The international platform of Ludwig-Maximilians University's Center for Economic Studies and the Ifo Institute

CESifo Working Papers

www.cesifo.org/wp

Revealed Preferences in a Sequential Prisoners' Dilemma: A Horse-Race Between Five Utility Functions

Topi Miettinen
Michael Kosfeld
Ernst Fehr
Jörgen Weibull

CESIFO WORKING PAPER NO. 6358
CATEGORY 13: BEHAVIOURAL ECONOMICS
FEBRUARY 2017

An electronic version of the paper may be downloaded
- *from the SSRN website: www.SSRN.com*
- *from the RePEc website: www.RePEc.org*
- *from the CESifo website: www.CESifo-group.org/wp*

ISSN 2364-1428

CESifo
Center for Economic Studies & Ifo Institute

CESifo *Working Paper No. 6358*

Revealed Preferences in a Sequential Prisoners' Dilemma: A Horse-Race Between Five Utility Functions

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. We find that subjects' beliefs about others' choices are fairly accurate on average. Using the elicited beliefs, we compare the explanatory power of a few current models of social and moral preferences. The data show 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 observed behavior. We use the estimated preference parameters to identify biases in subjects' expectations. We find a consensus bias (whereby subjects believe others behave like themselves) and a certain optimism (whereby subjects overestimate probabilities for favorable outcomes), the former being about twice as strong as the second.

JEL-Codes: C720, C900, D030, D840.

Keywords: cooperation, prisoners' dilemma, other-regarding preferences, categorical imperative, consensus effect, optimism.

Topie Miettinen
Hanken School of Economics & HECER
Arkadiankatu 7
P.O. Box 479
Finland – 00101 Helsinki
topi.miettinen@hanken.fi

Michael Kosfeld
Goethe-University Frankfurt
Department of Economics
Grüneburgplatz 1
Germany – 60323 Frankfurt
kosfeld@econ.uni-frankfurt.de

Ernst Fehr
University of Zurich
Department of Economics
Blümlisalpstrasse 10
Switzerland – 8006 Zurich
ernst.fehr@econ.uzh.ch

Jörgen Weibull
Stockholm School of Economics
Department of Economics
PO Box 6501
Sweden – 11383 Stockholm
jorgen.weibull@hhs.se

January 30, 2017
We thank Anna Dreber-Almenberg, Magnus Johannesson, Astri Muren, Tuomas Nurminen and Rickard Sandberg for helpful comments. The first author gratefully acknowledges financial support from the Yrjö Jahnsson Foundation. 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).

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 Altmann et al. (2008), there is a tendency for within-subject positive correlation between first-mover cooperation and second-mover conditional cooperation. That paper points out that if beliefs about others' behavior had been correct, there would instead be a negative correlation according to established preference models (such as the inequity aversion model of Fehr-Schmidt, 1999). Follow-up papers have noted a more general tendency for inconsistency between within-subject behavioral correlations and the predictions of various preference models (see e.g. Blanco et al. 2011). There are also studies that have started to analyze belief biases as potential explanations for anomalous behavioral correlations. Gächter et al. (2012) and Blanco et al. (2014) suggest a role both for optimism (Weinstein, 1980) and for a consensus effect (Ross et al. 1977), respectively.[1]

We here report results from a simple laboratory experiment based on a sequential prisoner's dilemma (one player moves first and the other player observes the first move and then makes a move). 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. 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 main contribution is to use the subjects' stated beliefs about each others' behavior in a comparison of the explanatory power of five current models of social preferences. In all five models we assume risk neutrality. Our simplest model, *Homo economicus* –maximization of own expected payoff–explains about 28% of our observations. Unconditional *Altruism*, where a positive weight is placed on the other party's payoff, can explain about 44% of the observations. The Fehr-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) can explain about 60% of the observations. The fourth model was a version of the Charness and Rabin (2002) model of a concern for social efficiency and inequity conditioned upon the two parties' relative outcomes (just as the *Inequity aversion* model). The (slightly simplified) version tested here, which we call the *Social welfare* model, explains 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

[1] For studies of the role of optimism in economics, see Hey (1984), Puri and Robinson (2007), Bellemare et al. (2008), Gächter et al. (2012), Muren (2012), Spinnewijn (2015), and Dillenberger et al. (2016).

include reciprocity models, such as the more complex version of Charness and Rabin (2002), as well as Dufwenberg and Kirchsteiger (2004), Falk and Fishbacher (2006), and Cox et al. (2008). Due to their complexity, our data is insufficient to identify the relevant parameters of those models.[2]

In our "horse-race" there is a fifth and final model that is similar to the latter in that it is not purely consequentialistic, but this one is identifiable: the *Homo moralis* model of Alger and Weibull (2013), which attaches a positive weight to a version of Kant's categorical imperative, explains about 83% of the observations.[3] Hence, the *Social welfare* and *Homo moralis* models share the "first prize" in this horse-race between motivational models. We note, however, that while the *Social welfare* model has two free parameters, the *Homo moralis* model has only one.

