter values) when players face uncertainty about each others’ social preferences than when they do not; and (2) if the introduction of uncertainty can be used as a means of selecting the efficient equilibrium outcome among the multiple equilibria that are supported under

---

¹PD games are strategically equivalent to public good games as well as to games of imperfect competition among firms where actions are strategic substitutes. Sequential move versions are also used to characterized firm-worker wage-effort decisions and the notion of “gift-exchange”.

²This result relates with that of Rabin (1993) for psychological games. Indeed, he predicts that players behave according to the above best response function if they are sufficiently 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.
complete information. In the case of the simultaneous-move game we show that uncertainty allows for partial cooperation to occur and that mutual cooperation is the unique equilibrium outcome that can be sustained when both players have sufficiently high concerns about fairness; this finding stands in contrast to the complete information game where multiple equilibria exist under the same conditions.

We next analyze the sequential version of the PD game. Under complete information, the second mover “mimics” the first mover’s action when he is sufficiently concerned about fairness, but defects otherwise. Given this best response, the first mover cooperates when the second mover is sufficiently concerned about fairness but defects otherwise, supporting, as a consequence, a unique equilibrium outcome. Therefore, the introduction of uncertainty in the sequential-move PD game does not render equilibrium-selection benefits. Nonetheless, we show that incomplete information in this case does allow for partial cooperation to occur for social preference parameter values that would allow for only mutual defection under complete information.

We further investigate the role of uncertainty regarding social preferences in twice-repeated versions of the simultaneous and sequential-move PD games. In this context, we analyze a signaling game in which players use their actions to communicate their social preferences to other individuals. In particular, we suppose that one player privately observes his own concern for fairness, while the social preferences of the other player are common knowledge. Players interact in a simultaneous prisoner’s dilemma game in the first period, at the end of which payoffs are accrued, allowing every player to infer his opponent’s chosen action. In the second and final period, players interact in a simultaneous game again, and payoffs are distributed. We first identify a separating equilibrium in which only highly concerned players cooperate in the first-period game whereas unconcerned players defect. Specifically, we show that when the informed player is sufficiently concerned about fairness, he may choose to cooperate during the first period of the game in order to convey his social concerns to the uninformed player, inducing the latter to cooperate in the subsequent period. This result suggests that uninformed subjects might respond to kind actions (cooperation) with kind actions, even when they only sustain preferences for fairness, but they do not have intrinsic concerns for intentions or reciprocity in their preferences.

We also identify a pooling equilibrium in which both types of informed players cooperate during the first-period game of the simultaneous PD game. An informed player who is unconcerned with fairness chooses to cooperate, pooling with the uniformed player who is highly concerned with

---

3Our results under complete information relate with those in Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) which, in the context of sequential-move psychological games, show that the second mover behaves according to the above best response if he is sufficiently motivated by reciprocity. Similarly to Rabin (1993), however, these papers assume that individuals’ reciprocity parameters are common knowledge among the players. In contrast, we extend our analysis to the case in which every player observes his own concern for fairness but not his opponents’.

4In this regard, our results relate to those in Fong (2009) where, in the context of the gift-exchange game under incomplete information about altruism, he shows that the donor might exaggerate his gifts to the recipient in order to reveal his strong altruistic concerns. Such “gift exaggeration” triggers reciprocal behavior afterwards even when agents have no concerns for reciprocity. Similarly, Cason and Mui (2009) provide experimental data showing that, for the divide-and-conquer game, players can convey their social preferences to other individuals through their actions in the repeated game (or through actual communication), helping to ameliorate the first-mover’s transgression against two responders.
fairness in order to “mislead” the uninformed player about his actual concerns for fairness. If priors are sufficiently high, this misleading strategy attracts the uninformed player towards cooperation in the subsequent second period at which point the unconcerned player 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, whereby subjects defect in the last period of their interaction, despite a previous history of cooperation; see for instance, Selten and Stoecker (1986) and Andreoni and Miller (1993) for the prisoner’s dilemma, McKelvey and Palfrey (1992) for the centipede game and Camerer and Weigelt (1988) and Brandts and Figueras (2003) for a borrower-lender game. Importantly, this informational explanation, does not rely on subjects’ inability to understand the rules of the game, or failure to backward induct but rests instead on the existence of incomplete information about other players’ social preferences.

Healy (2007) identifies a similar result in the context of a finitely-repeated gift-exchange game where the firm manager does not observe the worker’s type (assumed to be either reciprocator or selfish–Healy’s results do not derive from any preference specifications). 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. In particular, we examine a twice-repeated sequential-move prisoner’s dilemma game where the first mover is uninformed about the second-mover’s social preferences.\footnote{Indeed, note that 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.} We show that a pooling equilibrium not only emerges in the simultaneous but also in the sequential version of the game. Specifically, the privately informed player (second mover) uses a “backstabbing” strategy by disguising himself as an individual with high concerns about fairness during the first-period game (cooperating) in order to defect in the second-period 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. Specifically, our results predict that more (less) cooperation can be sustained in the simultaneous- than in the sequential-move version of the PD game if the first-mover’s concerns for fairness are relatively high (low, respectively). Importantly, we show that these results are not only applicable to the case in which only one player is uninformed about his opponent’s social preference parameters, but also to settings where all players are uninformed about each others’ social preferences.

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 the sequential-move version of the game. By contrast, our model analyzes equilibrium play under more general time and information structures: the simultaneous and sequential version of the game, both under complete and incomplete information, which allows for a richer set of comparisons. Additionally, we allow for signaling to occur in the finitely repeated game which, as suggested above, can help
to rationalize several experimental observations.

Finally, note that the behavioral assumptions in this paper are related to those in Kreps et al. (1982). In particular, they show that cooperation can be sustained in the finitely-repeated prisoner’s dilemma game as long as both players believe that there is a small probability that his opponent is “irrational,” i.e., plays a conditionally cooperative tit-for-tat strategy. In particular, Kreps et al. (1982) consider 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. In contrast, they show that their conclusions would not hold in a context of one-sided uncertainty in which it is common knowledge that defection constitutes a dominant strategy for the uninformed player. In this paper we show that Kreps et al.’s (1982) cooperative results can be extended to environments with one-sided uncertainty as well. Specifically, this occurs when it is common knowledge that the uninformed player’s best-response is to “mimic” his opponent’s actions. In that setting, the “backstabbing” equilibrium predicts cooperation during the first period of play by players unconcerned about fairness who attempt to convince their uninformed opponents that they are concerned about fairness when in fact they are not. In this 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 that information about the informed player’s type can be conveyed to the uninformed player via signaling; see Kreps et al. (1982) page 251.


The structure of the paper is as follows. The following section presents the model. Section three compares equilibrium outcomes in the simultaneous-move game under complete and incomplete information. Section four makes a similar comparison for the sequential-move version of the game. In section five we investigate the extent of information transmission about privately observed social preferences between players in twice-repeated versions of the simultaneous and sequential-move PD games. Section six concludes.

## 2 Model

Consider the 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 game.

<table>
<thead>
<tr>
<th></th>
<th align="right">Player 1</th>
<th align="right"></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td align="right">C</td>
<td align="right">D</td>
</tr>
<tr>
<td><strong>Player 2</strong></td>
<td align="right">a,a</td>
<td align="right">c,b</td>
</tr>
<tr>
<td><strong>Player 1</strong></td>
<td align="right">b,c</td>
<td align="right">d,d</td>
</tr>
</tbody>
</table>

We shall focus on the game played by players who possess Fehr and Schmidt (1999)-type social preferences, a now standard specification which, for the case of two players, reduces to:

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

---

6 In this 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 that information about the informed player’s type can be conveyed to the uninformed player via signaling; see Kreps et al. (1982) page 251.

where $x_i$ is player $i$’s payoff, and $x_j$ is the other player $j$’s payoff. Parameter $\alpha_i$ represents the disutility from allocations that are disadvantageously unequal for player $i$ (i.e., due to envy about player $j$’s higher payoff), while $\beta_i$ 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 concerns dominate guilt concerns, i.e., $\alpha_i \geq \beta_i$ and $1 > \beta_i \geq 0$.\(^8\) We shall contrast this case of “social preferences” (alternatively described throughout this paper as “concerns for fairness” or “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:

\[
\begin{array}{c|cc}
\text{Player 1} & \text{C} & \text{D} \\
\hline
\text{C} & a, a & c - \alpha_1(b - c), b - \beta_2(b - c) \\
\text{D} & b - \beta_1(b - c), c - \alpha_2(b - c) & d, d \\
\end{array}
\]

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 $i$. In contrast, if player $i$’s concern for fairness is relatively high, $\beta_i \geq \frac{b-a}{b-c}$, his best response is to match player $j$’s action: cooperate when $j$ cooperates and defect otherwise.

