Journal of Economic Behavior and Organization 173 (2020) 1–25

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Revealed preferences in a sequential prisoners' dilemma: A horse-race between six utility functions☆

Topi Miettinen[a,*], Michael Kosfeld[b], Ernst Fehr[c], Jörgen Weibull[d,e]

[a] Hanken School of Economics and HECER, Arkadiankatu 7, P.O. Box 479, Helsinki FI-00101, Finland
[b] Faculty of Economics and Business, Goethe-Universität Frankfurt, Theodor-W.-Adorno-Platz 4, Frankfurt D-60323, Germany
[c] Department of Economics, University of Zürich, Blümlisalpstrasse 10, Zürich CH-8006, Switzerland
[d] Department of Economics, Stockholm School of Economics, PO Box 6501, Stockholm SE-11383, Sweden
[e] Institute for Advanced Study in Toulouse, 21 Allée de Brienne, Toulouse F-31000, France

ARTICLE INFO

Article history:
Received 11 July 2019
Revised 21 February 2020
Accepted 25 February 2020

JEL classification:
C72
C9
D03
D84

Keywords:
Cooperation
Prisoners' dilemma
Other-regarding preferences
Categorical imperative
Consensus effect
Optimism
Salience

ABSTRACT

We experimentally investigate behavior and beliefs in a sequential prisoner's dilemma. Each subject had to choose an action as first mover and a conditional action as second mover. All subjects also had to state their beliefs about others' second-mover choices. Using these elicited beliefs, we apply the transparent Selten–Krischker approach to compare the explanatory power of a few current models of social and moral preferences. We find clear differences in explanatory power between the preference models, both without and with control for the number of free parameters. The best-performing models explain about 80% of the observed behavior. We compare our results with those obtained from a conventional maximum-likelihood approach, and find that the results by and large agree. We also present a structural model of belief formation. We find a consensus bias—whereby subjects believe others behave like themselves—and payoff-salience driven optimism—whereby subjects overestimate the probabilities for favorable outcomes.

1. Introduction

Alternative specifications of social preferences have been discussed and analyzed in the behavioral and experimental economics literature. Recently, a lively debate has emerged about how potential belief biases influence behavior, in particular concerning conditional cooperation in sequential prisoners' dilemmas and trust games. As noted by Brandts and Charness (2000), and Altmann et al. (2008), there is a tendency for within-subject positive correlation between first-mover coopera-

☆ We thank two anonymous referees, the editor of this journal, as well as Anna Dreber-Almenberg, Manuel Bagues, Magnus Johannesson, Astri Muren, Tuomas Nurminen and Rickard Sandberg for helpful comments, and Tuomas Nurminen for excellent research assistance. The first author gratefully acknowledges financial support from the Yrjö Jahnsson Foundation (grant 7011), and Norwegian Research Council (grant 250506). The last author gratefully acknowledges financial support from the Knut and Alice Wallenberg Research Foundation and the Agence Nationale de la Recherche (Chaire IDEX ANR-11-IDEX-0002-02).

* Corresponding author.
E-mail address: topi.miettinen@hanken.fi (T. Miettinen).

https://doi.org/10.1016/j.jebo.2020.02.018
0167-2681/

tion and second-mover conditional cooperation. The latter of these papers points out that if beliefs about others' behavior had been correct, there would instead be a negative correlation according to many established preference models (such as the inequity aversion model of Fehr and Schmidt, 1999). Blanco et al. (2011) note, more generally, that within-subject correlations across various games and decision nodes are not consistent with the inequity-aversion model if subjects' beliefs are correct. There are also studies that have started to analyze belief biases as potential explanations for such inconsistencies. Gächter et al. (2012); Blanco et al. (2014) and Rubinstein and Salant (2016) suggest a role both for optimism (Weinstein, 1980) and for a (false) consensus effect (Ross et al., 1977), respectively.[1]

We here report results from a simple laboratory experiment based on a sequential prisoner's dilemma, that is, a dilemma in which one player moves first and the other player observes the first move and then makes a move (Clark and Sefton, 2001; Brandts and Charness, 2000). Subjects were randomly and anonymously matched in pairs. Each subject had to choose an action, C or D, both as first mover, and as second mover after each of the first mover's possible two actions.[2] All subjects also had to state their beliefs about others' second-mover choices. After this, we randomly assigned the roles as first and second mover within each pair, and the subjects' chosen actions were implemented and payoffs paid. Subjects were also paid according to the accuracy of their beliefs about other's choices. Our design is close to that of Fischbacher et al. (2001) and Fischbacher and Gächter (2010), who elicit and classify preference types by way of analyzing second-mover choices.

Our main contribution is to use subjects' stated beliefs about each others' behavior in a comparison of the predictive power of five current models of social preferences, and one model of Kantian morality as a motivating factor. We evaluate the predictive power of the models in several ways. Our main evaluation is based on the Selten and Krischker (1983) difference measure, but we also use the standard maximum-likelihood approach, and we compare alternative approaches with each other. An advantage of the Selten and Krischker approach, which uses the difference between "hit rate" and "hit area" as a measure of predictive power, is its transparency and usefulness even when data is relatively sparse (see Section 4.8 for a description of this approach). In all six models, we assume risk neutrality (see below).

Before calculating the models' Selten-Krischker scores, we examine what share of the subject population behaves in a way that is compatible with each model. The simplest model, *Homo oeconomicus* – maximization of own expected payoff – has a hit rate of about 28% of our observations. Unconditional *Altruism*, where a positive weight is placed on the other party's payoff, has a hit rate of about 44% of the observations. Fehr and Schmidt (1999) *Inequity aversion* model, in which negative weights are given to payoff differences between the two parties, with a bigger weight when the difference is to the subject's disadvantage, has a hit rate of about 60% of the observations. The fourth model is a version of Charness and Rabin (2002) model of a conditional concern for social welfare, conditioned upon the two parties' relative outcomes (just as the *Inequity aversion* model). The (slightly simplified) version tested here, which we call the *Conditional welfare* model, has a hit rate of about 82% of the observations. These four models all view decision makers as only concerned about the distribution of payoffs, and not how payoff distributions come about.

Models of the latter kind include reciprocity models, such as the more complex version of Charness and Rabin (2002), as well as Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006), and Cox et al. (2008). We consider the reciprocity model of Charness and Rabin (2002) model in our comparison. The model has a hit rate of about 83% of the observations.[3] The sixth and final model is the *Homo moralis* model of Alger and Weibull (2013, 2016), a model that attaches a positive weight to a version of Kant's categorical imperative. This mode turns out to have a hit rate of about 83% of the observations.[4] Hence, the *Conditional welfare, Reciprocity*, and *Homo moralis* models share the "first prize" in this preliminary horse-race between motivational models. Obviously, it is not sufficient to compare mere hit-rates since one model may be much more permissive than another. In particular, a "model" of indifference is consistent with any observation and thus has hit rate one, while not being helpful for prediction (Andersen et al., 2010; Selten and Krischker, 1983). Another related aspect not captured by pure hit-rates is model complexity. In the present context, we note that while the *Conditional welfare* model has two free parameters, and the *Reciprocity* model has three, the *Homo moralis* model has only one, and *Homo oeconomicus* none. We therefore run several horse races to study the robustness of our preliminary hit-rate results by way of using the Selten and Krischker (1983) method. In one test, we allow for considerable individual heterogeneity in the preference parameters for each model by way of dividing the subject pool into 8 groups. Second, we allow for less individual heterogeneity, by dividing the subject pool into 4 groups. Third, we compare the hit-rates and Selten–Krischker estimates to more traditional maximum-likelihood estimates. When comparing models in terms of maximum likelihoods, we penalize models for their numbers of parameters.

Altogether, we run eight horse races between our six models. It turns out that the *Conditional welfare* and *Reciprocity* models are among the best performing, though in some races they are beaten by either the *Homo moralis* or the *Inequity*

[1] See Hey (1984); Puri and Robinson (2007); Bellemare et al. (2008); Gächter et al. (2012); Muren (2012); Spinnewijn (2015), and Dillenberger et al. (2017) for studies on the role of optimism in economic behavior, and see Nosenzo and Tufano (2017), and Engelmann et al. (2019) for studies of the consensus effect.

[2] That is, we use the so-called strategy method (Selten, 1967). Brandts and Charness (2000) study whether this method, as compared with actual decisions at the moment (in the "hot" state), triggers differences in rates of conditional cooperation. They find no statistical significance. Yet, statistically significant differences have been observed in other contexts, see Iriberri and Rey-Biel (2011). See Brandts and Charness (2011) for a survey of the literature on the strategy method.

[3] Levine (1998) model of conditional altruism and spite is difficult to identify with our small data set, and thus is not analyzed in this paper.

[4] Kant (1964): "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."

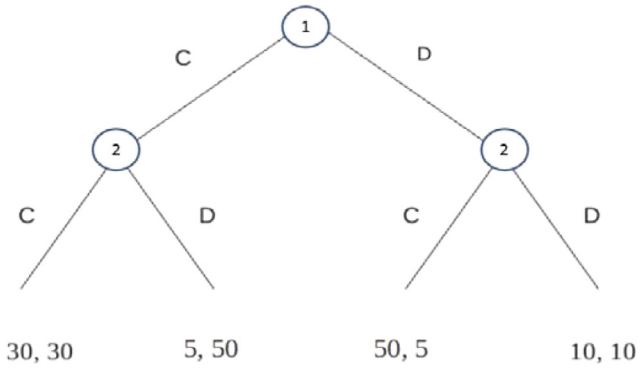

Fig. 1. A sequential prisoners' dilemma.

aversion model. This robustness in the rankings from the different races suggests that the Selten–Krischker approach does not compromise on the capacity to identify relevant patterns in the data, as compared with more traditional methods. Our results complement the literature on structural estimation of social preferences, see Bruhin et al. (2018) and the references therein.

Our second contribution is to shed light on subjective beliefs about others' behavior. We find that, on average, subjects' beliefs are fairly accurate. Yet, individual beliefs differ in a consistent manner. These differences can be explained in a unified way as a combination of a consensus bias, whereby individuals tend to believe that others act like themselves, and an optimism bias, whereby individuals tend to believe that favorable outcomes are more likely than unfavorable outcomes. Indeed, our data suggests that the subjects' stated own subjective beliefs as first movers, about second-mover cooperation rates, can be explained as the sum of three terms: the true empirical rate, own cooperation rate, and an optimism term. The last term we define in terms of the salience of the payoff influence from the opponent's response action, with salience defined in line with Bordalo et al. (2012, 2013). Our estimates for the average weights given to these three terms are approximately 0.53 for the true rate, 0.27 for own rate, and 0.19 for the optimism term.[5]

In summary, the paper sheds light on the two key motivational factors behind strategic behavior; beliefs and preferences. In all of our analysis, we make the simplifying assumption that agents are risk-neutral, both with respect to own and other players' payoff uncertainty. Recent evidence suggests that other-regarding preferences in the face of risk are more complicated than suggested by simple risk-neutrality (see Fudenberg and Levine, 2012; Trautmann and Vieider, 2012; Miettinen et al., 2020 and the references in these papers). No consensus has yet emerged regarding how risk and other-regarding attitudes should be integrated. We here proceed by abstracting from risk-aversion altogether, and discuss this limitations further in the concluding section.

The rest of the material is organized as follows. Section 2 describes the experimental design, Section 3 reports observations about average behaviors and beliefs, Section 4 specifies the different preference models and analyzes their predictions. Section 5 makes a model comparison, Section 6 presents results from our random-utility maximum likelihood estimations, and Section 7 analyzes belief biases. Section 8 concludes. Background calculations are provided in the Appendix at the end of the paper.

2. Experimental design

At the beginning of the experiment, subjects were given written instructions containing all the details of the experiment. To ensure the understanding of the experimental procedures all subjects had to answer several control questions. The experiment did not start until all subjects had answered all questions correctly. In addition, key aspects of the experiment were orally summarized. Subjects interacted in a Prisoner's Dilemma game form as is illustrated in Fig. 1.

The elicitation of subjects' preferences proceeded as follows. All subjects were randomly and anonymously matched into pairs, i.e., no subject knew the identity of her opponent. Each subject was asked to make a second-mover choice between C and D, both for the case when the other player—the first mover—plays C and D. In addition, each subject had to make an unconditional choice between C and D as first mover. In order to rule out possible sequencing effects, half of the subjects made their first-mover choice first and their second-mover choices second, while the other half made their second-mover choices first.

When the subjects had made their choices, each subject was asked to state his or her belief about the conditional choices of the opponent. More precisely, we asked each subject for his or her estimate of the probability that the second mover will

[5] It is a coincidence that the sum of these weights is close to unity. No restriction has been imposed on their sum, see Section 7.

Table 1
Subjects' first-mover choices (rows) and second-mover choices (columns).

	CC	CD	DC	DD	
C	7	33	1	13	0.56
D	2	3	5	32	0.44
	0.09	0.38	0.06	0.47	

Table 2
Subjects' beliefs about second-mover choices.

	Expected coop-rate cond. on C		Expected coop-rate cond. on D	
Behavioral class	**mean**	**s.d.**	**mean**	**s.d.**
DD	0.35	0.31	0.21	0.29
CD	0.67	0.26	0.12	0.23
CC	0.61	0.31	0.19	0.26
DC	0.27	0.35	0.39	0.55

cooperate if he or she as first mover cooperates or defects, respectively. The quadratic scoring rule was used to make the elicitation of beliefs incentive compatible.[6]

After both subjects in a pair had made their choices and stated their beliefs, in each session one subject threw a die to determine for whom of the subjects the unconditional decision and for whom the conditional decision was payoff relevant. Finally, subjects were informed about their and their opponent's payoff relevant decision and the resulting payoff they earned in the experiment.

In total 96 subjects participated in the experiment. All subjects were students either at the University of Zürich or the ETH Swiss Federal Institute of Technology in Zürich.[7] No subject participated in more than one session. All decisions had monetary consequences, where 10 payoff units represented 3 Swiss Francs (1 CHF = 0.59 USD at the time of the experiment). On average subjects received 27.70 Swiss Francs, including a show-up fee of 10 Swiss Francs. All decisions were made on a computer screen. We used the experimental software z-Tree (Fischbacher, 2007).

3. Average behaviors and average beliefs

We use the second-mover choices to categorize the participants into four *behavior classes* as follows: *unconditional cooperators*, who cooperate irrespective of the first-mover choice, *conditional cooperators*, who reciprocate the choice of the first mover, those who do the opposite of the first mover, and *unconditional defectors*, who defect irrespective of the first-mover's choice. We find 9, 36, 6, and 45 participants in each of these classes. In percentages, this amounts to population shares of approximately 9%, 38%, 6%, and 47%. We call those (few) who do the opposite of the first-mover*mismatchers*. See Table 1, where CC indicates unconditional cooperation, CD conditional cooperation, DC mismatching, and DD unconditional defection.