Our second contribution is to shed light on subjective beliefs about others' behavior. We find that on average subjects' beliefs are fairly accurate. Yet, the stated beliefs differ in a consistent manner. The differences can be explained in terms of a consensus bias (belief that others act like oneself) and optimism (belief that more favorable outcomes are more likely). Indeed, our data suggests that the stated subjective beliefs about second-mover cooperation rates can be explained as a combination of three terms: the true rate (about 52% weight), a consensus bias (about 33%), and optimism (15%). In summary, the paper sheds light on the two main motivational factors behind strategic behavior; beliefs and preferences.

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, and Section 5 concludes.

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 Figure 1 below.

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

[2] The Levine (1998) model of conditional altruism and spite is difficult to identify with our data on similar grounds, and thus is not analyzed in this paper.

[3] Kant (1785): "Act only according to that maxim whereby you can, at the same time, will that it should become a universal law."

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.

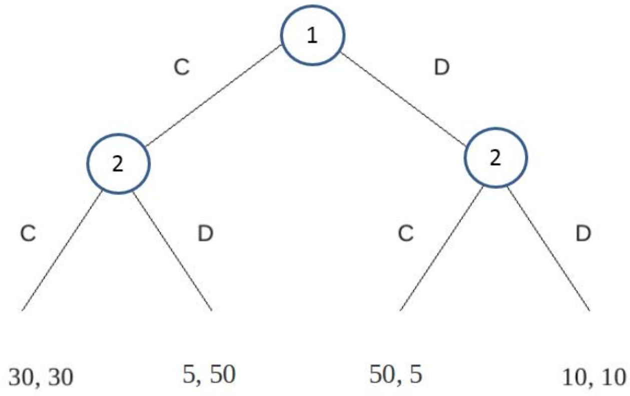

Figure 1: A sequential prisoners' dilemma.

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 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.[4]

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. The experimental sessions were run in 2003. 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).

[4]Notice that subjects' inputs to the belief questions can take any value between 0 and 100 indicating the likelihood, in percentage terms, that the opponent chooses C.

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, *mismatchers*, 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%, see Table 1 below, where CC indicates unconditional cooperation, CD conditional cooperation, DC mismatching, and DD unconditional defection.

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

	CC	**CD**	**DC**	**DD**	
C	7	33	1	13	56%
D	2	3	5	32	44%
	9%	38%	6%	47%	

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.

We also study the participants' *average beliefs* about the second-mover cooperation. 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 are 47% and 16%. There is thus some upward bias in participants' expectations about other participant's second-mover cooperation rates. There are important differences in beliefs across the four behavior classes, however (ordered according to prevalence in our subject pool):

> The 45 *unconditional defectors* (playing DD) expect on average 35% cooperation rate conditional on cooperation, and 21% cooperation rate conditional on defection,
>
> The 36 *conditional cooperators* (playing CD) expect on average 67% cooperation rate conditional on cooperation, and 12% cooperation rate conditional on defection,
>
> The 9 *unconditional cooperators* (playing CC) expect on average 61% cooperation rate conditional on cooperation, and 19% cooperation rate conditional on defection,

The 6 *mismatchers* (playing DC) expect on average 27% cooperation rate conditional on cooperation, and 39% cooperation rate conditional on defection

Thus, in all but one of these eight conditional beliefs there is a *false consensus effect* (Ross et al., 1977, Blanco et al., 2014), 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. This biased belief could instead be categorized as "optimism", in the sense of exaggerating the likely success of one's own behavior.[5]

Fourth, and finally, we examine statistically the cooperation rates expected by each of the two main behavior classes; unconditional defectors and conditional cooperators. The beliefs of the latter show a consensus effect while the beliefs of the unconditional defectors does not. Our results can be summarized as follows:

RESULT 1: *Unconditional defectors* expect a more cooperative reaction to defection than *conditional cooperators* do. The expected cooperation rate after defection is negatively correlated with the subject's own response to cooperation.[6]

RESULT 2: *Conditional cooperators* expect a significantly more cooperative reaction to cooperation than *unconditional defectors* do. The expected cooperation rate after cooperation is positively correlated with the subject's own response to cooperation.[7]

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 false 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

[5] More on this in Section 4.1.