### 3 Simultaneous-move game

In this section we briefly analyze equilibrium predictions for the simultaneous-move Prisoner’s Dilemma (PD) game under complete information about social preferences, in order to compare them afterwards with equilibria under incomplete information. As shown in Duffy and Munoz-Garcia (2010), if either (or both) players have relatively low concerns about guilt (equivalently, low concerns for fairness), 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. However, when both individuals are sufficiently concerned about fairness (the shaded area of Figure 1), we can identify three different Nash equilibria: one in which both players defect, one in which both players cooperate, and a mixed strategy Nash equilibrium where players randomize. The introduction of sufficient concerns about fairness for both players then transforms the payoff structure of the game from a Prisoner’s Dilemma to a Pareto-rankable coordination game. In particular, every player’s best response is to select the same action as his

\(^8\)Intuitively, $\alpha_i \geq \beta_i$ implies that players (weakly) suffer more from inequality directed at them than inequality directed at others. On the other hand, $\beta_i \geq 0$ means that players dislike being better off than others (this assumption rules out cases in which individuals are status seekers but serves to simplify the analysis). Finally, $\beta_i < 1$ suggests that when player $i$’s payoff is higher than that of player $j$’s by one unit (e.g. a dollar), player $i$ is never willing to give up more than one unit in order to reduce this inequality. For a more detailed explanation of these assumptions, see Fehr and Schmidt (1999).
opponent, but both players strictly prefer (C,C) to (D,D).\footnote{Note that this best response function is similar to what Cooper et al. (1996) call “best response altruists,” namely players for whom cooperate is their best response to cooperation, but defect is their best response to defection.}

![Figure 1. Equilibria in the simultaneous game under complete information.](image)

### 3.1 Simultaneous-move game under incomplete information

This section examines how the introduction of incomplete information regarding players’ social preferences affects equilibrium play in the simultaneous-move version of the game. In particular, we consider the case where every player $i$ knows his own preference parameters, $\alpha_i, \beta_i$, but is uncertain of those for the other player $j \neq i$. We further suppose that the envy parameter $\alpha_i$ is distributed according to a commonly known cumulative distribution function $F(\alpha_i)$ with associated density $f(\alpha_i) > 0$ for all $\alpha_i \geq 0$ for both players $i = \{1, 2\}$. Similarly, the guilt parameter $\beta_i$ is assumed to be distributed according to the commonly known cumulative distribution function $G(\beta_i)$ with associated density $g(\beta_i) > 0$ for all $\beta_i \geq 0$ for both individuals. For simplicity, we assume that probability distributions over preferences are independent across players.\footnote{The distribution of fairness concerns could be non-independent among individuals if, for instance, player $i$’s observation of his own concern about fairness informs him about the conditional probability that his opponent’s social preferences take values in a certain interval. In section 6 we discuss the implications of this possibility.}

As observed in the simultaneous move PD game under complete information, when a player’s own concern about guilt is sufficiently low, $\beta_i < \frac{b-a}{b-c}$, defection becomes a strictly dominant strategy for him, for any concerns (preference parameters) of his opponent, $(\alpha_j, \beta_j)$. This same argument is also applicable to the incomplete information game. By contrast, when a player’s own concerns for fairness are sufficiently high, $\beta_i \geq \frac{b-a}{b-c}$, his decision to cooperate depends on the expectation he forms about his opponent’s social preferences. In particular, player $i$ cooperates when he is highly
concerned about fairness, \( \beta_i > \frac{b-a}{b-c} \), if and only if \( EU_i \left( C|\beta_i > \frac{b-a}{b-c} \right) \geq EU_i \left( D|\beta_i > \frac{b-a}{b-c} \right) \), or
\[
\left[ 1 - G \left( \frac{b-a}{b-c} \right) \right] a + G \left( \frac{b-a}{b-c} \right) [c - \alpha_i(b - c)] \geq \left[ 1 - G \left( \frac{b-a}{b-c} \right) \right] [b - \beta_i(b - c)] + G \left( \frac{b-a}{b-c} \right) d
\]

Intuitively, player \( i \) obtains a payoff of \( a \) when both players cooperate, \((C,C)\), which occurs if player \( j \) is highly concerned about fairness, \( \beta_j \geq \frac{b-a}{b-c} \), an event that occurs with probability \( 1 - G \left( \frac{b-a}{b-c} \right) \). By contrast, player \( i \) experiences a disutility if he cooperates but his opponent defects, \((C,D)\), which happens when player \( j \)’s concerns for fairness are relatively low, or with probability \( G \left( \frac{b-a}{b-c} \right) \). Specifically, his payoff is only \( c - \alpha_i(b - c) \), because of the envy he experiences in outcome \((C,D)\). If, instead, player \( i \) chooses to defect, then he might be the only player defecting – outcome \((D,C)\) – which occurs when player \( j \) is highly concerned about fairness; in this \((D,C)\) outcome, player \( i \)’s payoff is \( b - \beta_i(b - c) \), due to the the guilt that player \( i \) experiences. Finally, if both players defect, \((D,D)\), player \( i \)’s payoff is \( d \). Solving for \( G \left( \frac{b-a}{b-c} \right) \), we have:

\[
G \left( \frac{b-a}{b-c} \right) \leq \frac{a - b + \beta_i(b - c)}{a + d - b - c + (\alpha_i + \beta_i)(b - c)} \equiv C^{Sm} (\alpha_i, \beta_i) \tag{Condition A}
\]

where \( C^{Sm} \) denotes the cut-off for the simultaneous move PD game. Condition A is satisfied for probability distributions \( G(\beta_j) \) that concentrate most of their weight on realizations above \( \beta_j = \frac{b-a}{b-c} \). That is, \( G \left( \frac{b-a}{b-c} \right) \leq C^{Sm} (\alpha_i, \beta_i) \), or alternatively, \( 1 - G \left( \frac{b-a}{b-c} \right) \geq 1 - C^{Sm} (\alpha_i, \beta_i) \). Note further that \( C^{Sm} (\alpha_i, \beta_i) \) is decreasing in \( \alpha_i \) and increasing in \( \beta_i \). Consequently, a low concern about envy but a high concern about guilt raises the cutoff \( C^{Sm} (\alpha_i, \beta_i) \), thus making cooperation more likely. We can now characterize the Bayesian equilibrium of this simultaneous-move game with incomplete information.

**Proposition 1.** In the simultaneous PD game where players are privately informed about their concerns for fairness, the following strategy profile can be supported as pure strategy Bayesian Nash equilibria of the game:

1. \((C,C)\) if both players’ concerns about fairness are sufficiently high, \( \beta_i, \beta_j \geq \frac{b-a}{b-c} \), and the probability distribution \( G(\cdot) \) satisfies condition A;

---

11 Note that if \( \beta_j \geq \frac{b-a}{b-c} \) then \( \alpha_j \) satisfies \( \alpha_j > \frac{b-a}{b-c} \) since \( \alpha_j \geq \beta_j \) by definition. This is applicable in the first term of both the left and right hand side of the above inequality. In the second term of both sides of the inequality we use the property that defection becomes an strictly dominant strategy for player \( j \) when \( \beta_j < \frac{b-a}{b-c} \), for any realization of \( \alpha_j \).

12 If a particular distribution function \( G_A(\beta_j) \) satisfies Condition A, then any distribution function \( G_B(\beta_j) \neq G_A(\beta_j) \) that first-order stochastically dominates \( G_A(\beta_j) \), i.e., \( G_B(\beta_j) \leq G_A(\beta_j) \) for every \( \beta_j \), must also satisfy Condition A.

13 Note that \( C^{Sm} (\alpha_i, \beta_i) \) is strictly positive for \( \beta_i \geq \frac{b-a}{b-c} \), which is the case when examining player \( i \)’s decision to cooperate or defect. (Recall that when \( \beta_i < \frac{b-a}{b-c} \) defection becomes a strictly dominant strategy.) Additionally, \( C^{Sm} (\alpha_i, \beta_i) \) is lower than one for any \( \alpha_i > 0 > \frac{b-c}{c} \), which is true by definition.

14 Note that assuming different probability distributions over types for players \( i \) and \( j \) would make condition A player-dependent, specifying that player \( i \) cooperates if the probability distribution on player \( j \)’s guilt, \( G_j(\beta_j) \), satisfies \( G_j(\beta_j) \leq C^{Sm} (\alpha_i, \beta_i) \).
2. \((C,D)\) if player \(i\) is relatively concerned about fairness but player \(j\) is not, \(\beta_i \geq \frac{b-a}{b-c} > \beta_j\), and the probability distribution \(G(\cdot)\) satisfies condition A. Similarly for \((D,C)\).

3. \((D,D)\) otherwise.