We then study the *average first-mover behavior* within each of these four behavior classes, and find important differences. Whereas only 29% of the unconditional defectors cooperate as first movers, as many as 92% of the conditional cooperators do. Of the two less frequent behavior classes, 78% of the unconditional cooperators cooperate as first movers, while only 17% of the mismatchers do so. Thus, in total 56% of all participants cooperate as first movers.

A key novelty in our model horse-race is that we allow for heterogeneous subjective beliefs. In summary, we find that participants, on average, believe that roughly 49% of the second movers will cooperate if the first mover cooperates, and that roughly 20% will cooperate if the first mover defects. According to our data, the true rates were 47% and 16%. There is thus, on average, a slight upward bias in participants' expectations about other participants' second-mover cooperation rates. There is, however, important heterogeneity in the subject pool. Table 2, illustrates this in terms of the four behavior classes.

In all but one of these eight conditional beliefs presented in there is a *consensus effect* (Ross et al., 1977; Blanco et al., 2014; Engelmann et al., 2019), that is, biased towards one's own behavior class. The one conditional belief, among the eight, that does not exhibit any consensus effect appears among the unconditional defectors. They expect a higher than actual cooperation rate conditional on defection (21% instead of 16%), although they defect themselves as second movers in the same situation.

[6] Subjects' inputs to the belief questions are allowed to take any value between 0 and 100, indicating the likelihood, in percentage terms, that the opponent chooses C. For a discussion of the use of the quadratic scoring role in economic experiments, see, e.g., Blanco et al. (2010); Schlag et al. (2015), and Trautmann and Kuilen (2015). The quadratic scoring rule is known to generate a bias if subjects are risk averse. In addition, subjects may play hedging strategies in the belief elicitation stage if choices and beliefs are paid for simultaneously.

[7] The experimental sessions were run in 2003.

We also examine statistically the cooperation rates expected by each of the two main behavior classes; unconditional defectors and conditional cooperators. Our results can be summarized as follows. First, unconditional defectors expect a more cooperative reaction to defection than conditional cooperators do; the expected cooperation rate after defection is negatively correlated with subjects' own response to cooperation.[8] Second, conditional cooperators expect a significantly more cooperative reaction to cooperation than unconditional defectors do; the expected cooperation rate after cooperation is positively correlated with subject's own response to cooperation.[9]

In another sequential prisoners' dilemma experiment, Altmann et al. (2008) found that conditional cooperation in the second-mover role was positively correlated with cooperation in the first-mover role, an observation they found puzzling since it is inconsistent with many models of other-regarding preferences under the hypothesis that the subjects have correct beliefs about each other's average behavior. Blanco et al. (2014) found evidence that the consensus effect might account for a major part of the puzzling variation. Indeed, that observation is consistent with our data for the beliefs elucidated from the conditional cooperators in our experiment. By contrast, the beliefs of the unconditional defectors in our data are on average not consistent with consensus bias. The bias in the expectations of the unconditional defectors about cooperation, conditional on defection, could instead be categorized as "optimism" in the sense of exaggerating the likely success (defined in terms of one's own preferences) of one's own behavior. We return to this issue in Section 4.7, where we propose a simple structural belief-formation model which derives belief biases in a unified way from personal preferences, thereby combining the consensus effect and optimism.

4. Other-regarding and moral preferences

The previous section, with evidence about beliefs that differ across behavior classes, highlights the importance of allowing for heterogeneous beliefs in studies of the explanatory power of other-regarding and moral preference models. We now turn to this main purpose of the study. In the six subsections to follow, we estimate model parameters so as to maximize the number of observations consistent with each model, thus obtaining the model's hit-rate. We show how that maximization problem can be illustrated graphically and intuitively. Moreover, once we have found the best fitting preference model that is consistent with the elicited behavior and subjective beliefs, we use the corresponding preference parameter estimates in Section 7 to propose a unified belief formation model which derives belief differences from differences in the revealed preference parameters. We then estimate best-fitting common weight parameters for optimism and consensus bias within this structural belief bias model. This is the second contribution of the paper.

When we estimate the parameters of a pre-specified preference model, all subjects are assumed to behave as if they maximized the subjectively expected value of that parametric goal function. We use the data we have about individual participants' subjective beliefs about other's choices when calculating the expected values. We consider six parametric (partly nested) families of goal functions, thus covering pure self-interest, inequity aversion (Fehr and Schmidt, 1999), a conditional concern for welfare and for reciprocity (Charness and Rabin, 2002), (unconditional) altruism (Becker, 1974; 1976), and Homo moralis (Alger and Weibull, 2013). For each model, we seek the parameter values that maximize the number of observations that are consistent with the model. In that task, we assume that (a) subjects do not make mistakes in the second-mover role (that is, they act in accordance with the hypothesized goal function), (b) subjective beliefs have been reported truthfully, (c) all individuals within each behavior class have the same parameter values in the hypothesized goal function (though we use their individually stated beliefs, which thus varies across individuals in the same behavior class), (d) they choose according to their predicted motivational goals in the first-mover role, (e) and they are risk-neutral.[10]

In the sequential prisoners' dilemma in Fig. 1, let $A_1 = A_2 = \{C, D\}$ be the set of actions available in player roles 1 and 2. Let $S_1 = \{C, D\}$ be the pure-strategy set of player role 1 (first mover), and $S_2 = \{CC, CD, DC, DD\}$ the pure-strategy set of player role 2 (second mover). Let $\pi_i(s)$, for $s = (s_1, s_2) \in S_1 \times S_2$, be the monetary payoff earned by a subject in player role $i = 1, 2$ when using pure strategy $s_i \in S_i$ against an opponent who uses strategy $s_j \in S_j$ (for $j \neq i$). We assume that subjects use pure strategies and believe that others use pure strategies too. Hence, the elicited beliefs about second-player moves are interpreted as beliefs about population fractions using different pure strategies.

[8] Consider the null hypothesis that the cooperation rate anticipated by the unconditional defectors is lower than that anticipated by the conditional cooperators. This hypothesis can be rejected at the 5%-level, at $p = 0.0323$ with Mann–Whitney U-test with continuity correction, and at $p = 0.0538$ with Student's t-test. Moreover, there is a significant negative correlation of -0.2085 (Pearson's product moment correlation coefficient) between a cooperative reaction to cooperation and the beliefs about cooperation in response to defection ($p = 0.0207$, one-sided).

[9] We find a significant (at $p = 0.01$, one-sided) positive correlation of 0.4562 (Pearson's product moment correlation coefficient). Moreover, we can reject the null hypothesis that the expectations of the unconditional defectors are higher than those of the conditional cooperators (Mann Whitney U test), at 1% level and by a large margin.

[10] While many studies estimate the best-fitting parameter values for the entire subject pool, our homogeneity assumption (c) allows for heterogeneity across the eight behavioral classes (see Table 1 for the behavioral classification). Estimating parameter values for each subject separately would be statistically questionable since we have so few data points per subject, and it would give too many degrees of freedom and would raise the hit-rates for many of the tested models to close to 1.

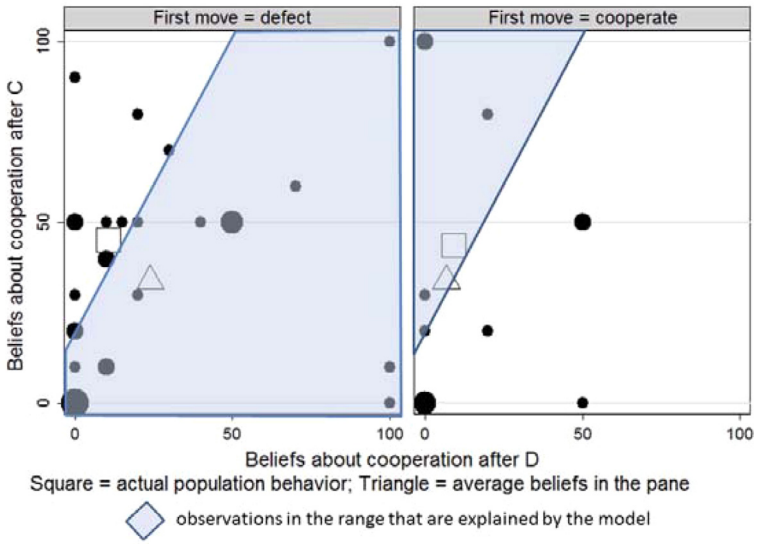

Fig. 2. Unconditional defectors. The explanatory power of the *Homo oeconomicus* model.

4.1. Homo oeconomicus

As our bench-mark goal function we take self-interest, that is, the goal function is then simply π_i for each player role $i = 1, 2$. This goal function evidently dictates unconditional defection in player role 2. In player role 1, it is optimal to cooperate if

$$30 \cdot p(C|C) + 5 \cdot p(D|C) \geq 50 \cdot p(C|D) + 10 \cdot p(D|D), \tag{1}$$

where $p(a_j|a_i) \in [0, 1]$ is the subject's (elicited) expectation about the second mover's action a_j if the subject takes first-mover action a_i. As indicated by (1), one can pin down the subjective expectations about second-mover cooperation rates that sustain cooperation in player role 1. Using $p(D|C) = 1 - p(C|C)$ and $p(D|D) = 1 - p(C|D)$, we can rewrite (1) as

$$p(C|C) \geq \frac{1}{5} + \frac{8}{5} \cdot p(C|D). \tag{2}$$

Fig. 2, illustrates this. Each point in each of the two panes represents the beliefs of a single unconditional defector. In the pane to the left, we have the beliefs of the unconditional defectors who defect as a first move. In the right pane, we have the beliefs of the unconditional defectors who cooperate as a first move. On the vertical axis we have the participant's belief about the cooperation rate conditional on cooperation (subjective estimate of the percentage of participants who react cooperatively to first-mover cooperation). On the horizontal axis we have the participant's belief about the cooperation rate conditional on defection. The size of each dot is proportional to the number of observations having the particular combination of beliefs. With beliefs above the upward-sloping straight line in each pane, first-mover cooperation is optimal for a *Homo oeconomicus* That is, with beliefs in the shaded areas, a *Homo oeconomicus* as first mover behaves optimally, given his or her beliefs, while subjects with beliefs in the white areas behave inconsistently with the *Homo oeconomicus* model.

The hollow square (at the same location in both panes) represents the average cooperation rate in the entire population, after defection (horizontally) and cooperation (vertically). (Both rates are expressed as shares and can thus be represented in the same manner as beliefs about cooperation.) The two hollow triangles represent average beliefs in each pane. Thus, the relative location of each panel's hollow triangle, with respect to the hollow square, reflects the direction of the bias in subjective beliefs (see Section 3). The figure shows that the distribution of individual beliefs is fairly similar in both (unconditional defector) panes. Hence, beliefs are not strongly correlated with choices in the first-mover role. The closer the dot lies to the top-left (bottom-right) corner, the stronger is the belief that the second mover conditionally cooperates (mismatches her action with that of the first mover). Thus, observations to the top-left (bottom-right) corner should be associated with first-mover cooperation (defection) if the self-interest model is correct.

One also sees that in the left pane there are many observations above the upward-sloping straight line (the white area in the left pane). These are the subjects who, if self-interested and risk-neutral, should cooperate but in fact do not. Similarly, in the right pane, there are many observations below the line – these are the subjects who, if self-interested and risk-neutral, should defect but do not (the white area in the right pane).

In sum, this goal function explains the behavior of 27 out of 96 subjects, that is, a "hit rate" of about 28%. All other goal functions nest *Homo oeconomicus* and add parameters. Hence they will do at least as well. The question is by how much.

4.2. Inequity aversion

An inequity averse decision-maker with preferences according to the model in Fehr and Schmidt (1999) has the following goal function:

$$U_i^{(FS)}(s) = \begin{cases} \pi_i(s) - \alpha \cdot \left[\pi_j(s) - \pi_i(s)\right] & \text{if } \pi_i(s) \leq \pi_j(s) \\ \pi_i(s) - \beta \cdot \left[\pi_i(s) - \pi_j(s)\right] & \text{if } \pi_i(s) > \pi_j(s) \end{cases} \tag{3}$$

for player roles $i = 1, 2$ and $j \neq i$, where α and β are nonnegative and $\alpha \geq \beta$. In other words, individuals (weakly) dislike inequity, and (weakly) more so when they are worse off than the other party.

In the second player's role, a decision-maker with such a goal function prefers to cooperate conditional on cooperation (i.e., use pure strategy CC or CD) if

$$30 \geq 50 - (50 - 5)\beta, \tag{4}$$

or equivalently, if $\beta \geq 4/9$. Likewise, the decision-maker prefers to defect conditional on defection (i.e. use pure strategy CD or DD) if

$$10 \geq 5 - (50 - 5)\alpha. \tag{5}$$

By hypothesis $\alpha \geq 0$, so all second movers with goal function $U_2^{(FS)}$ should defect in response to first-mover defection. Thus, when imposing the structural Fehr–Schmidt model, conditional cooperation as second mover is equivalent with $\beta \geq 4/9$; selfish second movers with $\beta = 0$ always defect.

According to $U_1^{(FS)}$, a first mover cooperates if

$$30 \cdot p(C|C) + [5 - (50 - 5)\alpha] \cdot p(D|C) \geq [50 - (50 - 5)\beta] \cdot p(C|D) + 10 \cdot p(D|D), \tag{6}$$

where $p(a_j|a_i)$ is the subject's elicited expectation about the second mover's choice a_j if the subject chooses a_i. This inequality highlights the importance of allowing for subjective expectations in other-regarding models. To see this, move the terms with α and β to the right-hand side to yield $45 \cdot [\alpha p(D|C) - \beta p(C|D)]$. Since $\alpha \geq \beta$ is assumed in the Inequity aversion model, this term is positive if $p(D|C) > p(C|D)$, an inequality that holds under correct beliefs. A model with correct beliefs about others' second-mover choices thus predicts that inequity averse individuals are less likely to cooperate as first movers than selfish individuals. This prediction goes against the empirical observations where unconditional defectors are less likely to cooperate as first movers than conditional cooperators (see Table 1 as well as Altmann et al. (2008) and Blanco et al. (2014)).