If Condition A is violated, the unique equilibrium outcome in the incomplete-information simultaneous move PD game, \((D,D)\), coincides with that in the complete information version of that game where one or both players are unconcerned about fairness. If Condition A is satisfied, however, outcomes other than \((D,D)\) can also be sustained as equilibria. Figure 2 illustrates the equilibrium possibilities in the simultaneous-move game under incomplete information when Condition A is satisfied. First, when both players are sufficiently concerned about fairness, the introduction of incomplete information in the stage game reduces the set of pure-strategy equilibria: from \((C,C)\) and \((D,D)\) under complete information in Figure 1 to only \((C,C)\) in Figure 2. Hence, incomplete information about players’ social preferences can aid in selecting the efficient, cooperative equilibrium \((C,C)\).

![Figure 2. Equilibria in the simultaneous game under incomplete information.](image)

When only one of the two players is highly concerned about fairness, then strategy profiles \((C,D)\) or \((D,C)\) can be supported as equilibria as indicated by the dashed areas in Figure 2. Essentially, under complete information, the highly-concerned player knows that he is dealing with an individual with low concerns for fairness for whom defection is a strictly dominant strategy. This leads him to also defect — inducing outcome \((D,D)\) — in order to avoid the disutility arising from envy in the \((C,D)\) outcome. In contrast, under incomplete information, the highly concerned player chooses to cooperate only according to the prior probability that his opponent is also concerned. Hence, when players’ concerns for fairness are relatively asymmetric, the introduction of incomplete information can facilitate cooperation from the most concerned player.
4 Sequential-move game

In this section we investigate cooperation in the one-shot, sequential-move version of the PD game. Under complete information, the second mover 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 when his own concern for fairness is low. Given the second mover’s response, the first mover chooses to cooperate when he anticipates that the second mover will respond by cooperating, which occurs only if $\beta_j \geq \frac{b-a}{b-c}$. We summarize the equilibrium of this sequential-move game under complete information in the following lemma.

Lemma 1. In the sequential-move PD game where players are informed about each others’ social preference parameters, the following strategy profile can be supported as the unique subgame perfect equilibrium of the game:

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 3. 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 the simultaneous version of the game (compare the shaded areas in Figures 3 and 1, respectively). In particular, cooperation can be sustained in the sequential-move version as long as the second mover is sufficiently concerned about fairness, $\beta_j \geq \frac{b-a}{b-c}$, as Figure 3 illustrates below, unlike in the simultaneous game, where such an outcome could only be sustained if both players’ fairness concerns are sufficiently high, i.e., $\beta_i, \beta_j \geq \frac{b-a}{b-c}$.

15 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 “matching” the first mover’s choice if he is sufficiently reciprocally motivated. When he is not, he responds to any action of the first mover with defection.

16 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 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, relatively to the standard PD game, 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 goes in line with our result, since the second mover has more incentives to respond to cooperation with cooperation if his concerns for fairness are relatively high (when the “temptation payoff” from defecting decreases), but to respond defecting any choice of the first mover when he is unconcerned about fairness (when the “temptation payoff” increases).
4.1 Sequential-move game under incomplete information

We now introduce incomplete information into the sequential-move PD game by considering the case in which both players are privately informed about their own concerns for fairness and cannot observe each other’s concerns. This situation arises in instances where players interact for the first time in a strategic setting that is novel to both of them. Notice that the second mover’s best response still coincides with that described in Lemma 1, since it depends on his own social preferences, and is independent of the first mover’s social preferences. The first mover, however, must now decide whether to cooperate or defect without knowing the second mover’s type, choosing to cooperate if

$$1 - G \left( \frac{b-a}{b-c} \right) a + G \left( \frac{b-a}{b-c} \right) (a - \alpha_i(b - c)) \geq 1 - G \left( \frac{b-a}{b-c} \right) d + G \left( \frac{b-a}{b-c} \right) d$$

By cooperating, the first mover – say, player $i$ – obtains a payoff of $a$ when the second mover responds with cooperation, but his payoff is reduced to $c - \alpha_i(b - c)$ when the second mover, player $j$, responds with defection. If, instead, the first mover chooses to defect, the second mover responds by defecting regardless of his concerns for fairness, yielding a payoff of $d$. Solving for $G \left( \frac{b-a}{b-c} \right)$ in the above inequality, we have: \(^{17}\)

$$G \left( \frac{b-a}{b-c} \right) \leq \frac{a - d}{(a - c) + \alpha_i(b - c)} \equiv C_{Seq} (\alpha_i) \quad \text{(Condition B)}$$

Similar to the simultaneous-move version of the PD game, when Condition B is satisfied, the first mover chooses to cooperate in the sequential version of the game as the probability distribution

\(^{17}\)Note that cutoff $C_{Seq} (\alpha_i)$ is lower than one for any $\alpha_i > 0 > \frac{c - d}{b - c}$, which holds by definition.
$G(\beta_j)$ concentrates most of its weight on relatively high values for $\beta_j$, implying that the second mover is most likely a reciprocating type. Cutoff $C^{\text{Seq}}(\alpha_i)$ depends upon the first mover’s envy parameter, $\alpha_i$, but is independent of his guilt parameter, $\beta_j$. Indeed, when the first mover chooses to cooperate, he faces the risk that the second-mover is unconcerned about fairness and defects, yielding a disutility to the first mover in the form of envy. If, instead, the first mover $i$ decides to defect, then the second mover $j$ responds by defecting with certainty, regardless of $j$’s own fairness concerns. This reciprocating response by the second mover eliminates the disutility from guilt that the first mover would experience under $(D,C)$, thus making $C^{\text{Seq}}(\alpha_i)$ independent of the first mover’s guilt parameter $\beta_i$. Finally, note that the cutoff $C^{\text{Seq}}(\alpha_i)$ is decreasing in $\alpha_i$, making it less likely that Condition $B$ is satisfied. Intuitively, an increase in the first mover’s envy makes cooperation more difficult to sustain in the sequential-move game, which parallels our result for the simultaneous version of the game. The next proposition describes equilibrium play in the sequential game under incomplete information.

**Proposition 2.** In the sequential-move PD game where both players are privately informed about their own concerns for fairness, the following strategy profile can be supported as the pure strategy PBE of the game:

1. The first mover cooperates if and only if Condition $B$ holds; 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.

Similar to Condition $A$ for the simultaneous-move game, both players in the sequential-move game defect under all parameter values when Condition $B$ is violated. When Condition $B$ is satisfied, however, the first mover cooperates under all parameter values, whereas the second mover cooperates only when his own concern about fairness are sufficiently high. Figure 4 illustrates the equilibrium possibilities in the sequential-move game with incomplete information when Condition $B$ is satisfied.
Comparison with the game under complete information. Let us compare the last results, as illustrated in Figure 4, with those for the sequential-move game under complete information, as illustrated in Figure 3. When condition $B$ holds, the cooperative outcome $(C,C)$ can be supported under the same parameters as in the complete information environment. Nonetheless, the partially cooperative outcome $(C,D)$ can be sustained under incomplete information provided that $\beta_j < \frac{b-a}{b-c}$. The latter outcome cannot be supported in the complete information version of the game.

Comparison with the simultaneous-move game. Suppose first that Conditions $A$ and $B$ are simultaneously satisfied. In that case, the cooperative outcome $(C,C)$ can be supported under a larger set of social preference parameters in the sequential-move version than in the simultaneous-move version of the incomplete information PD game. To see this, compare the shaded areas of Figures 4 and 2, respectively. Similarly, the partially cooperative outcome, $(C,D)$, can be sustained in the sequential-move game when both players’ concerns for fairness are relatively low, $\beta_i, \beta_j < \frac{b-a}{b-c}$, unlike in the simultaneous-move game, where only the non-cooperative outcome $(D,D)$ can be sustained. Suppose next that only Condition $A$ ($B$) holds. In that case, cooperation can only be sustained in the simultaneous-move game (only in the sequential-move game, respectively), if both players are (the second mover is) highly concerned about fairness. Comparing Conditions $A$ and $B$, one obtains that $C_{Sm}(\alpha_i, \beta_i) \leq C_{Seq}(\alpha_i)$ if and only if $\beta_i < \frac{b-d}{b-c}$, where $\frac{b-a}{b-c} < \frac{b-d}{b-c}$. Therefore, for relatively low fairness concerns, $\beta_i < \frac{b-d}{b-c}$, $C_{Sm}(\alpha_i, \beta_i) \leq C_{Seq}(\alpha_i)$, making cooperation more likely to occur in the sequential-move than in the simultaneous-move game, i.e.,

---

18 Note that Conditions $A$ and $B$ are simultaneously satisfied when the probability of player $j$’s guilt being high, i.e., $\beta_j \geq \frac{b-a}{b-c}$, is sufficiently large. Intuitively, this implies that player $j$ is most likely a reciprocator with whom player $i$ plays the $(C,C)$ outcome, both in the simultaneous and sequential-move game.