In our setting, we can illustrate how heterogeneity of beliefs reconciles the predictions of the theory with the empirical observations. To see this, rewrite the condition for the optimality of first-mover cooperation (6) more explicitly in terms of beliefs about second-mover cooperation:

$$p(C|C) \geq \frac{5 + 45\alpha}{25 + 45\alpha} + \frac{40 - 45\beta}{25 + 45\alpha} \cdot p(C|D). \tag{7}$$

Now any subjective beliefs consistent with this inequality are consistent with first-mover cooperation, and subjective beliefs which violate this inequality are consistent with first mover defection. As we will see, the differences in the distribution of beliefs between unconditional defectors and conditional cooperators differ remarkably. These differences reconcile the model predictions to the empirical pattern that conditional cooperators are more likely to cooperate as first-movers than unconditional defectors. In fact our model with subjective beliefs predicts that one half of the conditional cooperators and only one ninth of the unconditional defectors should cooperate as first movers, thus predicting a positive correlation between conditional and first-mover cooperation. Moreover, sixty out of eighty-one participants in these two behavioral classes also behave in the predicted manner as first-movers.

To show this and to evaluate the overall explanatory power of the inequity aversion model with subjective beliefs, we first need to estimate inequity-aversion parameters. In the estimations, we assume that all subjects who have chosen the same strategy profile $(s_1, s_2) \in S_1 \times S_2$ have the same preference parameters (but the parameters may differ between subjects who chose different strategy profiles). Since there are eight pure-strategy profiles in the sequential prisoners' dilemma, we have made equally many estimates of the parameter pair (α, β). Moreover, for each of these eight subject pools, we impose the theoretical restriction in Fehr and Schmidt (1999) that $\alpha \geq \beta \geq 0$. Hence, each subject has a (weak) dislike of inequity, and (weakly) more strongly so when the inequity is disadvantageous than advantageous.

At first, let us consider the choices of the participants who as second movers act as conditional cooperators (match their action with that of the first mover). There are 36 such participants in our experiment. Recall that these are participants whose choices cannot be explained by the self-interest model of the previous subsection. In Fig. 3, such participants' beliefs regarding others' cooperation in the second-mover role are depicted in the same manner as the beliefs of the unconditional defectors in the previous subsection. In this diagram, we have set $\alpha = \beta = 4/9$ in both panels.

Let us apply the same approach as we did for the *Homo oeconomicus* model. Points in the closed half-plane defined by (7) are consistent with first mover cooperation. The points in the opposite closed half-plane are consistent with first-mover defection. By adjusting the parameters α and β, we can influence the goodness of fit of the inequity aversion model. In general, a higher β lowers the slope of the line that separates these half-planes. Therefore a higher β increases the range of

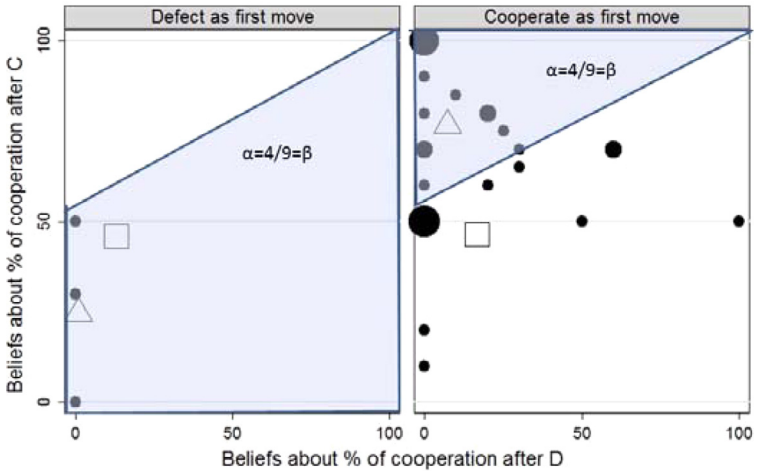

Fig. 3. Conditional cooperators. The explanatory power of the *Inequity aversion* model.

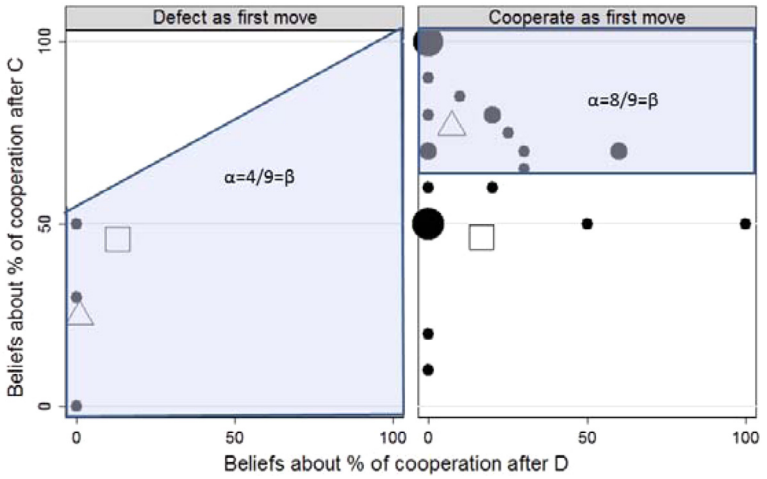

Fig. 4. Conditional cooperators. The explanatory power of the *Inequity aversion* model.

beliefs that are consistent with first-mover cooperation. A higher α shifts the intercept upwards and turns the slope flatter. Thus the effect of α is ambiguous.

Clearly and as shown above, if we impose $\alpha \geq 4/9 = \beta$, then all the second-move choices of the conditional cooperators are consistent with the Fehr–Schmidt model. What about the first-mover choices? It is easy to check that when $\alpha = 4/9 = \beta$, then the intercept of the black line in Fig. 3 lies at 5/9 (56%) and the slope is 4/9. Since all the three dots in the left pane lie below this line (see Fig. 3), all the defective first-mover choices by conditional cooperators (3 observations out of 3) are consistent with the Fehr–Schmidt model parameters $\alpha = 4/9 = \beta$. On the other hand, 14 of the 33 observations in the right-hand pane of figure are consistent with these parameter values of the Fehr–Schmidt model.

Since setting a lower value of α or β than 4/9 would imply that the second-mover choices of all the 36 conditional cooperators would become inconsistent with the model, we cannot improve the fit of the Fehr–Schmidt model by lowering either of the parameters. The question that remains is whether a higher value of α or β would allow increasing the fit of the model. If we start out from $\alpha = 4/9 = \beta$, increasing β alone is ruled out by the restriction that $\alpha \geq \beta$. Then again increasing α raises the intercept and lowers the slope in the right-hand-side of (7). Notice that a higher β also has a negative impact on the slope without affecting the intercept. Thus, we should in every case raise β to its maximal level where $\beta = \alpha$ in order to maximize the explanatory power. Among such parameter values (requiring $\alpha = \beta \geq 4/9$), we find that $\alpha = \beta = 8/9$ has the highest explanatory power for the conditional cooperators who also cooperate as the first move. This parameter pair yields the best fit in the right pane because (surprisingly) many subjects place a substantial probability for second-mover cooperation conditional on defection. The outcome with cooperation in response to defection results in a high material payoff for the first-mover. Thus, not choosing to collect this payoff can only be justified by an important aversion for advantageous inequality. With these parameter values, 21 out of 36 first-mover choices of the conditional cooperators can be explained (see Fig. 4).

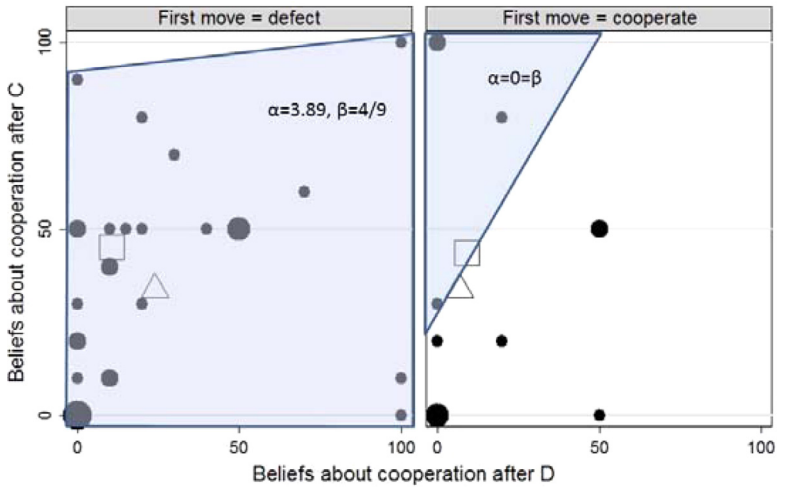

Fig. 5. Unconditional defectors. The explanatory power of the *Inequity aversion* model.

Let us next consider the unconditional defectors. According to (4), these reveal themselves having $\beta \leq 4/9$ (when inequity aversion is used as an identifying assumption). The second mover choices do not constrain α by any means. It turns out that all of the observations in the left pane of Fig. 2 can be explained if we impose $\alpha = 3.89$ (close to the maximal estimate in Fehr and Schmidt, 1999) and $\beta = 4/9$. Quite intuitively, a high α, indicating a strong aversion against disadvantageous inequality, makes first-mover cooperation highly unattractive, due to the risk of being exploited. This line is depicted in the left-hand pane of Fig. 5.[11] To maximize the number of observations explained in the right pane, one should instead set $\alpha = 0 = \beta$.

Let us finally briefly analyze the behavior of unconditional cooperators (who cooperate as second movers both in response to cooperation and in response to defection) and mismatchers (who defect as second movers in response to cooperation and cooperate in response to defection). Inequity aversion cannot explain any of these choices, since both of these types cooperate in response to defection. For this behavior to be optimal, we would need $10 \leq 5 - (50 - 5)\alpha$. This inequality is not satisfied for any feasible, i.e. non-negative, values of α and thus inequity aversion cannot explain the behavior of these types.

In sum, this two-dimensional class of goal functions–representing inequity aversion–is consistent with the behavior of 58 out of the 96 subjects' behavior, a "hit rate" of about 60%. The model can explain the behavior of all the 32 unconditional defectors who defect as the first move. The best-fitting parameter estimates for this behavioral profile are $\alpha = 3.89$ and $\beta = 4/9$. The model also explains the behavior 5 of the 13 unconditional defectors who cooperate as the first move (estimated parameters for this profile are $\alpha = 0 = \beta$), the behavior of all the 3 conditional cooperators who defect as the first move (estimated parameters: $\alpha = \beta = 4/9$), and the behavior of 18 of the 33 conditional cooperators who cooperate as a first move (estimated parameters: $\alpha = \beta = 8/9$). If the constraint $\alpha \geq \beta$ is removed, and only non-negativity of each parameter is required, then 79% of the observations are consistent with the model. Since cooperation in response to first-mover defection is inconsistent with this model (independently of what constraint on α is imposed), the behavior of unconditional cooperators and mismatchers cannot be explained by the model.

4.3. Conditional welfare

Suppose next that all individuals have a conditional concern for welfare, as expressed by the goal function in Charness and Rabin (2002). Applied to our setting, this goal function can be written in the form

$$U_i^{(CR)}(s) = \begin{cases} (1-\rho)\pi_i(s) + \rho\pi_j(s) & \text{if } \pi_i(s) \geq \pi_j(s) \\ (1-\sigma)\pi_i(s) + \sigma\pi_j(s) & \text{if } \pi_i(s) < \pi_j(s) \end{cases} \tag{8}$$

for player roles $i = 1, 2$ (and $j \neq i$) where ρ and σ are non-negative parameters such that $\sigma \leq 1/2$, $\rho \leq 1$, and $\sigma \leq \rho$. This goal function expresses a form of conditional altruism, whereby the weight placed on the other party's material outcome depends on who earns more.

It is easy to derive conditions for second-mover cooperation conditional on first-mover cooperation and defection. These are $\rho \geq 4/9$ and $\sigma \geq 1/9$, respectively. Thus unconditional defectors have $\rho \leq 4/9$ and $\sigma \leq 1/9$; conditional cooperators have

[11] Other parameter pairs can also explain the data points in the left-hand pane. However, the chosen parameter pair is the one among these that make the "hit area" as small as possible (see later discussion of this aspect).

$\rho \geq 4/9$ and $\sigma \leq 1/9$; unconditional cooperators have $\rho \geq 4/9$ and $\sigma \geq 1/9$, and mismatchers have $\rho \leq 4/9$ and $\sigma \geq 1/9$. First-mover cooperation is optimal if beliefs and preference parameters satisfy:

$$30 \cdot p(C|C) + [5(1-\sigma) + 50\sigma] \cdot p(D|C) \geq [50(1-\rho) + 5\rho] \cdot p(C|D) + 10 \cdot p(D|D), \tag{9}$$

or equivalently

$$p(C|C) \geq \frac{5-45\sigma}{25-45\sigma} + \frac{40-45\rho}{25-45\sigma} \cdot p(C|D), \tag{10}$$

We now select parameter values for this utility model, values for each of the eight behavioral classes which we observe. These parameter values are selected so as to maximize the hit-rate of the utility model in each class, and, as a secondary objective (when hit rates are the same), to minimize the hit-area in each class (without reducing the hit-rate from its maximal level). Preference parameters are constrained by the bounds set by the second-mover choices. The analysis (see Appendix A) reveals that this two-dimensional class of goal functions—representing a conditional concern for welfare—explains the behavior of 79 of the 96 subjects, a hit rate of 82%. The model explains the behavior of 22 of the 32 unconditional defectors who defect as first mover (the best-fitting parameter estimates for this behavioral profile are $\sigma = \rho = 0$), the behavior of 12 of the 13 unconditional defectors who cooperate as first mover (with $\sigma = 1/9$ and $\rho = 4/9$), the behavior of 1 of the 3 conditional cooperators who defect as first mover (with $\sigma = 1/9$ and $\rho = 8/9$), the behavior of all conditional cooperators who cooperate as first mover (with $\sigma = 1/9$ and $\rho = 4/9$), the behavior of 1 of the 2 unconditional cooperators who defect as first mover (with $\sigma = 1/9$ and $\rho = 8/9$), the behavior of all unconditional cooperators who cooperate as first mover (with $\sigma = 19/90$ and $\rho = 1/2$) and the behavior of 4 of the 5 mismatchers who defect as first mover (with $\sigma = 1/9$ and $\rho = 4/9$).