19 This implies that relatively low concerns for fairness, $\beta_i < \frac{b-d}{b-c}$, are consistent with the previously used cutoff $\beta_i < \frac{b-a}{b-c}$, given that $\frac{b-a}{b-c} < \frac{b-d}{b-c}$. 

---
condition \( A, G\left( \frac{b-a}{b-c} \right) \leq C^{Sm} (\alpha_i, \beta_i) \), holds under a larger set of parameter values than does condition \( B, G\left( \frac{b-a}{b-c} \right) \leq C^{Seq} (\alpha_i) \). By contrast, for sufficiently high concerns for fairness, \( \beta_i \geq \frac{b-d}{b-c} \), we obtain \( C^{Sm} (\alpha_i, \beta_i) \geq C^{Seq} (\alpha_i) \), and cooperation becomes more likely in the simultaneous-move version of the incomplete information PD game. Intuitively, high concerns about guilt raise cutoff \( C^{Sm} (\alpha_i, \beta_i) \) —inducing player \( i \) to cooperate under larger parameter conditions— but do not affect the value of \( C^{Seq} (\alpha_i) \), as described above. Hence, when player \( i \)’s guilt concern is relatively high, i.e., \( \beta_i \geq \frac{b-d}{b-c} \), player \( i \)’s cooperation becomes more likely in the simultaneous than in the sequential version of the game. Otherwise, cooperation can be sustained under a larger set of parameter values in the sequential-move game.

5 Signaling private concerns about fairness

In this section we examine the conditions under which a player can use his actions to reveal his privately observed social preferences to other uninformed players with the aim of promoting cooperation in subsequent periods. We also investigate conditions under which an informed player with relatively low concerns for fairness might cooperate during the first period to “disguise” himself to his uninformed opponent as an individual highly concerned about fairness thereby inducing the latter to cooperate while he takes the opportunity to defect. Specifically, suppose that nature selects player \( i \)’s concern for fairness, \( \beta_i \), and this information is revealed to player \( i \) alone. For simplicity we assume that \( \beta_i \) is distributed according to a discrete probability distribution, specifically:

\[
\beta_i = \begin{cases} 
\beta_i^H & \text{with probability } q, \\
\beta_i^L & \text{with probability } 1 - q
\end{cases}
\]

for every player \( i = \{1, 2\} \), 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, \( \beta_j \), is perfectly observable by all players. In order to focus on the possibility that player \( i \) signals his guilt concern to player \( j \), let us further assume that both individuals’ envy concerns, \( \alpha_i \) and \( \alpha_j \), are common knowledge among the players. Hence, player \( i \) holds private information about his guilt parameter alone, since the precise value of \( \alpha_i \), either \( \alpha_i^H \geq \beta_i^H \) or \( \alpha_i^L \geq \beta_i^L \), is common knowledge.

5.1 Twice-repeated simultaneous-move PD game

The timing of the twice-repeated “signaling” game is as follows. First, before any interaction between the players, player \( i \) privately observes his social preferences, but the uninformed player \( j \) does not (he only knows the prior probability distribution of \( \beta_j \)). 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, uninformed player \( j \) can use this information to update his beliefs about player
i’s type. Given these beliefs, players play a second and final stage of the simultaneous move PD game and payoffs are accrued.

Suppose further that player j is sufficiently concerned about fairness, i.e., $\beta_j \geq \frac{b-a}{b-c}$. If this were not the case, then 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 have any effect on player j’s future play. Since the game only involves two subsequent periods, we assume no time discounting. Under these assumptions, the following proposition describes a separating Perfect Bayesian Equilibrium (PBE) of this signaling game where the player i cooperates (defects) in the first-period stage game when his fairness concerns are high (low, respectively).

**Proposition 3.** Suppose $q < \frac{d-c+\alpha_j(b-c)}{a+d-c-b+(\alpha_j+\beta_j)(b-c)} \equiv q^{\text{Sim}}(\alpha_j, \beta_j)$. Then there exists a separating PBE in which the informed player i cooperates in the first period when he is concerned about fairness but defects otherwise if and only if $\alpha^H_i < \frac{a+d-c-2d}{b-c}$ and $\alpha^L_i + \beta^H_i \geq \frac{c+b-2d}{b-c}$. In the second period, player i cooperates when he is concerned about fairness but defects otherwise. The uninformed but highly concerned player j defects in the first period and cooperates in the second period if and only if he observes that player i cooperated during the first period, given beliefs $\mu(\beta^H_i | C) = 1$ and $\mu(\beta^H_i | D) = 0$.

In this strategy profile, player i cooperates in the first period of the game only when his concerns about fairness are sufficiently high. Note that a player i with high concerns for fairness will cooperate in the first period despite knowing that the player j, (who is commonly known to be highly concerned about fairness) will defect in the first period due to j’s low prior, $q < q^{\text{Sim}}(\alpha_j, \beta_j)$, yielding an equilibrium outcome of (C,D), where player i’s payoff is reduced to $c - \alpha^H_i (b - c)$ due to envy. Player i’s cooperation during the first stage communicates to the uninformed player j that i’s concern for fairness is high (like player j’s) and consequently, both players will cooperate in the second and final period.\(^{21}\) Hence, the highly concerned player i incurs a first-period envy cost in order to reveal his type to player j. In return, he obtains a future benefit since he induces the cooperative outcome (C,C) in the second and final period of the game.\(^{22}\) Two conditions must hold for this separating equilibrium to be supported. First, a player i who is unconcerned with fairness cannot have incentives to deviate towards cooperation in the first period in order to mimic a highly-concerned player. In particular, the unconcerned player i does not deviate to cooperation if the disutility from the envy he would bear during the first period (when the uninformed player j defects), and the disutility from guilt he suffers in the second period (when the uninformed player j

\(^{20}\)Note that we use “separating” (“pooling”) equilibrium to refer to those strategy profiles in which the highly-concerned player i (both types of player i, respectively) cooperate during the first-period game.

\(^{21}\)When both players are highly concerned about fairness and player i’s high concerns are revealed to player j through equilibrium play in the separating equilibrium, both (C,C) and (D,D) can be supported as equilibrium outcomes in the second-period game, as described in Lemma 1. For simplicity, we assume that players can resort to some coordination mechanism, such as social norms or a stochastic randomization, by which players coordinate in the efficient cooperative outcome (C,C).

\(^{22}\)Note that if the informed player i discounts future payoffs, cooperative outcome (C,C) becomes less attractive and the separating equilibrium can be sustained under more restrictive parameter conditions. A similar argument is applicable to the pooling equilibrium we describe below.
cooperates) is sufficiently high, i.e., if $\alpha_i^L + \beta_i^L \geq \frac{(b-a)c - 2d}{b-c}$. On the other hand, the concerned player
$i$ cooperates in the first period, bearing a disutility from envy in order to convey his type to the
uninformed but highly concerned player $j$ thereby convincing the player $j$ to play the cooperative
outcome in the second period of the game. The concerned player $i$ finds this strategy profitable
only if the disutility from the envy he experiences in the first period (when his opponent defects)
is sufficiently small, i.e., if $\alpha_i^H < \frac{a+c-2d}{b-c}$. Note that this condition is compatible with the initial
assumption that $\alpha_i^H \geq \beta_i^H > \frac{b-a}{b-c}$ only if $2a-b > 2d-c$. The latter inequality is satisfied when the
benefit from promoting the cooperative outcome $(C,C)$ in the second-period game and obtaining a
payoff of $a$, is sufficiently high. Indeed, when that benefit is sufficiently high, player $i$ cooperates
during the first period, as prescribed in the separating equilibrium. In contrast, when that benefit
is relatively small, player $i$ prefers to avoid the first-period cost from envy and chooses to defect.
In the latter case, the separating equilibrium cannot be supported, and a pooling equilibria where
both types of player $i$ defect becomes the only PBE in pure strategies that can be sustained in the
signaling game.

The following proposition describes a different pooling equilibrium in which both types of player
$i$ cooperate in the first-period stage game.\footnote{For robustness, we show that both separating and pooling equilibria survive the Cho and Kreps’ (1987) Intuitive Criterion under all parameter values; see Appendix 1. That appendix also provides conditions under which the “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 the Cho and Kreps’ (1987) Intuitive Criterion.}

**Proposition 4.** Suppose $q > q^{Sim}(\alpha_j, \beta_j)$. Then there exists a pooling PBE in which player $i$
cooperates in the first and second period when he is concerned about fairness, but cooperates only
in the first period when he is unconcerned about fairness. 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, given beliefs $\mu(\beta_i^H|C) = q \geq q^{Sim}(\alpha_j, \beta_j) > \mu(\beta_i^H|D)$.