4.4. Reciprocity

In a sequential prisoner's dilemma, some form of reciprocity may explain second-mover behavior. Second movers may become altruistic towards first movers who play C and spiteful against first movers who play D. A fully developed model of reciprocity would need to introduce a type space for players, and let second movers make Bayesian updating about first movers' types conditional on observed choice.[12] Charness and Rabin (2002) did propose a simple reciprocity model of this kind, whereby the positive concern for the material well-being of the other party falls if the latter "misbehaved". Technically, this means that the non-negativity constraint on the parameter σ in their conditional-welfare model is dropped if the first mover played D. The second mover's concern for a first mover who misbehaved is captured by $\xi\sigma$, where $\xi \leq 1$. Since a new parameter, ξ, is introduced, the Charness–Rabin Reciprocity model has three parameters. We here study the explanatory power of this version of the Charness–Rabin model, to be called the *Reciprocity* model. In the present context, we take "misbehavior" to mean first-mover defection. As indicated above, this means that second-mover choices, conditional on first-mover defection, does not reveal as much about the sigma-parameter than in the *Conditional welfare* model, since the inference is confounded by the negative reciprocity motivation captured by the parameter ξ. Therefore, second-mover choices limit σ less, and a greater number of first-mover choices are consistent with the *Reciprocity* model than with the *Conditional welfare* model. With our data, it turns out that precisely 1 more observation (the only unexplained observation is in the right hand pane of Fig. 9) is consistent with the *Reciprocity* model. Thus, its hit-rate equals 83%, instead of the 82% obtained by the *Conditional welfare* model.

4.5. Altruism

A standard utility function used to represent (unconditional) altruism (see e.g. Becker (1976)) is

$$U_i^{(B)}(s) = \pi_i(s) + \theta \cdot \pi_j(s) \tag{11}$$

for some $\theta \in (0, 1)$. In other words, individuals care positively about the material outcome for the other party. Evidently, this is equivalent with a concern for welfare, $U_i^{(B)}(s) = (1-\theta) \cdot \pi_i(s) + \theta \cdot [\pi_1(s) + \pi_2(s)]$. Hence, this model is nested by Charness and Rabin (2002) model, obtained from (8) by requiring $0 < \rho = \sigma < 1/2$.[13]

A decision maker with such a utility function prefers to defect conditional on a first mover's defection if $10 + 10\theta \geq 5 + 50\theta$, or equivalently if $\theta \leq 1/8$. Similarly, the decision maker prefers to defect conditional on a first mover's cooperation if $30 + 30\theta \leq 50 + 5\theta$, or equivalently if $\theta \leq 4/5$.

An altruistic first mover prefers to cooperate if

$$(30 + 30\theta) \cdot p(C|C) + (5 + 50\theta) \cdot p(D|C) \geq (50 + 5\theta) \cdot p(C|D) + (10 + 10\theta) \cdot p(D|D). \tag{12}$$

Substituting $p(D|D) = 1 - p(C|D)$ and $p(D|C) = 1 - p(C|C)$ yields

$$(30 + 30\theta) \cdot p(C|C) + (5 + 50\theta) \cdot (1 - p(C|C)) \geq (50 + 5\theta) \cdot p(C|D) + (10 + 10\theta) \cdot (1 - p(C|D)). \tag{13}$$

[12] For models of conditional and interdependent preferences, see also Levine (1998); Dufwenberg and Kirchsteiger (2004); Weibull (2004); Falk and Fischbacher (2006); Cox et al. (2008), and Gul and Pesendorfer (2010).

[13] Eq. (8) can likewise be re-written in the form of conditional altruism.

and thus we need

$$p(C|C) \geq \frac{1-8\theta}{5-4\theta} + \frac{8-\theta}{5-4\theta} \cdot p(C|D) \tag{14}$$

for first-mover cooperation by an altruist.

These predictions illustrate that for an altruist unconditional defection is consistent with $\theta \leq 1/8$, and unconditional cooperation is consistent with $\theta \geq 4/5$. Mismatching by altruists is consistent with $1/8 < \theta < 4/5$. Conditional cooperation is never consistent with altruism since cooperating in reaction to first-mover cooperation reveals that θ must exceed 4/5, which would clearly imply second-mover cooperation in reaction to first-mover defection, a contradiction. In Appendix B, we provide diagrams and show the hit-rate maximizing parameter values which, as a secondary objective, also minimize the Selten–Krischker score for the altruism model.

The analysis reveals that this one-dimensional class of goal functions–representing altruism–can explain the behavior of 42 out of the 96 subjects, a hit rate of about 44%. The model explains the behavior of 22 of the 32 unconditional defectors who defect as the first move (parameter estimate $\theta = 0$ for this behavioral profile), the behavior of 9 of the 13 unconditional defectors who cooperate as a first move (with $\theta = 1/8$), the behavior of all 7 unconditional cooperators who cooperate as a first move (with $\theta = 4/5$), and the behavior of 4 of the 5 mismatchers who defect as a first move (with $\theta = 1/8$). The behavior of the conditional cooperators is inconsistent with the model.

4.6. Homo moralis

Alger and Weibull (2013) define a class of utility functions, that they call *Homo moralis*, for symmetric interactions. A sequential prisoners' dilemma is asymmetric; the strategy sets differ between the two player roles. However, a subject in our experiment is equally likely to be in the first-mover as in the second-mover role. And such an interaction is symmetric.[14] In this setting, a behavior strategy for a player is a triplet $x = (x_1, x_2, x_3)$, a point in the unit cube $X = [0, 1]^3$, where x_1 is the probability of playing C when in the first-mover position, x_2 the probability of playing C when in the second-mover position after the opponent has played C, and x_3 the probability of playing C in the second-mover position after the opponent has played D.[15]

With an equal chance to be assigned the first- or second-mover position in the sequential prisoners' dilemma, the expected material payoff, $\pi(x, y)$, for any player who uses strategy $x \in X$ against an opponent who uses strategy $y \in X$ is

$$\pi(x,y) = \frac{1}{2}[(25y_2+5)x_1 + (40y_3+10)(1-x_1)] + \frac{1}{2}[(50-20x_2)y_1 + (10-5x_3)(1-y_1)]. \tag{15}$$

The first term on the right-hand side is the probability that the player moves first, multiplied by the expected material payoff when playing C with probability x_1 in this role, given that the other player reciprocates C with probability y_2 and reciprocates D with probability $1 - y_3$. The second term is the probability that the player has the second move, multiplied by the expected material payoff when reciprocating C with probability x_2 and reciprocating D with probability $1 - x_3$, given that the other player, who makes the first move, chooses C with probability y_1.

The *ex ante* utility of a *Homo moralis* with degree of morality $\kappa \in [0, 1]$ is defined as

$$U^{(AW)}(x,y) = (1-\kappa)\pi(x,y) + \kappa\pi(x,x), \tag{16}$$

which gives

$$\begin{aligned} U^{(AW)}(x,y) = {} & \frac{\kappa}{2} \cdot [35x_1 + 35x_3 + 5x_1x_2 - 35x_1x_3 + 20] \\ & + \frac{1-\kappa}{2} \cdot [(25y_2 - 40y_3 - 5)x_1 - 20y_1x_2 - 5(1-y_1)x_3 + 40(y_1+y_3) + 20] \end{aligned} \tag{17}$$

What will a *Homo moralis* do as first mover? As in the other models, we use the subject's elicited beliefs, so $y_2 = p(C|C)$ and $y_3 = p(C|D)$ **.** As is easily verified, it is optimal for a *Homo moralis* of degree of morality κ to play C as a first mover if and only if[16]

$$p(C|C) \geq \frac{1}{5} + \frac{8}{5} \cdot p(C|D) - \frac{x_2 + 7(1-x_3)}{5-5\kappa} \cdot \kappa. \tag{18}$$

We note that for $\kappa = 0$ it is the same condition as for *Homo oeconomicus*. For high enough κ, the condition is always met, implying that *Homo moralis* of sufficiently high degree of morality plays C as first mover.

[14] Formally, the interaction can be represented by a game tree in which "nature" makes a first move that allocate the player roles, with probability 1/2 for each role allocation, followed by two copies of the tree in Fig. 1, with player labels reversed in one of the copies.

[15] Hence, each of the eight pure-strategy profiles in the sequential prisoners' dilemma corresponds to each of the eight corners of the unit cube X.

[16] We obtain this and the following inequalities by taking the partial derivatives of $U^{(AW)}(x, y)$ with respect to each component of the vector x, see appendix.

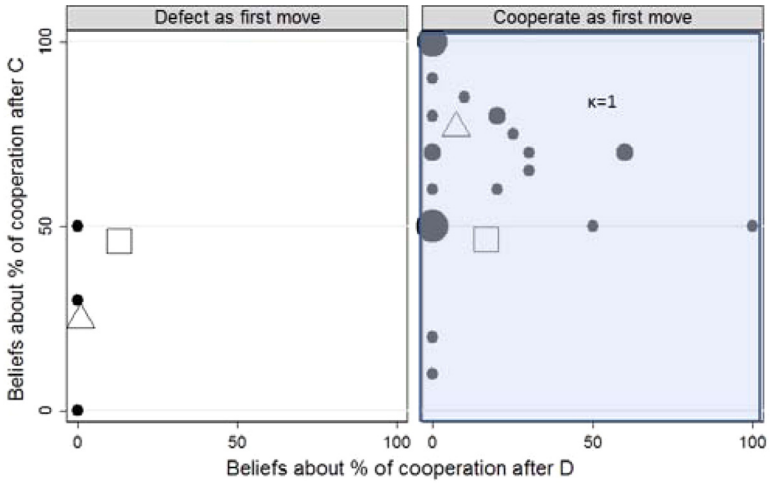

Fig. 6. Conditional cooperators. The explanatory power of the *Homo moralis* model.

In the second-mover role, the subject knows the first mover's pure strategy (which is either *C*, represented as $y_1 = 1$, or *D*, represented as $y_1 = 0$) and plays a best reply. For a *Homo moralis* of degree of morality κ, cooperation in reaction to cooperation (that is, to play $x_2 = 1$ when $y_1 = 1$) is optimal iff

$$\kappa x_1 \geq 4(1-\kappa). \tag{19}$$

Likewise, cooperation in reaction to defection (to play $x_3 = 1$ when $y_1 = 0$) is optimal iff

$$7(1-x_1)\kappa \geq 1-\kappa. \tag{20}$$

Let us first consider conditional cooperators who play *C* as first movers. In order to be optimal for a *Homo moralis* of degree of morality κ, this strategy, $x = (1, 1, 0)$, has to satisfy $\kappa \geq 4/5$, $0 \leq 1-\kappa$, and

$$p(C|C) \geq \frac{8}{5} \cdot p(C|D) - \frac{9\kappa - 1}{5(1-\kappa)}. \tag{21}$$

Clearly the latter inequality is met (strictly) for every $\kappa \geq 4/5$, so to play *C* as first mover and conditional cooperation as second mover is optimal for *Homo moralis* if and only if $\kappa \geq 4/5$, irrespective of beliefs about the opponent's second move. Conversely, it is never optimal to play *D* as a first move and conditional cooperation as one's second move—that is, use the strategy $x = (0, 1, 0)$—because then (19) requires $\kappa = 1$, (20) requires $\kappa \geq 1/8$ and the converse of (21) has to hold, which we just noted fails for all such κ -values. Hence, this strategy is never optimal. See Fig. 6.

We next study the unconditional defectors who defect as a first move (strategy $x = (0, 0, 0)$). Condition (19) is met for all κ, and the reverse of (20) is satisfied iff $\kappa \leq 1/8$. Condition (18) boils down to

$$p(C|C) \leq \frac{8}{5} p(C|D) + \frac{1-8\kappa}{5(1-\kappa)}. \tag{22}$$

Setting $\kappa = 0$ yields the maximal hit-rate, see the left pane in Fig. 7. For unconditional defectors who play C as their first move, shown in the right pane of Fig. 7, $\kappa = 0.42$ maximizes the hit rate.

Consider an unconditional cooperator who cooperates as a first move (strategy $x = (1, 1, 1)$). Condition (20) requires $\kappa = 1$ and from (19) we get that $\kappa \geq 4/5$ is required. Hence, $\kappa = 1$ and condition (18) is met, so that all observations in the right pane of Fig. 8 can be explained.

Consider then an unconditional cooperator who defects as a first move (strategy $x = (0, 1, 1)$). Condition (20) boils down to $\kappa \geq 1/8$, condition (19) requires $\kappa = 1$, and condition (18) fails, and thus none of the two observations in the left pane of Fig. 8 can be explained.

Finally, consider mismatchers, and start with those mismatchers who cooperate as a first move (strategy $x = (1, 0, 1)$). Condition (20) requires $\kappa = 1$ and (19) requires $\kappa \leq 4/5$. Hence, no data point in the right pane of Fig. 9 is explained. However, for mismatchers who defect as first movers (strategy $x = (0, 0, 1)$), (19) requires $\kappa = 1$, and (20) requires $\kappa \geq 1/8$. Hence, $\kappa = 1$ is necessary. Moreover, the separating straight line (from (18)) is

$$p(C|C) = \frac{1}{5} + \frac{8}{5} \cdot p(C|D). \tag{23}$$

We note that this is independent of κ and that it is the same as for *Homo oeconomicus*. Thus all the observations in the left pane of Fig. 9 can be explained.

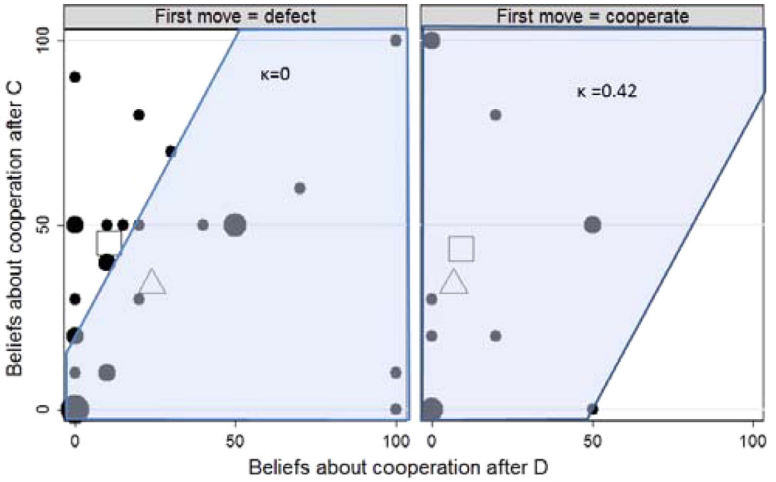

Fig. 7. Unconditional defectors. The explanatory power of the *Homo moralis* model.

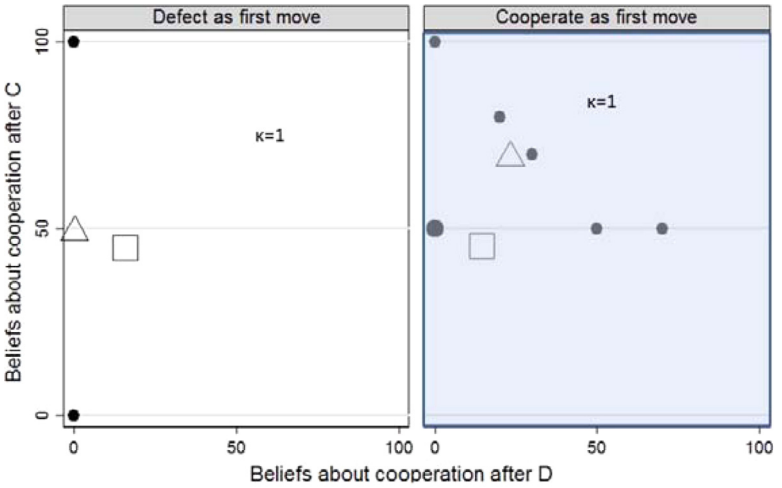

Fig. 8. Unconditional cooperators. The explanatory power of the *Homo moralis* model.