In the cooperative pooling equilibrium both types of player $i$ cooperate in the first period.
The uninformed player $j$ cooperates in both the first period (given his relatively high prior, $q$)
and the second period, as long as he observes that player $i$ cooperated in the first stage. Hence,
by cooperating in the first period, the highly concerned player $i$ induces the outcome $(C,C)$ in
the second period of the game, whereas the unconcerned player $i$ defects so that the final period
outcome is $(D,C)$. The latter “backstabbing” behavior of the unconcerned player $i$ is a commonly
observed outcome in the final period of experiments involving finitely-repeated, simultaneous-move
Prisoner’s Dilemma games as found in the literature cited in the introduction. In particular, subjects
initially cooperate but choose to defect in the last period, even when their opponent cooperated
has cooperated in all prior periods.\footnote{Appendix 2 confirms that both the separating and the pooling equilibria identified in propositions 3 and 4 survive Cho and Kreps’ (1987) Intuitive Criterion.}
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 is similar to that found in Healy (2007). Specifically, in the context of a finitely-repeated, sequential-move game (the gift-exchange game), Healy shows that both types of workers, “selfish” and “reciprocators,” respond to high wage offers exerting high levels of effort, continuing to do so until the last periods of interaction, when selfish workers slack.\footnote{In particular, this result corresponds to proposition 1 in Healy (2007) where the past actions of all players are observable, but the workers’ types are not.}

To facilitate a comparison of our results with those of Healy (2007), we next examine a twice-repeated game of incomplete information in which the stage game where players interact is the sequential prisoner’s dilemma, rather than the simultaneous-move version studied in the previous section. Here we suppose that the second mover is privately informed about his concerns for fairness, either high, $H_2 \equiv b > c$, or low, $L_2 < b > c$, with probabilities $q$ and $1 - q$, respectively. By contrast the first mover’s high concern for fairness, $H_1 \equiv b > c$, is assumed to be common knowledge among players. Under these assumptions, the following proposition provides conditions under which a cooperative pooling outcome can be a PBE.\footnote{This proposition compares the pooling equilibrium in the simultaneous and sequential versions of the signaling game. Nonetheless, the proof shows that the separating equilibrium described in section 5.1 for the simultaneous PD game can also be supported as a PBE of the sequential version of the game under similar parameter values.}

**Proposition 5.** Suppose $q > q^{\text{Seq}}_1 (\alpha_1) = \frac{d - c + \alpha_1 (b - c)}{a - c + \alpha_1 (b - c)}$. Then there exists a pooling PBE in the twice-repeated sequential-move PD game under incomplete information where:

1. The uninformed first mover cooperates in the first period but cooperates in the second period only after observing that the second mover cooperated during the first-period game; otherwise he defects, given beliefs $\mu (\beta_2^H | C) = q > q^{\text{Seq}}_1 (\alpha_1) > \mu (\beta_2^L | D)$.

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

Furthermore, $q^{\text{Seq}}_1 (\alpha_1) > q^{\text{Sim}}_1 (\alpha_1, \beta_1)$ if and only if the uninformed player is highly concerned about fairness, $\beta_1^H \geq \frac{b - d}{b - c}$.

Therefore, when players interact in the twice-repeated sequential-move PD game, the pooling strategy profile described in section 5.1 can be supported as a PBE by replacing cutoff $q^{\text{Sim}}_1 (\alpha_1, \beta_1)$ for $q^{\text{Seq}}_1 (\alpha_1)$. Hence, the cooperative pooling equilibrium can be sustained under a larger set of parameter values in the simultaneous prisoner’s dilemma game as compared with the sequential version, i.e., $q > q^{\text{Seq}}_1 (\alpha_1) > q^{\text{Sim}}_1 (\alpha_1, \beta_1)$, if the uninformed first mover is highly concerned about fairness, i.e., $\beta_1^H > \frac{b - d}{b - c}$. Intuitively, when the 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, the informed player cannot sequentially respond to the uninformed player’s action, leading to an expected payoff of \( q[b - \beta_1^H (b - c)] + (1 - q)d \) from playing defection. Therefore, the uninformed player obtains a higher payoff from playing defection in the sequential than in the simultaneous-move game if his guilt feeling is substantial, i.e., if \( \beta_1^H \geq \frac{b - d}{b - c} \). Under this condition, defection becomes more attractive in the sequential-move than in the simultaneous-move version of the game. The cooperation that the pooling equilibrium prescribes, therefore, can only be sustained in the sequential game if prior beliefs are relatively high, i.e., \( q > q^\text{Seq} (\alpha_1) > q^\text{Sim} (\alpha_1, \beta_1) \). This result yields the testable implication that cooperation is more easily sustained during the first periods of a finitely-repeated interaction if players are asked to interact in a simultaneous-move as opposed to a sequential-move PD game as studied in Healy (2007). Furthermore, note that a given reduction in the uninformed player’s envy, \( \alpha_1 \), produces a larger decrease in \( q^\text{Sim} (\alpha_1, \beta_1) \) than in \( q^\text{Seq} (\alpha_1) \), 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 game. If, in contrast, the uninformed player is not highly concerned about social preferences, i.e., \( \frac{b - a}{b - c} < \beta_1^H < \frac{b - d}{b - c} \), cutoffs instead satisfy \( q^\text{Sim} (\alpha_1, \beta_1) > q^\text{Seq} (\alpha_1) \), and the cooperative equilibrium can be sustained under larger parameter conditions in the sequential than in the simultaneous-move game. This result coincides with that of incomplete information games where all players are uninformed about each others’ concerns for fairness. Hence, when \( \beta_1^H \geq \frac{b - d}{b - c} \) holds, cooperation can be supported under larger parameter values in the simultaneous than in the sequential version of the game, both when all players are uninformed and when only one player is.

**Changes in the informational structure.** For comparison purposes, we examine equilibrium play in the sequential PD game where the player holding private information about social preferences is switched, from the second to the first mover. The following corollary shows that under this informational setting only the “cooperative” pooling equilibrium can be supported.

**Corollary 1.** If players interact in a sequential-move PD game where the first mover is privately informed about his concerns for fairness, only the pooling strategy profile described in Proposition 5 can be supported as a PBE of the signaling game. Moreover, it can be sustained for all second mover’s beliefs \( \mu (\beta_1^H | C) = q \in (0, 1) \) and \( \mu (\beta_1^H | D) \in (0, 1) \).

In particular, the uninformed player (second mover) can react to any deviation towards defection of the first mover, both during the first and second period of the game, since he observes the first-mover’s action before choosing his. This time structure, hence, protects the uninformed second mover from any potential exploitation attempt of his opponent. This differs from the sequential PD game analyzed in Proposition 5, where the uninformed player was the first mover. In that

---

Note that a given equilibrium (either separating or pooling) can only be supported under the same parameter values if the uninformed player’s concern for fairness is exactly \( \beta_1^H = \frac{b - d}{b - c} \). The results in this section can be easily extended to continuous probability distributions, \( G(\beta_1) \), as described in sections 3 and 4. Note that under a continuous probability distribution, the probability that concerns for fairness takes the exact realization \( \beta_1^H = \frac{b - d}{b - c} \) approaches zero.
context he could still be exploited by the second mover in the last period of the game.\textsuperscript{28}

6 Conclusions

A large experimental literature has provided evidence that subjects appear to exhibit other-regarding or “social” preferences as opposed to the more standard self-regarding preferences that are typically assumed by theories. Social preferences have been formally modeled in an effort to explain why experimental data often depart from equilibrium predictions. To date, all models of social preferences have assumed that the players can perfectly observe one another’s social preference parameters. This might be a reasonable assumption in strategic environments where players have been interacting with one another for long periods of time, allowing their preference parameters to be revealed through their prior action choices. Nonetheless, in contexts where such a long history of play is not available, incomplete information regarding the social preferences of other players seems a more reasonable assumption.

In this paper we examine how equilibrium play in the simultaneous and sequential-move Prisoner’s Dilemma game is affected by the introduction of incomplete information regarding players’ social preferences. Our results are mainly evaluated according to two criteria: (1) whether the efficient cooperative outcome can be supported under larger sets of parameter values relative to the complete information environment; and (2) whether the introduction of uncertainty helps select a unique equilibrium outcome. We then investigate information transmission when players interact in a twice-repeated simultaneous prisoner’s dilemma game. A separating equilibrium can be supported in which a player with high concerns about fairness bears the cost of cooperating in order to reveal his type to his uniformed opponent, thus promoting cooperation in subsequent periods. We also identify a pooling equilibrium in which a player unconcerned about inequity aversion initially cooperates in order to mislead the uninformed player. Specifically, this misleading strategy induces the uninformed player to cooperate in the subsequent game, when the unconcerned player takes the opportunity to defect, yielding outcome \( (D, C) \). This “backstabbing” equilibrium might explain subjects’ “end-game” behavior where they behave non-cooperatively in the final play of experimental games such as Prisoner’s Dilemma. We also demonstrate that this type of equilibrium can be supported in the sequential version of the stage game, although under different parameter conditions. This helps us compare our results with those in the literature and predict which version of the game can sustain cooperation under a larger set of parameter values.