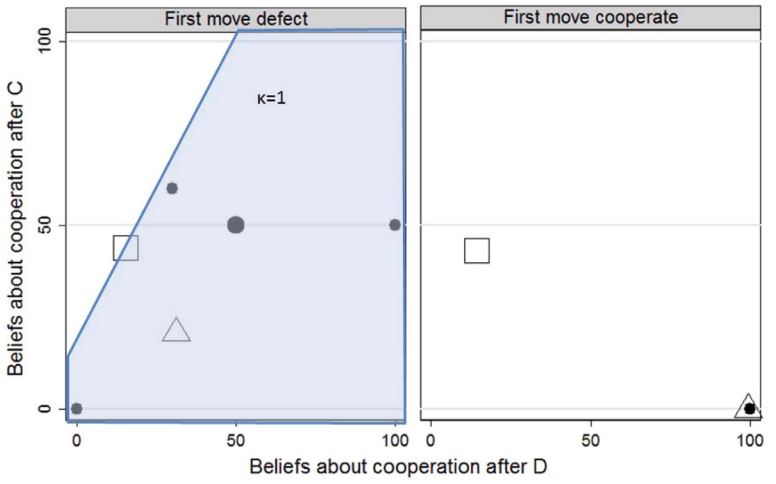

Fig. 9. Mismatchers. The explanatory power of the *Homo moralis* model.

Table 3
Analysis with 8 homogeneous groups.

Model	Hit rate	Hit area	# param.
Homo oeconomicus	0.28	0.12	0
Inequity aversion	0.60	0.28	2
Conditional welfare	0.82	0.44	2
Reciprocity	0.83	0.47	3
Altruism	0.44	0.39	1
Homo moralis	0.83	0.55	1

In sum, this one-dimensional class of goal functions—representing a mix of selfishness and Kantian morality—can explain the behavior of 80 out of the 96 subjects.

5. Model comparison

The "raw" explanatory power of the six models, expressed as their "hit rate", the percentage of the subject pool whose behavior can be explained, are given in the first column of the Table 3. These hit rates suggest that *Reciprocity, Conditional welfare,* and *Homo moralis* are the most capable of explaining our data.[17] These hit rates should of course be taken with a grain of salt. We will here briefly discuss two concerns: each model's number of free parameters, and "hit area", respectively.

First, the hit rates are not compensated for the number of parameters in the preference model at hand, neither for the parameter constraints that some models contain. The more parameters, the easier it is for a model to "hit" a data target, but the more constraints one imposes on the parameter values the harder it is. The number of parameters are given in the fourth column in the table. The *Conditional welfare* model, has the same number of parameters as the *Inequity aversion* model. If all parameter constraints of the*Inequity* aversion and the *Conditional welfare* model were relaxed, then these two models would obtain exactly the same hit rate. Indeed, the two models would then be mathematically identical.[18] We also note that while the *Homo oeconomicus* model has no free parameter, the *Altruism* and *Homo moralis* models have only one each. The *Reciprocity* model has the largest number of parameters. Hence, the playing field is not "horizontal" in terms of numbers of parameters and parameter constraints. For a"fair" comparison, the models should be given some "handicap" depending on their numbers of free parameters, also adjusted for constraints upon these. This is usually done by means of so-called BIC and AIC scores, a topic to which we will turn in the next section.

Secondly, the hit rates are not compensated for the models' "hit areas", that is, the size of the area in each panel that is consistent with the model in question. This point was made by Selten and Krischker (1983), and their compensation rule was axiomatized in Selten (1991). The compensation rule is to subtract the "hit area" from the "hit rate", where, generally speaking, the hit area is the area of the set of outcomes that agree with the model in question, expressed as a share of the total area of all possible outcomes. To be more precise, let $r \in [0, 1]$ be the relative frequency of a model's successful predictions, and let $a \in [0, 1]$ be the model's relative area of predictions that are possible under the model. A measure of predictive power is a function m that maps pairs (r, a) to the unit interval. Selten (1991) proposes different axioms for such measures. One axiom ("rate monotonicity") is that if two models have the same hit area but one has a higher hit rate than the other, then the first has higher predictive power. Another axiom ("area monotonicity") is, likewise, that if two models have the same hit rate but one has a smaller hit area, then this has higher predictive power. A third axiom ("equivalence of trivial theories") is that any model that has unit hit rate and unit hit area is just as useless as a model that has zero hit rate and zero hit area. A fourth axiom ("aggregation") is that if a model is used in two experiments with different hit rates and hit areas, say (r_1, a_1) and (r_2, a_2), then the overall predictive power of the model should be a weighted arithmetic mean of $m(r_1, a_1)$ and $m(r_2, a_2)$.[19] These four axioms are clearly met by the simple difference measure, $m^*(r, a) \equiv r - a$, the difference between the hit rate and the hit area. Selten (1991, Theorem 2)) establishes that, up to positive affine transformations, the difference measure m^* is the only measure of predictive power that meets these four axioms. We note that the Selten-Krischker score m^* takes values in the interval $[-1, 1]$, where the highest score, 1, is given to models with unit hit rate on a hit area of size zero, the score zero is given to all models for which the hit rate equals the hit area, and the lowest score, -1, is given to models with zero hit rate on a hit area of unit size.

It is this measure, m^*, that we here apply. However, there is a snag with applying it directly to the numbers in Table 3. The problem is that the parameters used for establishing the numbers in that table were chosen so as to maximize each

[17] Notice, however, that without the inequality restrictions $\sigma \leq \rho \leq 1/2$, the *Conditional welfare* model would have a hit rate of 95% (91 of the 96 observations could be explained), and with an alternative specification of the beliefs in the *Homo moralis* model (consensus beliefs about first-mover behavior), also that model's hit rate would be 85%.

[18] With such parameter relaxation, 95% of the observations could be explained, instead of 60%.

[19] Formally, the axioms are: (A1) $m(r, a)$ is strictly increasing in r, (A2) $m(r, a)$ is strictly decreasing in a, (A3) $m(0, 0) = m(1, 1)$, and (A4) $m(\lambda r_1 + (1-\lambda)r_2, \lambda a_1 + (1-\lambda)a_2) = \lambda m(r_1, a_1) + (1-\lambda)m(r_2, a_2)$ for all $\lambda \in [0, 1]$.

Table 4
Analysis with 4 homogeneous groups.

Model	Hit rate	Selten–Krischker score
Homo oeconomicus	0.28	0.16
Inequity aversion	0.53	0.28
Conditional welfare	0.76	0.26
Reciprocity	0.76	0.26
Altruism	0.40	0.02
Homo moralis	0.70	0.33

model's hit rate, with only lexicographic consideration of its hit area.[20] For a fair comparison between models, one would need to instead optimize the parameters for each model so as to maximize the difference between its hit rate and hit area. However, such parameter optimization is a non-trivial integer programming task. Given the small amount of data, it would seem exaggerated to bring in such brute force here. We therefore instead treat each of the four second-move strategies (CC,CD,DC,DD) as one homogeneous subpopulation, thus analyzing the population as consisting of 4 rather than 8 homogeneous groups. This approach is motivated by the observation that, unlike first-mover choices, second-mover choices are clear manifestations of preferences, without the confounding effect of beliefs. This evidently reduces the hit rate of all models except *Homo oeconomicus* (which has no free parameter to be adapted).

To have 4 instead of 8 subpopulations has the significant methodological advantage of making maximization of the Selten–Krischker score equivalent with the maximization of the hit rate. The reason is that the hit area in the left pane in each of the figures above then equals the "miss area" in the associated right-hand pane, irrespective of parameter values. Hence, by maximizing the hit rate for each model under this stronger homogeneity constraint, we obtain the maximal Selten–Krischker score for each model, thus levelling the field for the model race. The results are shown in Table 4.

We note the somewhat lower hit rates for all parametrized models. In comparison with Table 2, the ranking is fairly stable but has changed slightly. The *Conditional welfare* and *Reciprocity* models are still winners in terms of hit rates, while *Homo moralis* has fallen somewhat behind. However, in terms of the Selten–Krischker score, *Homo moralis* comes out first, followed by *Inequity aversion*. Thus the latter model is thus a more powerful competitor against *Conditional welfare* and *Reciprocity* in terms of the Selten–Krischker score than in terms of raw hit rates.

6. A random-utility specification

We finally make a random-utility specification of each of the six utility functions, assuming that variation in individual behavior is partly driven by idiosyncratic variations in preferences. More specifically, we assume that each individual's true preferences are those given by a parametrized utility function—any one of the six varieties we study—plus a random term accounting for idiosyncratic taste differences not captured by the candidate utility functions. The random "noise" terms are assumed to be i.i.d. Gumbel (or doubly exponentially) distributed, and thus all choice probabilities are logistic functions of the utilities assigned to the choice alternatives by the candidate utility model to be tested.

More specifically, for a subject ω facing a choice between C and D, either as first-mover or second-mover, the subject's utility for each alternative i is taken to be of the form $u_i + \varepsilon_{i\omega}$, where u_i is the deterministic utility assigned to alternative $i = C, D$ by a candidate utility function. The second term, $\varepsilon_{i\omega}$, is an idiosyncratic noise term, taken to be i.i.d. Gumbel distributed, with mode zero and standard deviation $\sigma = \pi/(\tau\sqrt{6}) \approx 1.28/\tau$, for some $\tau > 0$. We will call τ the *precision* parameter. It follows that if a candidate utility model specifies expected utilities u_C and u_D to the two choice alternatives at hand, then the probability for a randomly drawn subject to choose C takes the logistic form

$$\Pr[X_\omega = C] = \frac{e^{\tau u_C}}{e^{\tau u_C} + e^{\tau u_D}} \tag{24}$$

where $X_\omega \in \{C, D\}$ is the choice of subject ω. Using this formula, we provide maximum-likelihood estimates, for each model, of all parameters jointly (including τ).

First, for each utility function we assume that the preference heterogeneity in the whole subject pool is fully captured by the random noise terms. Table 5 shows each utility model's ML-estimated parameters, along with the associated AIC and BIC scores. The higher the estimated precision τ, or the lower the AIC or BIC score, the "better" is the candidate utility model at predicting observed behavior. (See Appendix E for detailed calculations.)

Thus, when requiring that all subjects in the whole subject pool should be treated as identical except for i.i.d. idiosyncratic differences, the *Altruism* and *Homo moralis* do not explain more than the *Homo oeconomicus* model. This may seem surprising, given the ample amount of deviations from the predictions of the self-interest model. This illustrates that the self-interest model is useful when simple models are needed for aggregate prediction. Moreover, the parameter σ in the *Conditional welfare* and *Reciprocity* models adds no predictive value. The best performing model according to the precision

[20] As indicated earlier, we have chosen parameter values so as to maximize each model's hit rate, and, if multiple optimal parameter vectors exist, chosen a parameter vector that minimizes the hit area among these.

Table 5
Analysis with one homogeneous group.

Model	Precision	AIC	BIC	Parameter estimates
Homo oeconomicus	0.029	389	393	–
Inequity aversion	0.060	350	361	$\alpha = \beta = 0.40$
Conditional welfare	0.152	360	371	$\rho = 0.44$ and $\sigma = 0$
Reciprocity	0.084	362	377	$\rho = 0.44$, $\sigma = 0$ and $\xi = -191$
Altruism	0.029	391	398	$\theta = 0$
Homo moralis	0.029	391	398	$\kappa = 0$

Table 6
Analysis with 4 homogeneous groups.

Model	Average precision	BIC	AIC
Homo oeconomicus	0.090	316	316
Inequity aversion	0.088	263	234
Conditional welfare	0.315	208	179
Reciprocity	0.318	218	174
Altruism	0.099	323	309
Homo moralis	0.157	289	274

parameter τ is the *Conditional welfare* model, followed by the *Reciprocity* and *Inequity aversion* models. If one instead used the AIC or BIC scores, then the Inequity aversion comes out as winner, followed by Conditional welfare and Reciprocity.[21] Why are the θ, κ, and σ estimates (virtually) zero? The reason is that about 85% of the participants responded to first-mover defection by defection. To maximize the likelihood of such choices, players should place little or no weight on the other's material payoff—since placing weight on the other just reduces the likelihood of observing defection. Richer interactions than the simple sequential prisoners' dilemma would be needed in order to differentiate these models.

In sum; for economic analysis based on a representative agent—that is, allowing only for idiosyncratic i.i.d. taste variations in the population—the simple *Homo oeconomicus* model does not fare so badly in terms of aggregate predictive power. The best such models, as measured by the AIC and BIC scores obtained in this experiment, are the *Reciprocity* model and the *Inequity* aversion models.

The representative agent model does not tell us much about the potential variety of motivations behind behavior in the population. Rather, all variation is then by assumption unobservable. To better understand whether the variation in motivation could be more finely categorized and understood, we therefore also consider a random utility specification which allows for precisely as much heterogeneity as the hit-rate analysis reported in Table 4. Thus, we now assume that the preference heterogeneity within each of the four behavior groups is fully captured by the random terms $\varepsilon_{i\omega}$, but treat different behavior groups as different populations with potentially different parameter specifications (in each of the candidate models). The results are reported in Table 6, where the precision of each model is the population-weighted average of the precisions (τ) for each of the four behavior groups (see Appendix A3 for details).

Comparing these results with those obtained by way of the Selten–Krischker approach, under the same homogeneity assumption (Table 3), we note that now *Conditional welfare* and *Reciprocity* take the lead, followed by *Homo moralis* and *Inequity aversion*.

7. Belief biases

Section 3 showed that expectations about cooperation differ from actual observed cooperation rates and that the beliefs are systematically correlated with behavior. Unconditional defectors, for instance, had a significantly higher expectation about the cooperation rate conditional on defection than conditional cooperators. In this subsection, we further analyze the observed differences in beliefs. We here suggest a unified structural model that permits us to explain how a subject's beliefs may be influenced by his or her preferences. Again, these results should be taken with a grain of salt, not least because of the potential incentive-compatibility difficulties with the quadratic scoring belief elicitation (see footnote 8). The point we want to make here is simple. Namely, that while both optimism and false consensus have been suggested as explanations for first-mover cooperation in social dilemmas, such biases should ideally be derived from individual preferences. This is precisely what we do in our simple structural analysis. More high-powered analyses have to be left for future research.

To illustrate the calculations, consider a first mover's beliefs about the second mover's reaction to the first move. We hypothesize that i' s subjective belief about the second mover's cooperation rate, conditional on the first mover's action, is a linear combination of the true second-mover cooperation rate ($\bar{x}_2$ and $\bar{x}_3$, respectively), i's own second-mover behavior (x_2

[21] This parameter is only relevant when the first mover has defected. Having a more negative value of ξ increases the likelihood of observing defection in response to defection. Indeed, 79 of the 96 subjects do so. Thus, in a homogeneous population, the maximum likelihood estimate of ξ is very negative. Table A2 shows that the estimates for the reciprocity parameter are more plausible in a heterogeneous population.

and x_3, respectively), and i's potential payoff gain, as follows:

$$\hat{p}_i(C|C) = \delta \cdot \bar{x}_2 + \gamma \cdot x_2 + \omega \cdot \frac{u_i(C,C) - u_i(C,D)}{|u_i(C,C)| + |u_i(C,D)|}, \tag{25}$$

and

$$\hat{p}_i(C|D) = \delta \cdot \bar{x}_3 + \gamma \cdot x_3 + \omega \cdot \frac{u_i(D,C) - u_i(D,D)}{|u_i(D,C)| + |u_i(D,D)|}. \tag{26}$$

Here δ is the weight on the true cooperation rate, γ the weight on the subject's own second-mover behavior, and ω the weight on a normalized payoff-difference term. We place no restrictions on the sign or size of these three parameters. In the normalized payoff-difference term, $u_i(a_i, C)$ denotes i's own utility if the second mover reacts cooperatively to i's action a_i, and $u_i(a_i, D)$ is i's utility if the second mover instead defects. Thus, at a given decision node of the opponent, this term biases the first mover's beliefs in the direction of overestimating the probability that the opponent's action will yield a high payoff (here by cooperating rather than defecting). This optimism effect is stronger when the payoff difference is more salient in the sense of Bordalo et al. (2012, p. 1250, and 2013, p. 809). We note that the optimism effect after cooperation vanishes if $u_i(C,C) = u_i(C,D)$; then $\hat{p}_i(C|C) = \delta\bar{x}_2 + \gamma x_2$, and likewise after defection. We also note that the normalized payoff difference is invariant under linear re-scaling of utilities, but not under addition of a constant to all utilities.[22] We estimate these utilities parametrically according to the *Conditional welfare* model.[23]

In order to illustrate the numbers involved, let us briefly consider an unconditional defector i (that is, an individual who always defects as second mover). If such an individual i also defects as a first mover, then his or her estimated ρ_i-value is zero (see Appendix). Thus, for such an individual we have $u_i(C,C) = 30$ and $u_i(C,D) = 5$, indicating a preference for second-mover cooperation in response to cooperation. The optimism term in (25) accordingly becomes $\omega \cdot (25/35) \approx 0.71\omega$, and, since this term is positive, there is an upward-biasing effect of optimism on the predicted conditional cooperation rate. Similar calculations for all behavioral classes yield two prediction errors for each subject, defined as the square of the deviation between the subjective and objective cooperation rates, $\left[\hat{p}_i(C|C) - p_i(C|C)\right]^2$ and $\left[\hat{p}_i(C|D) - p_i(C|D)\right]^2$, respectively. Minimization of these prediction errors results in the following aggregate population estimates $\hat{\delta} \approx 0.53$, $\hat{\gamma} \approx 0.27$ and $\hat{\omega} \approx 0.19$. The first two coefficients are significant at the 1% level and the third at the 5% level (see Appendix).

8. Conclusion

We here compare the explanatory power of six established preference models and identify subjects' belief biases. We do this by way of a novel approach that is, arguably, intuitive, visually appealing and operational, applied to a simple two-stage game, a sequential prisoners' dilemma. The six preference models are *Homo oeconomicus, Altruism* (Becker, 1976), *Inequity aversion* (Fehr and Schmidt, 1999), *Conditional welfare* (Charness and Rabin, 2002), *Reciprocity* (Charness and Rabin, 2002), and*Homo moralis* (Alger and Weibull, 2013).

Using maximum likelihood techniques based on subjects' elicited beliefs about each others' behavior, we find that, in terms of "hit rates" (share of subjects whose behaviors are compatible), *Reciprocity* and *Conditional welfare* come out as winners, closely followed by *Homo moralis*. If account is taken also for the "hit area" (share of behavior space allowed for), measured by what we call the "Selten–Krischker score", *Homo moralis* takes the lead, the second place goes to *Inequity aversion*, closely followed by *Conditional welfare* and *Reciprocity*. When a random utility approach is adopted and account is taken for the number of parameters in each of the six preference models, measured by the BIC score, *Conditional welfare* and *Reciprocity* come out as the winners, closely followed by *Inequity Aversion*.

All of these results have to be taken with a big grain of salt, as explained above.[24] In certain one person or strategic decisions, such as the dictator game or the ultimatum game, underlying and not recognized risk-aversion can generate a bias favoring other-regarding models with an altruistic component (see Gauriot et al., 2018, and Cameron, 1999, respectively). This is because risk-aversion means decreasing marginal utility of money, and may thus induce behavior which may appear quite altruistic while in fact it is motivated by little or no altruism. In our horse race, the handicapped models without an altruistic component will then be *Homo oeconomicus* and *Homo moralis*. However, self-interested risk aversion may in our experiment also operate in the opposite direction, since risk aversion may reduce the likelihood to cooperate as first mover. To verify this, consider any parametric expected-utility function (such as CRRA or CARA). For the sake of illustration, assume rational expectations, so that the probability weights agree with the true probabilities of cooperation. The derivative of the difference between the two sides of inequality (1), with respect to the risk aversion parameter, is then negative for most of the feasible parameter values. Since self-interested risk-aversion alone can induce effects both favoring and undermining other-regarding preference models, and since there is no consensus in the literature of the effects of risk-aversion in other-regarding contexts (see Trautmann and Vieider (2012); Miettinen et al. (2020), and their references), the present analysis is based on the (arguably simplistic) assumption of risk neutrality.

[22] If a positive constant is added to all utilities, then the normalized payoff difference shrinks, thus diminishing the salience of the difference.

[23] We use this model since it is one of the "winners" and since it is well-known. Ideally, one would like to take the best-fitting model for each behavioral category. For each behavioral class, we take the estimated parameters of the model and calculate the payoffs for the relevant action profiles.

[24] For the sake of consistency, we here summarize the results when the same 4-class approach is adopted, irrespective of method.

We also find that the subjects' average belief about each others' average behaviors comes fairly close to the truth. However, individual subjects differ quite a lot in beliefs and their individual biases show some correlation with their own behavior. Using maximum-likelihood methods, we identify a "false consensus" bias (Ross et al., 1977; Blanco et al., 2014), whereby subjects believe others behave more like themselves than they actually do, and a certain degree of optimism (Weinstein, 1980; Hey, 1984), whereby subjects overestimate probabilities for favorable outcomes (as evaluated in terms of their own preferences). We find that in our subject pool the consensus bias is about one and a half times stronger than the optimism bias.

All six preference models are more complex than the basic *Homo oeconomicus* model. Hence, their explanatory power needs to be traded off against their simplicity and versatility. Yet, even the simple sequential prisoners' dilemma that we use exemplifies that the gains in explanatory power by increasing the model complexity slightly may be considerable. On our experimental data, one can approximately triple the explanatory power by adding just one or two free parameters. We have no data on the "psychological realism" of the six preference models, but, arguably, each has a fairly clear intuitive and distinctive psychological appeal. Self-interest, altruism, inequity aversion, reciprocity, a concern for welfare and morality, all appear often in people's stated motivations for the actions they take in life.

There are, of course, many caveats to our analysis and results. In addition to the simplifying risk-aversion assumption, discussed above, two other limitations are particularly important. First, we make strong simplifying error specifications when calculating handicaps for the number of parameters (see Appendix). Second, our data set is quite limited in that each subject plays a single game (protocol) only once.

In future research, it would thus be interesting to collect more comprehensive data sets where each participant participates in several strategic interactions (as in Blanco et al., 2011; Boschini et al., 2013; Dreber et al., 2014). This would allow to test the accuracy of out-of-sample predictions based on preference parameters estimated from a sample of the observations. Another important avenue would be to collect more data on subjective beliefs about others' behaviors, and potentially even others' "types", so that one can estimate various reciprocity models and models of interdependent preferences (Levine, 1998; Dufwenberg and Kirchsteiger, 2004; Weibull, 2004; Falk and Fischbacher, 2006; Cox et al., 2008; Gul and Pesendorfer, 2010), or even models of concerns for social image or self-image (Bénabou and Tirole, 2006; Ellingsen et al., 2012; Malmendier et al., 2014).

Declaration of Competing Interest

We have now conflicting interests.

Appendix A. Hit rates for the Conditional welfare model

In the following diagrams (Figure A1, A2, A3 and A4), the straight lines indicate the combinations of beliefs about second-mover behavior for which the first mover is indifferent between cooperation and defection given the preference parameter values that allow to explain the highest number of observations. As in the inequity aversion model, there are two parameters of whom only one is relevant at a time—in each of the second-mover decision nodes. From the first-mover perspective, parameter ρ (σ) is relevant conditional on first-mover defection (cooperation), since the second mover will then earn either the same amount as the first mover or less (more). A higher ρ and σ renders first-mover cooperation more likely. Thus, in all of the left panes below, the hit rate is maximized by choosing a minimal ρ and σ. Likewise, in all of the right panes, the

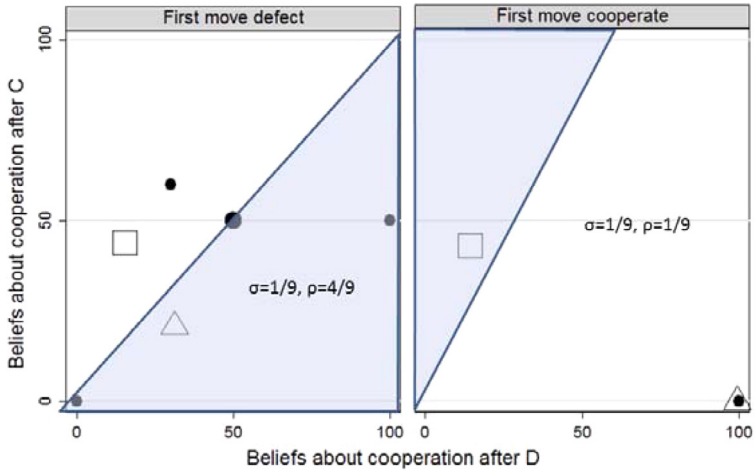

Fig. A1. Mismatchers. The explanatory power of the *Conditional welfare* model.

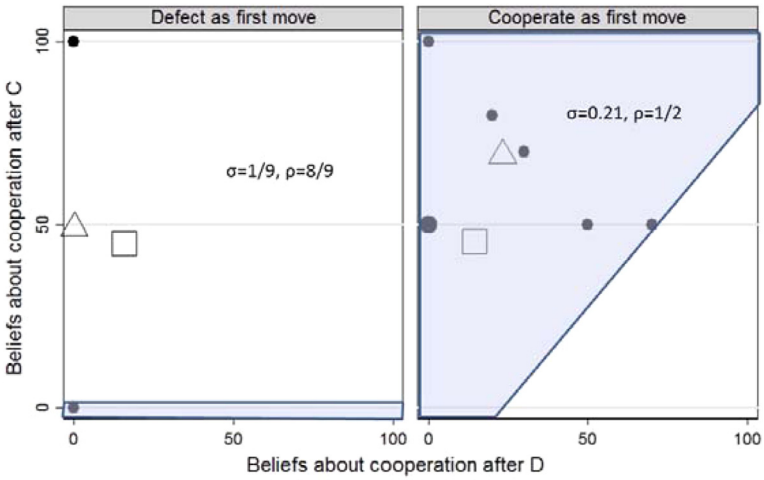

Fig. A2. Unconditional cooperators. The explanatory power of the *Conditional welfare* model.

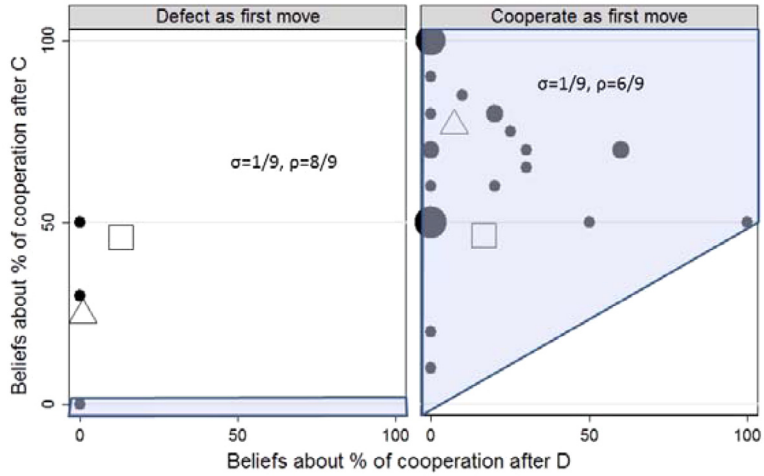

Fig. A3. Conditional cooperators. The explanatory power of the *Conditional welfare* model.

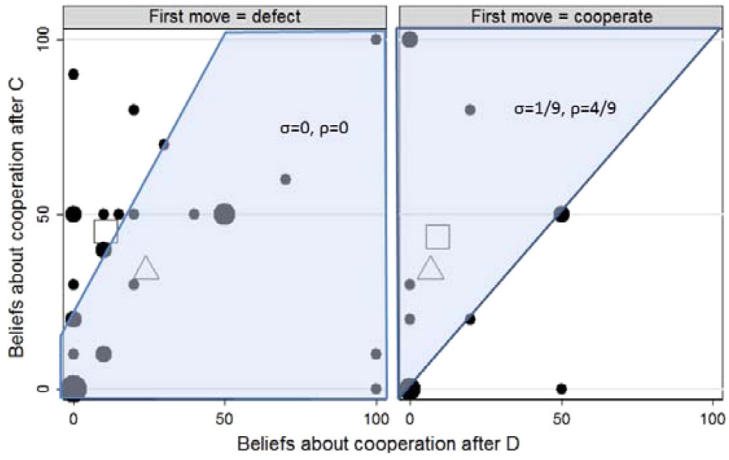

Fig. A4. Unconditional defectors. The explanatory power of the *Conditional welfare* model.