In this paper we assume that players’ social preferences are independently distributed. This assumption, however, can be relaxed if every individual’s private concerns about fairness informs him about the conditional probability that his opponent’s social preferences take values in a certain interval. Instead of considering \( G \left( \frac{b-a}{b-c} \right) \) in conditions \( A \) and \( B \), every player \( i \) could consider

\textsuperscript{28}Note that, in the context of the simultaneous PD game, our equilibrium results in section 5.1 would not be affected if we modify which player holds private information about his social preferences, either player \( i \) or \( j \), since our results in those propositions are valid for any player \( i = \{1, 2\} \) and \( j \neq i \).
the conditional probability $G\left(\frac{b-a}{b-c} | \beta_i \right)$, i.e., the probability that his opponent is relatively concerned about fairness, given that his own concern is $\beta_i$. In particular, if $G\left(\frac{b-a}{b-c} | \beta_i \right) < G\left(\frac{b-a}{b-c} \right)$, then cooperation is more likely to be sustained relative to the case where players’ social preferences are independently distributed. Intuitively, by observing $\beta_i$, player $i$ can infer that player $j$’s concerns are relatively high, making player $i$ more likely to cooperate than when social concerns are independently distributed. Of course, we come to the opposite conclusion is the case where $G\left(\frac{b-a}{b-c} | \beta_i \right) > G\left(\frac{b-a}{b-c} \right)$.

For compactness, we focus on a game – the Prisoner’s Dilemma– with strong competitive pressures, where the conflict between individual and social preferences is intense. Nonetheless, our analysis can be directly applied to other strategic environments where social preferences have been extensively studied, such as coordination and anti-coordination games. This would provide a better understanding of whether incomplete information about concerns for fairness induces more cooperation and how agents use their actions to reveal or conceal their social preferences to other players.

7 Appendix

7.1 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 $\mu(\beta_i^H | D) = q < q^{Sim}(\alpha_j, \beta_j)$ and $\mu(\beta_i^H | C) < q^{Sim}(\alpha_j, \beta_j)$; 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 $\alpha_i^L + \beta_i^L \geq \frac{c+b-2a}{b-c}$. 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 $\mu(\beta_i^H | D) = q < q^{Sim}(\alpha_j, \beta_j) < \mu(\beta_i^H | C)$.

c. Player $i$ defects in the second period when he is concerned about fairness, but cooperates in equilibrium otherwise given $\beta_i \leq \frac{b-d}{b-c}$. 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 $\mu(\beta_i^H | D) = q \geq q^{Sim}(\alpha_j, \beta_j) > \mu(\beta_i^H | C)$. 

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 \( \mu(\beta^H_{i}|C) \equiv \mu \in [0,1] \) and \( \mu(\beta^H_{i}|D) = q \). Given these beliefs, let us now analyze player \( j \)’s best response during the second period of the game. In particular, after observing \( D \) 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 \( q \geq \frac{d-c+\alpha_i(b-c)}{a+d-c-b+(\alpha_j+\beta_{ij})(b-c)} \equiv q^{\text{Sim}}(\alpha_j, \beta_j) \). Note that this cutoff strategy coincides with the one player \( j \) uses when selecting between \( C \) and \( D \) at the beginning of the first period. After observing \( C \) in the first period (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 \( \mu \geq q^{\text{Sim}}(\alpha_j, \beta_j) \). Let us now investigate the informed player \( i \)’s actions during the first period:

1. If \( q, \mu \geq q^{\text{Sim}}(\alpha_j, \beta_j) \) 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 \( a + a \geq b - \beta^H_i(b - c) + a \), which holds since \( \beta^H_i \geq \frac{b-a}{b-c} \) by definition. Hence, the prescribed strategy profile cannot be supported as a pooling PBE if \( q, \mu \geq q^{\text{Sim}}(\alpha_j, \beta_j) \).

2. If \( q, \mu < q^{\text{Sim}}(\alpha_j, \beta_j) \) 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 \( c - \alpha^H_i(b - c) + d \leq d + d \), which implies \( \alpha^H_i \geq 0 > \frac{c-d}{b-c} \), which holds by definition. On the other hand, the unconcerned player \( i \) defects if \( c - \alpha^L_i(b - c) + d \leq d + d \), which implies \( \alpha^L_i \geq 0 > \frac{c-d}{b-c} \), which also holds by definition. Therefore, this pooling strategy profile can be sustained as a PBE if \( q, \mu < q^{\text{Sim}}(\alpha_j, \beta_j) \).

3. If \( q \geq q^{\text{Sim}}(\alpha_j, \beta_j) > \mu \) 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 \( a + d \leq b - \beta^H_i(b - c) + a \), which is satisfied if \( \beta^H_i \leq \frac{b-d}{b-c} \), where \( \frac{b-a}{b-c} \leq \beta^H_i \leq \frac{b-d}{b-c} \). On the other hand, the unconcerned player \( i \) defects if \( a + d \leq b - \beta^L_i(b - c) + b - \beta^L_i(b - c) \), which holds if \( \beta^L_i < \frac{2b-a-d}{2(b-c)} \) which is satisfied since \( \beta^L_i < \frac{2b-a-d}{2(b-c)} \). Hence, this pooling strategy profile can be supported as PBE if \( q \geq q^{\text{Sim}}(\alpha_j, \beta_j) > \mu \).

20
4. If \( q < q^\text{Sim}(\alpha_j, \beta_j) \leq \mu \) 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 \( c - \alpha_i^L(b - c) + a \leq d + d \), 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 \( q < q^\text{Sim}(\alpha_j, \beta_j) \leq \mu \), 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 \( \alpha_i^L + \beta_i^L \geq \frac{c + b - 2d}{b - c} \). \( \blacksquare \)

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 as 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, \( \alpha_i^H \geq \frac{a + c - 2d}{b - c} \) and \( \alpha_i^L + \beta_i^L \geq \frac{c + b - 2d}{b - c} \). Finally, in the pooling equilibrium described in part \( (c) \), the uninformed player \( j \) interprets a deviation towards cooperation as most likely coming from an unconcerned player \( i \), i.e., \( \mu (\beta_i^H | C) < q^\text{Sim}(\alpha_j, \beta_j) \), 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 the Cho and Kreps’ (1987) Intuitive Criterion. By contrast, we demonstrate that all other equilibria do survive this refinement criterion under relatively large parameter conditions.

7.2 Appendix 2 - Equilibrium refinement

Lemma A. All PBEs described in Propositions 3-4 survive 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 \( (c) \) under all parameter values, and the pooling equilibrium described in part \( (a) \) if social preferences satisfy \( \alpha_i^H < \frac{a + c - 2d}{b - c} \) and \( \alpha_i^L + \beta_i^L \geq \frac{c + b - 2d}{b - c} \).

Proof. Proposition 3. In the separating PBE of the game no action by the informed player \( i \) is regarded as off-the-equilibrium since all actions are used by some type of player \( i \) with positive probability i.e., cooperation is selected by the highly concerned player \( i \) while defection is chosen
by the unconcerned type of player $i$. The uninformed player $j$ does not sustain off-the-equilibrium beliefs to be restricted using the Cho and Kreps’ Intuitive Criterion. Therefore, the separating PBE survives the Intuitive Criterion.

**Proposition 4.** Let us first check the pooling PBE where both types of player $i$ cooperate in the first period under $q \geq q^{Sim}(\alpha_j, \beta_j)$. Regarding the unconcerned player $i$, if he deviates towards defection, the highest payoff he can obtain is $b - \beta_i^T(b - c) + b - \beta_i^L(b - c)$, which exceeds his equilibrium payoff of $a + b - \beta_i^L(b - c)$. Regarding the highly concerned player $i$, if he deviates to defection, the highest payoff he can obtain is $a + a$, which coincides with his equilibrium payoff 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 $\mu(\beta_i^H|D) = 0$. Given these beliefs, player $j$ defects in the second period after observing defection (since it can only come from an unconcerned player $i$), yielding a total payoff for the unconcerned player $i$ of $b - \beta_i^T(b - c) + d$, which does not exceed his equilibrium payoff given that $a + b - \beta_i^L(b - c) \geq b - \beta_i^T(b - c) + d$ 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 payoff he can obtain is $c - \alpha_i^H(b - c) + a$, which exceeds his equilibrium payoff 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 $\beta_i^H \geq \frac{b - a}{b - c}$ if $b - c < 2(a - d)$]. Regarding the unconcerned player $i$, if he deviates towards cooperation the highest payoff he can obtain is $c - \alpha_i^L(b - c) + b - \beta_i^T(b - c)$, which exceeds his equilibrium payoff of $d + d$ only if $\alpha_i^L + \beta_i^T < \frac{c + b - 2d}{b - c}$. 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 $\alpha_i^L + \beta_i^T < \frac{c + b - 2d}{b - c}$ hold, both types of player $i$ have incentives to deviate, and the uninformed player $j$ cannot restrict his off-the-equilibrium-path beliefs. 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 $\alpha_i^L + \beta_i^T < \frac{c + b - 2d}{b - c}$ does not, the concerned player $i$ has incentives to deviate towards cooperation but the unconcerned does not. Hence, the uninformed player $j$ restricts his beliefs to $\mu(\beta_i^H|C) = 1$ and $\mu(\beta_i^H|D) = 0$, and cooperates in the second period after observing $C$ but defects after observing $D$. Given this response by player $j$ during the second period of the game, the highly concerned player $i$ has incentives to cooperate in the first period since he obtains $c - \alpha_i^L(b - c) + a$, which exceeds his equilibrium payoff of $d + d$, given that $\alpha_i^L < \frac{a + c - 2d}{b - c}$ holds. Hence the pooling PBE where both players defect given $q, \mu < q^{Sim}(\alpha_j, \beta_j)$ violates the Intuitive Criterion if $\alpha_i^H < \frac{a + c - 2d}{b - c}$ and $\alpha_i^L + \beta_i^T < \frac{c + b - 2d}{b - c}$.