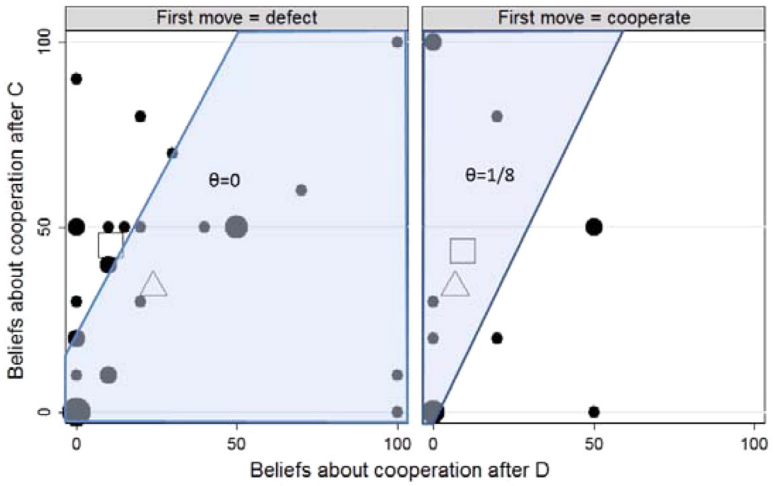

Fig. B1. Unconditional defectors. The predictive power of the altruism model.

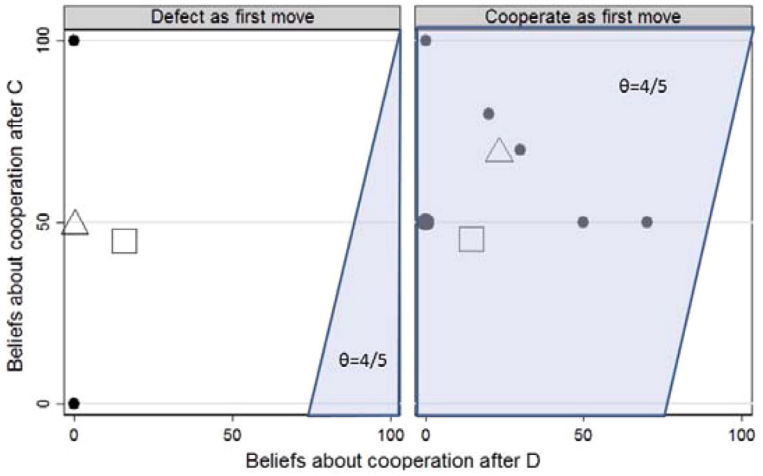

Fig. B2. Unconditional cooperators. The explanatory power of the *Altruism* model.

hit rate is maximized by choosing a maximal ρ and σ.[25] The values are yet constrained by what the second-mover choices reveal about the preferences.

Appendix B. Hit rates for the altruism model

The diagrams Fig. B1 show the hit-rate maximizing parameter values, which, as a secondary objective, also minimize the Selten-Krischker score for the altruism model.

To understand the first-mover choices of the unconditional defectors, their altruism parameter must satisfy $\theta \leq 1/8$. For the unconditional defectors who also defect in player role 1, the hit-rate maximizing θ-value induces a high intercept and slope. By choosing $\theta = 0$ the intercept equals 1/5 and the slope is 8/5, in which case 22 of the 32 observations in the left pane of Figure 10 can be accounted for. For the unconditional defectors who cooperate as a first move, a hit-rate maximizing θ-value induces a small intercept and small slope. With $\theta = 1/8$ the intercept equals zero and the slope equals $(63/8)\cdot(2/9) = 63/36$ so that 9 of the 13 observations in the right pane of Fig. B1 can be accounted for.

The unconditional cooperators have $\theta \geq 4/5$. All of the first-mover cooperators' choices in the right pane of Fig. B2, can be explained if we choose $\theta = 4/5$. Yet, for unconditional cooperators who defect as a first mover, choosing $\theta = 4/5$ maximizes the prospects of explaining the observations. However, the implied intercept of -3 and slope of 4 cannot explain

[25] In the figures we have maximized the hit-rate and lexicographically minimized the hit area, the area consistent with the model (conditional on not lowering the hit-rate). This obviously generates pressure to adjust the parameters in the opposite direction. See the end of section 4.8 for a "fair" comparison of the models, one that takes into account each model's hit area.

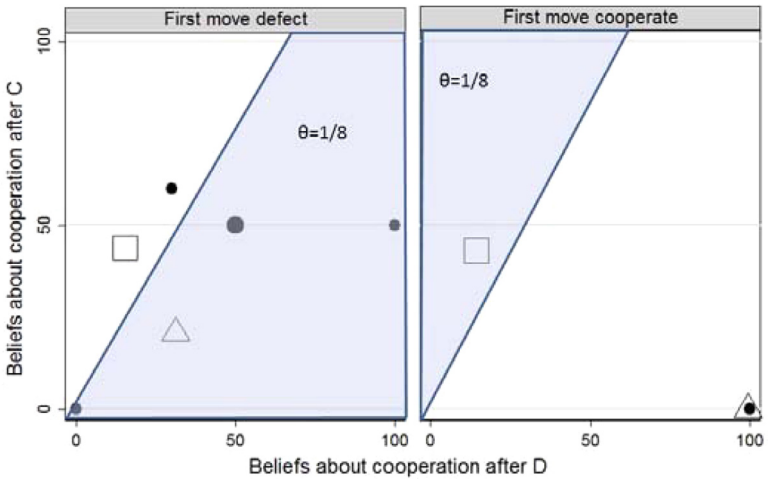

Fig. B3. Mismatchers. The explanatory power of the *Altruism* model.

either of the 2 observations in the left pane of Fig. B2. For the mismatchers with $\theta = 1/8$ the intercept equals zero and the slope equals $(63/8)\cdot(2/9) = 63/36$, so that all but one of the observations among the mismatchers who defect in the first-mover role can be accounted for (Fig. B3). There is no parameter value that allows explaining the unique mismatcher who cooperates as the first move.

Appendix C. Homo moralis decisions

Partial derivatives of the *Homo moralis* utility function with respect to each of the components of own strategy x:

$$\frac{\partial U^{(AW)}(x,y)}{\partial x_1} = \frac{5(5y_2 - 8y_3 - 1)(1-\kappa) + 5x_2\kappa + 35(1-x_3)\kappa}{2},$$

$$\frac{\partial U^{(AW)}(x,y)}{\partial x_2} = \frac{-20y_1(1-\kappa) + 5x_1\kappa}{2},$$

$$\frac{\partial U^{(AW)}(x,y)}{\partial x_3} = \frac{-5(1-y_1)(1-\kappa) + 35(1-x_1)\kappa}{2}. \tag{27}$$

In particular: the partial derivative with respect to x_1 is nonnegative if and only if

$$(1-\kappa)(5y_2 - 8y_3 - 1) + \kappa x_2 + 7\kappa(1-x_3) \geq 0 \tag{28}$$

which gives (18).

Appendix D. Estimates of belief biases

The obtained least-square estimates of the three coefficients δ (true rate), γ (consensus), and ω (optimism) in (25) and (26) are given in Table D1.

Table D1
Least-square estimates of the belief parameters.

VARIABLES	(1) no constant exp.coop.rate	(2) no constant and $\delta+\gamma+\omega=1$ exp.coop.rate	(3) non-constrained exp.coop.rate
truerate	0.530***	0.551***	0.554***
	(1.149)	(0.0962)	(0.151)
consensus	0.270***	0.269***	0.285***
	0.0479	0.0476	0.0499
optimism	1.188**	0.181**	0.317**
Constant			−8.469
			(7.714)
Observations	192	192	192
R-squared	0.656		0.307

Standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.01$.

Appendix E. Random-utility specification

We recall that a random variable X is *Gumbel* distributed with parameters $\tau > 0$ and $\nu \in \mathbb{R}$, or *Gumbel*(τ, ν), if its c.d.f. Φ is of the form

$$\Phi(x) = e^{-e^{-\tau(x-\nu)}} \qquad \forall x \in \mathbb{R}. \tag{29}$$

If $X_1, \ldots, X_n$ are statistically independent *Gumbel*(τ, ν_i) random variables, then the induced choice probabilities are

$$q_i = \frac{e^{\tau u_i}}{\sum_{j=1}^{n} e^{\tau u_j}} \qquad \text{for } i = 1, \ldots, n. \tag{30}$$

This is the versatile and much used *multinomial logit model* (McFadden, 1974). Assuming that all noise terms are i.i.d. Gumbel with mode $\nu = 0$ and model-specific precision $\tau > 0$, we obtain choice probabilities

$$q_C = \frac{e^{\tau u_C}}{e^{\tau u_C} + e^{\tau u_D}} = \frac{1}{1 + e^{-\tau(u_C - u_D)}} \tag{31}$$

and $q_D = 1 - q_C$. Here u_C and u_D are the expected utilities (according to the model in question) associated with the two choices at hand. A model with a high precision estimate τ is thus a model with good predictive power (as $\tau \to +\infty$: $q_C \to 1$ iff $u_C > u_D$).

E1. Homo oeconomicus

As our bench-mark goal function we took the goal function to be π_i for each player role $i = 1, 2$. This goal function evidently dictates unconditional defection in player role 2. By contrast, if the other party has made a first move C, then the random-utility specification assigns probability

$$q_{C|C} = \frac{1}{1 + e^{-(30-50)\tau}} = \frac{1}{1 + e^{20\tau}} \tag{32}$$

to the second move C. Likewise, after a first move D, the probability for the second move C is

$$q_{C|D} = \frac{1}{1 + e^{5\tau}}. \tag{33}$$

In player role 1, it is optimal to cooperate, according to the deterministic part of the goal function, if and only if $u_C \geq u_D$, where

$$u_C = 30 \cdot p(C|C) + 5 \cdot p(D|C) \quad \text{and} \quad u_D = 50 \cdot p(C|D) + 10 \cdot p(D|D), \tag{34}$$

and $p(a_j|a_i) \in [0, 1]$ is the subject's (elicited) expectation about the second mover's action a_j if the subject takes action a_i. Likewise, in the random-utility specification, the probability for the first move C is

$$q_C = \frac{1}{1 + e^{-[25p(C|C) - 40p(C|D) - 5]\tau}} \tag{35}$$

Using the data for all three choices made by all 96 subjects and their stated beliefs about second-mover choices, we obtain the maximum-likelihood estimate $\hat{\tau} \approx 0.029$ for the Homo oeconomicus model. This random-utility model has one free parameter, and its AIC and BIC scores are approximately 389 and 393.

E2. Inequity aversion

In the second player's role, a decision-maker with such a utility function prefers C conditional on first-mover C if

$$30 \geq 50 - (50 - 5)\beta, \tag{36}$$

and prefers D conditional on first-mover D if

$$10 \geq 5 - (50 - 5)\alpha. \tag{37}$$

By hypothesis $\alpha \geq 0$, so all second movers defect in response to first-mover defection. In the random-utility specification of this model, we likewise obtain

$$q_{C|C} = \frac{1}{1 + e^{-(45\beta - 20)\tau}} \quad \text{and} \quad q_{D|D} = \frac{1}{1 + e^{-(45\alpha + 5)\tau}}. \tag{38}$$

According to $U_1^{(FS)}$, a first mover cooperates if $u_C \geq u_D$, where

$$u_C = 30 \cdot p(C|C) + [5 - (50 - 5)\alpha] \cdot p(D|C) \tag{39}$$

and

$$u_D = [50 - (50 - 5)\beta] \cdot p(C|D) + 10 \cdot p(D|D). \tag{40}$$

In the random-utility version of the model, a first mover cooperates with probability

$$q_C = \frac{1}{1 + e^{-[(25+45\alpha)\cdot p(C|C) - (40-45\beta)\cdot p(C|D) - 45\alpha - 5]\tau}}. \tag{41}$$

Using the data for all subjects, we obtain the following ML-estimated parameter values: $\hat{\alpha} \approx \hat{\beta} \approx 0.40$, and $\hat{\tau} \approx 0.060$. The equality between the α and β estimates is due to the imposed constraint $\alpha \geq \beta$.[26] The precision parameter value being about double the size of that for *Homo oeconomicus*, we conclude that the inequity aversion model gives a better fit to the data than *Homo oeconomicus*. Its AIC and BIC values are approximately 350 and 361. Hence, also they suggest the superiority of this model over *Homo oeconomicus*, despite the fact that the AIC and BIC scores do not account for the fact that we here have imposed constraints on the parameter values for the inequity aversion model.

E3. Conditional welfare

Suppose next that all individuals have a conditional concern for welfare, as expressed by the utility function (8) taken from Charness and Rabin (2002), where ρ and σ are non-negative parameters such that $\sigma \leq 1/2$, $\rho \leq 1$, and $\sigma \leq \rho$. Along the same lines as in the inequity-aversion model, it is easily verified that the random-utility specification of this model gives second-mover probabilities

$$q_{C|C} = \frac{1}{1 + e^{-(45\rho - 20)\tau}} \quad \text{and} \quad q_{D|D} = \frac{1}{1 + e^{-(5-45\sigma)\tau}}. \tag{42}$$

According to $U_1^{(CR)}$, first-mover cooperation is optimal if beliefs and preference parameters satisfy

$$p(C|C) \geq \frac{5 - 45\sigma}{25 - 45\sigma} + \frac{40 - 45\rho}{25 - 45\sigma} \cdot p(C|D), \tag{43}$$

Within the random-utility framework, the probability for first-mover cooperation is

$$q_C = \frac{1}{1 + e^{-[(25-45\sigma)\cdot p(C|C) - (40-45\rho)\cdot p(C|D) + 45\sigma - 5]\tau}} \tag{44}$$

We obtain the following ML-estimated parameter values: $\hat{\rho} \approx 0.44$, $\hat{\sigma} \approx 0$, and $\hat{\tau} \approx 0.151$. It appears that the parameter σ hits its non-negativity constraint, so this parameter could be deleted from the model, or else one could relax the model by allowing σ to take negative values. We note that the noise parameter $\hat{\lambda}$ is lower than that for the inequity aversion model. In this respect, the Conditional welfare model has the best fit of the three models considered so far. Its AIC and BIC values are approximately 360 and 371, that, is slightly worse than for the inequity-aversion model.

E4. Reciprocity

In the random-utility specification of this model, the choice probabilities are the same as in the Conditional welfare model, except for

$$q_{D|D} = \frac{1}{1 + e^{-(5-45\xi\sigma)\tau}}. \tag{45}$$

We obtain the following ML-estimated parameter values: $\hat{\rho} \approx 0.44$ (the same as for the Conditional welfare model), $\hat{\sigma} \approx 0$, $\hat{\xi} \approx -191$, and $\hat{\tau} \approx 0.08$ (the same as for the Conditional welfare model). Apparently, both the σ and ξ parameters have hit their non-negativity constraint, and thus could be dropped from the model (or else let free). The AIC and BIC values for this model are slightly higher than for the Conditional welfare model, 362 and 377, respectively. Clearly, dropping the parameters σ and ξ from the model would boost these scores.