3. When neither condition $\alpha_i^H < \frac{a + c - 2d}{b - c}$ nor $\alpha_i^L + \beta_i^T < \frac{c + b - 2d}{b - c}$ hold, then neither type of player
i has incentives to deviate. Consequently, the uninformed player j cannot restrict his off-the-equilibrium-path beliefs, and no type of player i has incentives to change his equilibrium action. 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 \( \mu(\beta_i^L|D) = 1 \), since cooperation can only come from the unconcerned player. Consequently, player j cooperates in the second period after observing D but defects after observing C. Given 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 payoff 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 payoff 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 \( \alpha_i^H > \frac{a + c - 2d}{b - c} \) but \( \alpha_i^L + \beta_i^L < \frac{c + b - 2d}{b - c} \).

**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 payoff he can obtain is \( c + \alpha_i^H(b - c) + a \), which exceeds his equilibrium payoff 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 payoff he can obtain is \( c + \alpha_i^L(b - c) + b - \beta_i^L(b - c) \), which exceeds his equilibrium payoff 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 payoff he can obtain is \( a + a \), which exceeds his equilibrium payoff of \( b - \beta_i^H(b - c) + a \) since \( \beta_i^H > \frac{b - a}{b - c} \) by definition. Hence, the highly concerned player has incentives to deviate. Regarding the unconcerned player i, if he deviates towards cooperation, the highest payoffs he can achieve is \( a + b - \beta_i^L(b - c) \), which does not exceed his equilibrium payoff 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 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 payoff 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.

Starting from the pooling equilibrium in which both types of player i defect, described in Proposition A(c), the highly concerned player i can “separate” from the unconcerned type by
deviating towards cooperation. This allows the uninformed player \( j \) to assign full probability to player \( i \) being concerned (unconcerned) after observing cooperation (defection, respectively). Player \( j \) hence cooperates after observing cooperation from player \( i \) but defects otherwise. Given these updated beliefs for the uninformed player, the highly concerned player \( i \) cooperates during the first period, violating the pooling equilibrium in which both types of player \( i \) defect.

### 7.3 Proof of Proposition 1

First, note that when \( \beta_i < \frac{b-a}{b-c} \), defection becomes a strictly dominant strategy for player \( i \) regardless of player \( j \)'s social preferences. In contrast, when \( \beta_i \geq \frac{b-a}{b-c} \), defection is not a strictly dominant strategy, and player \( i \) cooperates only if \( G \left( \frac{b-a}{b-c} \right) \) satisfies condition A. If condition A holds and both players’ concerns satisfy \( \beta_i, \beta_j \geq \frac{b-a}{b-c} \), then they both cooperate and \((C,C)\) is the unique Bayesian Nash equilibrium of the game. If both players are concerned but condition A is violated, then they both defect, yielding outcome \((D,D)\) as the unique equilibrium. If condition A is satisfied but only one player is relatively concerned about fairness, then he is the only player cooperating, yielding either outcome \((C,D)\) or \((D,C)\). If only one player is relatively concerned but condition A does not hold, then \((D,D)\) is the unique equilibrium of the game. Finally, if both players are relatively concerned, \( \beta_i, \beta_j < \frac{b-a}{b-c} \), then outcome \((D,D)\) can be supported as the unique Bayesian Nash equilibrium of the game. ■

### 7.4 Proof of Lemma 1

From the text we know that the second mover best response 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

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

When the second mover’s concerns about fairness satisfy \( \beta_j \geq \frac{b-a}{b-c} \), the second mover responds by selecting the action selected by the first mover. Therefore, the first mover’s payoff from cooperating (which is responded with cooperation) is \( a \), while that from defecting (which is responded 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 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. Under this case, if the first mover selects cooperation his payoff 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 concerns satisfy \( \beta_j < \frac{b-a}{b-c} \). ■
7.5 Proof of Proposition 2

The second mover’s best response function coincides with that in Lemma 1. Regarding the first mover, he cooperates if condition $B$ holds, given that he is uninformed about the second mover’s concern for fairness. Hence, if condition $B$ holds, the first mover cooperates, which triggers a cooperative response from the second mover if $\beta_j > \frac{b-a}{b-c}$, yielding outcome (C,C). If condition $B$ holds but $\beta_j < \frac{b-a}{b-c}$, then the first mover is the only cooperator, yielding outcome (C,D). Finally, if condition $B$ is not satisfied, the first mover defects and the second mover responds by defecting, regardless of his concern for fairness, yielding outcome (D,D).

7.6 Proof of Proposition 3

Let us first analyze the separating equilibrium 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 equilibrium 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). Since $j$ is assumed to be highly concerned about fairness, $\beta_j > \frac{b-a}{b-c}$, 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 $i$ 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 of the 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 game:

1. If $q > q^{Sim}(\alpha_j, \beta_j)$, the uninformed player $j$ cooperates during the first period. On the one hand, the informed, highly concerned player $i$ thus cooperates if $a + a \geq b - \beta_i^H(b-c) + d$, since defection is responded to with defection in the subsequent period. That is, if $\beta_i^H \geq \frac{b-d-2a}{b-c}$. And since $\beta_i^H > \frac{b-a}{b-c} > \frac{b-d-2a}{b-c}$, the above condition is satisfied and therefore the informed, highly concerned player $i$ cooperates in the first period, as prescribed in this separating equilibrium. On the other hand, the informed, unconcerned player $i$ cooperates if $a + [b - \beta_i^L(b-c)] \geq [b - \beta_i^L(b-c)] + d$, i.e., $a \geq d$. Hence the unconcerned player $i$ also cooperates in the first period, which violates the prescribed strategy profile in this separating equilibrium.
2. If \( q < q^{\text{Sim}}(\alpha_j, \beta_j) \), the uninformed player \( j \) defects during the first period. On the one hand, the informed, 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 informed, unconcerned player \( i \) defects since \( c - \alpha_i^L(b - c) + c - \alpha_i^L(b - c) \leq d + d \), given that this implies \( \frac{c - d}{b - c} \leq \alpha_i^L \), which is true by definition. We therefore can support a separating equilibrium in which player \( j \) cooperates if and only if he is highly concerned about fairness if the prior probability \( q \) satisfies \( q < q^{\text{Sim}}(\alpha_j, \beta_j) \) and envy satisfies \( \alpha_i^H \leq \frac{a + c - 2d}{b - c} \).

Let us now analyze the separating equilibrium 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 game, player \( j \)'s beliefs in this separating equilibrium 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 first period, he believes that his opponent is unconcerned (according to the strategy profile prescribed in this separating equilibrium), and hence his opponent will not cooperate in the second-period game. Player \( j \) defects as a consequence in the second-period PD game since \( d > c - \alpha_j(b - c) \). After observing \( D \) in the first period, player \( j \) believes that his opponent is highly concerned, and that hence his opponent will cooperate in the second period. Therefore, player \( j \) cooperates in the second period since \( a > b - \beta_j(b - c) \) given that \( \beta_j > \frac{b - a}{b - c} \) 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 \( q \geq q^{\text{Sim}}(\alpha_j, \beta_j) \). Let us now investigate the informed player \( i \)'s actions during the first-period game:

1. If \( q > q^{\text{Sim}}(\alpha_j, \beta_j) \), the uninformed player \( j \) cooperates during the first period. On the one hand, the informed, 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 \( a + d \geq b - \beta_i^H(b - c) + b - \beta_i^H(b - c) \) which implies \( \beta_i^L > \frac{2b - a - d}{2(b - c)} \), which cannot hold since \( \frac{2b - a - d}{2(b - c)} \geq \frac{b - a}{b - c} \) and \( \beta_i^L \leq \frac{b - a}{b - c} \). Hence, this separating strategy profile cannot be sustained if \( q > q^{\text{Sim}}(\alpha_j, \beta_j) \).