E5. Altruism

A decision maker with our altruistic utility function prefers to defect conditional on a first mover's defection if $10 + 10\theta \geq 5 + 50\theta$. Similarly, the decision maker prefers to defect conditional on a first mover's cooperation if $30 + 30\theta \leq 50 + 5\theta$. In the random-utility specification, this amounts to choice probabilities

$$q_{D|D} = \frac{1}{1 + e^{-(5-40\theta)\tau}} \quad \text{and} \quad q_{D|C} = \frac{1}{1 + e^{-(20-25\theta)\tau}} \tag{46}$$

In the deterministic utility model, an altruistic first mover prefers to cooperate if

$$(30 + 30\theta) \cdot p(C|C) + (5 + 50\theta) \cdot p(D|C) \geq (50 + 5\theta) \cdot p(C|D) + (10 + 10\theta) \cdot p(D|D). \tag{47}$$

[26] Letting both parameters free, we obtained $\hat{\alpha} \approx 0.18$ and $\hat{\beta} \approx 0.44$. The precision level rose from 0.06 to approximately 0.10, and the AIC and BIC scores were slightly reduced.

Table E1
ML-estimates of the models' precision parameter.

model	CC group	CD group	DD group	DC group
Homo oeconomicus	$\tau = 0$	$\tau = 0$	$\tau = 0.19$	$\tau = 0.04$
Inequity aversion	$\tau = 0$	$\tau = 0.07$	$\tau = 0.13$	$\tau = 0.04$
Conditional welfare	$\tau = 0.09$	$\tau = 0.57$	$\tau = 0.19$	$\tau = 0.12$
Reciprocity	$\tau = 0.12$	$\tau = 0.30$	$\tau = 0.19$	$\tau = 0.08$
Altruism	$\tau = 0.07$	$\tau = 0$	$\tau = 0.19$	$\tau = 0.09$
Homo moralis	$\tau = 0.39$	$\tau = 0.06$	$\tau = 0.19$	$\tau = 0.20$

Table E2
ML-estimates of the models' other parameters.

model	CC group	CD group	DD group	DC group
HE	–	–	–	–
IA	$\alpha = 9.1,\ \beta = 8.9$	$\alpha = \beta = 1.1$	$\alpha = 0.29,\ \beta = 0$	$\alpha = \beta = 0$
E	$\rho = 1,\ \sigma = 0.5$	$\rho = 0.95,\ \sigma = 0$	$\rho = \sigma = 0$	$\rho = \sigma = 0.24$
R	$\begin{cases}\rho = 1, \sigma = 0.07\\ \xi = 1\end{cases}$	$\begin{cases}\rho = 0.94, \sigma = 0\\ \xi = -32\end{cases}$	$\begin{cases}\rho = 0.016, \sigma = 0\\ \xi = 0.84\end{cases}$	$\begin{cases}\rho = \sigma = 0.034\\ \xi = 1\end{cases}$
A	θ =1	$\theta = 0.57$	$\theta = 0$	$\theta = 0.32$
HM	$\kappa = 1$	$\kappa = 1$	$\kappa = 0$	$\kappa = 0.69$

Substituting $p(D|D) = 1 - p(C|D)$ and $p(D|C) = 1 - p(C|C)$, we obtain the following expression for the random-utility specification:

$$q_C = \frac{1}{1 + e^{-[(25-20\theta)\cdot p(C|C) - (40-5\theta)\cdot p(C|D) - 5 + 40\theta]\tau}} \tag{48}$$

We obtain the following ML-estimated parameter values: $\hat{\theta} \approx 0$ and $\hat{\tau} \approx 0.029$ (the same noise level as for *Homo oeconomicus*). Apparently, the altruism parameter θ has hit its non-negativity constraint and could thus be dropped from the model (or else be let free). The AIC and BIC values for this model are slightly higher than for *Homo oeconomicus* (because of the additional but empirically redundant parameter), 391 and 398, respectively. Hence, altruism, if assumed to be of the same strength for all subjects, does not add explanatory power over and beyond the basic *Homo oeconomicus* model.

E6. Homo moralis

The *ex ante* utility function of a *Homo moralis* with degree of morality $\kappa \in [0, 1]$ is given in equation (17). It follows that, conditional upon being the first mover, the expected utility difference between playing C ($x_1 = 1$) and D ($x_1 = 0$) is

$$u_C - u_D = (1 - \kappa) \cdot [25p(C|C) - 40p(C|D) - 5] + \kappa \cdot [35 + 5x_2 - 35x_3] \tag{49}$$

In the random-utility specification, the associated choice probability is thus

$$q_C = \frac{1}{1 + e^{-[(40-25p(C|C)+40p(C|D)+5x_2-35x_3)\kappa + 25p(C|C) - 5 - 40p(C|D)]\tau}}. \tag{50}$$

For an individual with *Homo moralis* preferences, the expected utility gain from cooperating, rather than defecting, as a reaction to cooperation (that is, playing $x_2 = 1$ instead of $x_2 = 0$) is

$$5\kappa x_1 - 20(1 - \kappa). \tag{51}$$

Likewise, the expected utility gain from cooperating rather than defecting in reaction to defection (playing $x_3 = 1$ instead of $x_3 = 0$) is

$$35(1 - x_1)\kappa - 5(1 - \kappa). \tag{52}$$

Hence, in the random-utility specification, we obtain

$$q_{C|C} = \frac{1}{1 + e^{-[5x_1\kappa - 20(1-\kappa)]\tau}} \quad \text{and} \quad q_{C|D} = \frac{1}{1 + e^{-[(40-35x_1)\kappa - 5]\tau}}. \tag{53}$$

We obtain the following ML-estimated parameter values: $\hat{\kappa} \approx 0$ and $\hat{\tau} \approx 0.029$ (the same noise level as for *Homo oeconomicus*). Apparently, like in the altruism model, the morality parameter κ has hit its non-negativity constraint and thus could be dropped from the model (or else be let free). The AIC and BIC values for this model are slightly higher than for *Homo oeconomicus*, 391 and 398, respectively, the same as for the altruism model. Hence, morality, if assumed to be of the same strength for all subjects, does not add explanatory power over and beyond the basic *Homo oeconomicus* model.

E7. Parameter estimates for the heterogeneous four-group model

For each of the six utility models, all parameters have been estimated. The maximum-likelihood estimates of all models' precision parameters, one for each of the four behavior groups, are given in Table E1. The associated maximum-likelihood estimates of all other parameters are given in Table E2.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jebo.2020.02.018.

References

Alger, I., Weibull, J., 2013. Homo moralis – preference evolution under incomplete information and assortativity. Econometrica 81, 2269–2302.
Alger, I., Weibull, J., 2016. Evolution and kantian morality. Games Econ. Behav. 98, 56–67.
Altmann, S., Dohmen, T., Wibral, M., 2008. Do the reciprocal trust less? Econ. Lett. 99, 454–457.
Andersen, S., Harrison, G.W., Lau, M.I., Rutström, E.E., 2010. Behavioral econometrics for psychologists. J. Econ. Psychol. 31, 553–576.
Becker, G., 1974. A theory of social interactions. J. Polit. Econ. 82, 1063–1093.
Becker, G., 1976. Altruism, egoism, and genetic fitness: economics and sociobiology. J. Econ. Lit. 14, 817–826.
Bellemare, C., Kroger, S., Soest, A.V., 2008. Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. Econometrica 76, 815–839.
Bénabou, R., Tirole, J., 2006. Incentives and pro-social behavior. Am. Econ. Rev. 96, 1652–1678.
Blanco, M., Engelmann, D., Koch, A., Normann, H., 2014. Preferences and beliefs in a sequential social dilemma: a within-subjects analysis. Games Econ. Behav. 87, 122–135.
Blanco, M., Engelmann, D., Koch, A.K., Normann, H.T., 2010. Belief elicitation in experiments: is there a hedging problem? Exp. Econ. 13 (4), 412–438.
Blanco, M., Engelmann, D., Normann, H., 2011. A within-subject analysis of other-regarding preferences. Games Econ. Behav. 72, 321–338.
Bordalo, P., Gennaioli, N., Shleifer, A., 2012. Salience theory of choice under risk. Q. J. Econ. 127, 1243–1285.
Bordalo, P., Gennaioli, N., Shleifer, A., 2013. Salience and consumer choice. J. Polit. Econ. 121, 803–843.
Boschini, A., Muren, A., Persson, M., 2013. The social egoist. In: WP 2013: 14, Department of Economics, Stockholm University.
Brandts, J., Charness, G., 2000. Hot vs. cold: sequential responses and preference stability in experimental games. Exp. Econ. 2, 227–238.
Brandts, J., Charness, G., 2011. The strategy versus the direct-response method: a first survey of experimental comparisons. Exp. Econ. 14, 375–398.
Bruhin, A., Fehr, E., Schunk, D., 2018. The many faces of human sociality: Uncovering the distribution and stability of social preferences. J. Eur. Econ. Assoc. doi:10.1093/jeea/jvy018. (In press)
Cameron, L.A., 1999. Raising the stakes in the ultimatum game: experimental evidence from Indonesia. Econ. Inq. 37, 47–59.
Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. Q. J. Econ. 117, 817–869.
Clark, K., Sefton, M., 2001. The sequential prisoner's dilemma: evidence on reciprocation. Econ. J. 111, 51–68.
Cox, J., Friedman, D., Sadiraj, V., 2008. Revealed altruism. Econometrica 76, 31–69.
Dillenberger, D., Postlewaite, A., Rozen, K., 2017. Optimism and pessimism with expected utility. J. Eur. Econ. Assoc. 15, 1158–1175.
Dreber, A., Fudenberg, D., Rand, D., 2014. Who cooperates in repeated games: the role of altruism, inequity aversion, and demographics. J. Econ. Behav. Org. 98, 41–55.
Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games Econ. Behav. 47, 268–298.
Ellingsen, T., Johannesson, M., Möllerström, J., Munkhammar, S., 2012. Social framing effects: preferences or beliefs? Games Econ. Behav. 76, 117–130.
Engelmann, J.B., Schmid, B., De Dreu, C.K.W., Chumbley, J., Fehr, E., 2019. On the psychology and economics of antisocial personality. PNAS 116, 12781–12786.
Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games Econ. Behav. 54, 293–315.
Fehr, E., Schmidt, K., 1999. A theory of fairness, competition, and cooperation. Q. J. Econ. 114, 817–868.
Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10, 171–178.
Fischbacher, U., Gächter, S., 2010. Social preferences, beliefs, and the dynamics of free riding in public goods experiments. Am. Econ. Rev. 100, 541–556.
Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Econ. Lett. 71, 397–404.
Fudenberg, D., Levine, D., 2012. Fairness, risk preferences and independence: impossibility theorems. J. Econ. Behav. Org. 81, 606–612.
Gächter, S., Nosenzo, D., Renner, E., Sefton, M., 2012. Who makes a good leader? Cooperativeness, optimism and leading-by-example. Econ. Inq. 50, 867–879.
Gauriot, R., Heger, S., Slonim, R., 2018. Altruism or diminishing marginal utility? In: IZA Institute of Labor Economics, Discussion Paper No. 11721.
Gul, F., Pesendorfer, W., 2010. Interdependent preference models as a theory of intentions. J. Econ. Theory 165, 179–208.
Hey, J.D., 1984. The economics of optimism and pessimism. Kyklos 37, 181–205.
Iriberri, N., Rey-Biel, P., 2011. The role of role uncertainty in modified dictator games. Exp. Econ. 14, 160–180.
Kant, I., 1964. Grundlegung zur metaphysik der sitten. In: English: Groundwork of the Metaphysics of Morals. Harper Torch books, New York. (1785)
Levine, D., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dyn. 1, 593–622.
Malmendier, U., de Velde, V., Weber, R., 2014. Rethinking reciprocity. Annu. Rev. Econ. 6, 849–874.
McFadden, D., 1974. The measurement of urban travel demand. J. Publ. Econ. 3, 303–328.
Miettinen, T., Ropponen, O., Sääskilahti, P., 2020. Prospect theory, fairness, and the escalation of conflict at negotiation impasses. Scand. J. Econ. doi:10.1111/sjoe.12384. (In press)
Muren, A., 2012. Optimistic behavior when a decision bias is costly: an experimental test. Econ. Inq. 50, 463–469.
Nosenzo, D., Tufano, F., 2017. The effect of voluntary participation on cooperation. J. Econ. Behav. Org. 142, 307–319.
Puri, M., Robinson, D., 2007. Optimism and economic choice. J. Financ. Econ. 86, 71–99.
Ross, L., Greene, D., House, P., 1977. The 'false consensus effect': an egocentric bias in social perception and attribution processes. J. Exp. Soc. Psychol. 13, 279–301.
Rubinstein, A., Salant, Y., 2016. Isn't everyone like me? On the presence of self-similarity in strategic interactions. Judg. Decis. Mak. 11, 168.
Schlag, K.H., Tremewan, J., Weele, J.J.V.d., 2015. A penny for your thoughts: a survey of methods for eliciting beliefs. Exp. Econ. 18, 457–490.
Selten, R., 1967. Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Beiträge zur Experimentellen Wirtschaftsforschung. JCB Mohr (Paul Siebeck), Tübingen, BRD.
Selten, R., 1991. Properties of a measure of predictive success. Math. Soc. Sci. 21, 153–167.
Selten, R., Krischker, S., 1983. Comparison of two theories for characteristic function experiments. In: Aspiration Levels in Bargaining and Economic Decision Making. Springer, Berlin, pp. 259–264.
Spinnewijn, J., 2015. Unemployed but optimistic: optimal insurance design with biased beliefs. J. Eur. Econ. Assoc. 13, 130–167.
Trautmann, S., Vieider, F., 2012. Social influences on risk attitudes: applications in economics. In: Roeser, S., et al. (Eds.), Handbook of Risk thezory. Springer, New York. Chapter 29
Trautmann, S.T., Kuilen, G., 2015. Belief elicitation: a horse race among truth serums. Econ. J. 125, 2116–2135.
Weibull, J., 2004. Testing game theory. In: Huck, S. (Ed.), Advances in Understanding Strategic Behaviour: Game Theory, Experiments, and Bounded Rationality – Essays in Honor of Werner Güth. Palgrave MacMillan. Chapter 6
Weinstein, N.D., 1980. Unrealistic optimism about future life events. J. Person. Soc. Psychol. 39, 806–820.