2. If \( q < q^{\text{Sim}}(\alpha_j, \beta_j) \), the uninformed player \( j \) defects during the first period. On the one hand, the informed 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 informed 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 \( q < q^{\text{Sim}}(\alpha_j, \beta_j) \). ■
Let us analyze the pooling equilibrium 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, player \( j \)’s beliefs in this pooling equilibrium are \( \mu(\beta_i^H|C) = q \) (in equilibrium) and \( \mu(\beta_i^H|D) = \gamma \in [0,1] \) (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 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 \( q \geq q^{Sim}(\alpha_j, \beta_j) \). Note that this cutoff strategy coincides with the one that player \( j \) uses when selecting between \( C \) and \( D \) prior to playing the first period of the game. After observing \( D \) in the first period (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 in the first period of the 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 \( \beta_i^H \geq \frac{b-a}{b-c} \). 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 since \( \beta_i^L < \frac{b-a}{b-c} \). Therefore, this pooling strategy profile cannot be supported if \( q, \gamma > q^{Sim}(\alpha_j, \beta_j) \).

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 a \( C \) in the first period. On the one hand, the highly concerned player \( i \) cooperates since \( a + a \geq b - \beta_i^H(b-c) + d \), or \( \beta_i^H \geq \frac{b+2d-a}{b-c} \). Given that \( \frac{b-a}{b-c} > \frac{b+2d-a}{b-c} \), then condition \( \beta_i^H \geq \frac{b+2d-a}{b-c} \) 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 \( q > q^{Sim}(\alpha_j, \beta_j) > \gamma \).
4. If \( q < q^{\text{Sim}}(\alpha_j, \beta_j) \leq \gamma \) player \( j \) defects in the first period of the game and in the second period he defects only after observing \( C \). 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 \( q < q^{\text{Sim}}(\alpha_j, \beta_j) \leq \gamma \). ■

7.8 Proof of Proposition 5

Separating PBE. Let us first analyze the separating equilibrium where the second mover cooperates when being concerned but defects otherwise. First, note that after observing an action from the second mover in the first-period game, the first mover’s beliefs about \( \beta_2^H \) 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, but defects otherwise. After these choices, the second mover reciprocates the first mover in the second-period game if the second mover is concerned, or defects otherwise.

During the first period, the second mover cooperates when being concerned but defects otherwise, as prescribed. Hence, the 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 \( q \geq q^{\text{Seq}}(\alpha_1) \). Let us finally investigate the second mover’s action during the first-period game:

1. If \( q \geq q^{\text{Seq}}(\alpha_1) \), the first mover cooperates in the first-period 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 period. That is, if \( \beta_2^H \geq \frac{b - d - 2a}{b - c} \). And since \( \beta_2^H > \frac{b - a}{b - c} > \frac{b - d - 2a}{b - c} \) by definition, the above condition holds. Therefore the concerned second mover cooperates in the first period, as prescribed in this separating equilibrium. 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, which violates the prescribed strategy profile in this separating equilibrium.

2. If \( q < q^{\text{Seq}}(\alpha_1) \), the first mover defects in the first-period 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 \( \frac{c + b - 2d}{b - c} \leq \alpha_2^L + \beta_2^L \). We therefore can support a separating equilibrium in which the second mover cooperates only when he is concerned about fairness if the first-mover’s beliefs satisfy \( q < q^{\text{Seq}}(\alpha_1) \) and parameter values satisfy \( \alpha_2^H \leq \frac{a + c - 2d}{b - c} \) and \( \frac{c + b - 2d}{b - c} \leq \alpha_2^L + \beta_2^L \).

Pooling PBE. Let us now analyze the pooling equilibrium where both types of second mover cooperate in the first-period game. First, note that after observing an action from the second mover during the first period, first mover’s beliefs about \( \beta_2^H \) in this pooling equilibrium cannot
updated using Bayes’ rule and hence are \( \mu(\beta^H_2 | C) = q \) (in equilibrium) and \( \mu(\beta^H_2 | D) \equiv \gamma \in [0, 1] \) (off-the-equilibrium path). Given these beliefs, the first mover cooperates in the second period after observing that the second mover chose C (in equilibrium) in the first period if \( q \geq q^{\text{Seq}}(\alpha_1) \). If the first mover observes the second mover selecting D in the first period (off-the-equilibrium), then the first mover cooperates in the second-period game if and only if

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

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

1. If \( q, \gamma \geq q^{\text{Seq}}(\alpha_1) \) the first mover cooperates in both the first and second-period game after observing any action from the second mover. On one hand, the concerned second mover cooperates if \( a + a \geq b - \beta^H_2(b - c) + a \), which holds since \( \beta^H_2 \geq \frac{b-a}{b-c} \). On the other hand, the unconcerned second mover defects since \( a + b - \beta^H_2(b - c) \leq b - \beta^H_2(b - c) + b - \beta^H_2(b - c) \), which holds since \( \beta^H_2 \leq \frac{b-a}{b-c} \) by definition. Therefore, this pooling strategy profile cannot be supported if \( q, \gamma \geq q^{\text{Seq}}(\alpha_1) \).

2. If \( q, \gamma < q^{\text{Seq}}(\alpha_1) \) the first mover defects both in the first and second-period game, both after observing that the second mover selects C and D. The concerned second mover defects, however, since \( c - \alpha^H_2(b - c) + d \leq d + d \), which implies \( \alpha^H_2 \geq 0 \geq \frac{c-d}{b-c} \), which holds by definition. Hence, this pooling strategy profile cannot be sustained if \( q, \gamma < q^{\text{Seq}}(\alpha_1) \).

3. If \( q \geq q^{\text{Seq}}(\alpha_1) > \gamma \) the first mover cooperates in the first-period game and in the second period he cooperates only after observing C (in equilibrium). On one hand, the concerned second mover cooperates since \( a + a \geq b - \beta^H_2(b - c) + d \), or \( \beta^H_2 \geq \frac{b+d-2a}{b-c} \). Given that \( \frac{b-a}{b-c} > \frac{b+d-2a}{b-c} \), then condition \( \beta^H_2 \geq \frac{b+d-2a}{b-c} \) is satisfied from \( \beta^H_2 \geq \frac{b-a}{b-c} \). On the other hand, the unconcerned second mover cooperates since \( a + b - \beta^H_2(b - c) \geq b - \beta^H_2(b - c) + d \), or \( a \geq d \). Hence this pooling strategy profile can be supported as a PBE for \( q \geq q^{\text{Seq}}(\alpha_1) > \gamma \).

4. If \( q < q^{\text{Seq}}(\alpha_1) \leq \gamma \) the first mover defects in the first-period game and in the second period he defects only after observing C (in equilibrium). The concerned second mover defects, however, since \( c - \alpha^H_2(b - c) + d \leq d + a \), which implies \( \alpha^H_2 \geq 0 \geq \frac{c-a}{b-c} \) which is satisfied by definition. Thus, this pooling strategy profile cannot be sustained if \( q < q^{\text{Seq}}(\alpha_1) \leq \gamma \).

### 7.9 Proof of Corollary 1

**Separating PBE.** Let us first analyze the separating equilibrium where the first mover cooperates when being concerned but defects otherwise. First, note that after observing an action from the first mover in the first-period game, the second mover’s beliefs about \( \beta^1_1 \) are \( \mu(\beta^1_1 | C) = 1 \) and \( \mu(\beta^1_1 | D) = 0 \). Recall that the second mover is concerned about fairness by definition, \( \beta_2 \geq \frac{b-a}{b-c} \), and that this information is common knowledge. Hence, the second mover’s best response is to
“mimic” the action selected by the first mover, both in the first and second-period sequential PD games. Regarding the first mover, when being concerned he cooperates since \(a + a \geq d + d\). If unconcerned, the first mover defects (as prescribed) if \(a + d < d + d\), which does not hold. Hence, the separating strategy profile cannot be supported as a PBE of the game.

Pooling PBE. Let us now analyze the pooling equilibrium 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, second mover’s beliefs about \(\beta_1^H\) in this pooling equilibrium become \(\mu(\beta_1^H|C) = q\) and \(\mu(\beta_1^H|D) \equiv \gamma \in [0, 1]\). Because the second mover’s best response is to “mimic” the action selected by the first mover, we do not need to analyze equilibrium play under different beliefs, as we did in the proof of Proposition 5. Regarding the first mover, when being concerned he cooperates since \(a + a \geq d + d\). If unconcerned, the first mover also cooperates (as prescribed) given that \(a + d \geq d + d\). Hence, the pooling strategy profile can be supported as a PBE of the game, for all \(q\) and \(\gamma \in [0, 1]\).

References