[6] As for the findings, the null hypothesis that the cooperation rate anticipated by the unconditional defectors is significantly lower or equal to that anticipated by the conditional cooperators can be rejected at 5%-level (p=0.0323 with Mann-Whitney U-test with continuity correction; p=0.0538 with Student's t-test) and 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).

[7] 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), all at 1% level and by a large margin.

contrast, the beliefs of the unconditional defectors in our data are on average not consistent with consensus bias, as indicated by Result 1 above.

4 Other-regarding and moral preferences

We now turn to the main purpose of this study, namely, to use our data to see how well the observed behavior can be explained by certain other-regarding and moral preference models proposed in the literature. Moreover, once we have found the best fitting preference model for each subject, we use the corresponding parameter estimates in section 4.7 to estimate the strength of optimism and consensus bias in stated beliefs. This is the secondary 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 five parametric (partly nested) families of goal functions, thus covering pure self-interest, inequity aversion (Fehr and Schmidt, 1999), a conditional concern for efficiency (Charness and Rabin, 2002), (unconditional) altruism (Becker et al., 1974), 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.

In the sequential prisoners' dilemma in Figure 1, let $A_1 = \{C, D\}$ be the pure-strategy set of player role 1 (first mover), and $A_2 = \{CC, CD, DC, DD\}$ the pure-strategy set of player role 2 (second mover). Let $\pi_i(a_i, a_{-i})$ be the monetary payoff earned by a subject in player role $i = 1, 2$ when using pure strategy $a_i \in A_i$ against an opponent who uses strategy $a_{-i} \in A_{-i}$ (where $-i$ denotes player role $j \neq i$). For any goal function U that may guide the choices of a subject, the associated behavior is thus the solution(s) to the program

$$\max_{a_i \in A_i} \sum_{a_{-i} \in A_{-i}} p(a_{-i}) \cdot U_i(a_i, a_{-i}),$$

where $p(a_{-i})$ is the subject's belief, at the moment of decision-making, that the opponent will choose strategy $a_{-i} \in A_{-i}$. When a subject is in player role 1, we set the probability $p(a_{-i})$ according to his or her elicited beliefs. By contrast, when in player role 2, the subject will know, at the time of his or her decision, what choice the opponent (the first mover) has made, and hence we assign unit probability to that observed choice.

4.1 Homo economicus

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

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

where $\hat{\mu}(a_{-i}|a_i)$ is the subject's (elicited) expectation about the second mover's choice a_{-i} if the subject chooses 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 $\hat{\mu}(D|C) = 1 - \hat{\mu}(C|C)$ and $\hat{\mu}(D|D) = 1 - \hat{\mu}(C|D)$, we can rewrite (1) as

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

Figure 2, below, 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 economicus* That is, with beliefs in the shaded areas, a *Homo economicus* as first mover behaves optimally, given his or her beliefs, while subjects with beliefs in the white areas behave inconsistently with the *Homo economicus* model.

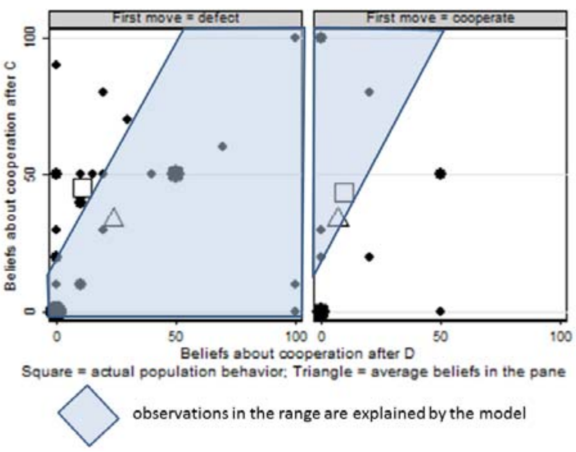

Figure 2: Unconditional defectors. The explanatory power of the *Homo economicus* model.

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. 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). Thus the darker area in each pane corresponds to the area where the observations are in line with the self-interest model and first-movers best-respond to their reported beliefs.

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 economicus* 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:

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

for $i = 1, 2$, 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 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 $\beta \geq 4/9$.

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

$$30 \cdot \hat{\mu}(C|C) + [5 - (50-5)\alpha] \cdot \hat{\mu}(D|C) \quad \geq \quad [50 - (50-5)\beta] \cdot \hat{\mu}(C|D) + 10 \cdot \hat{\mu}(D|D), \qquad (6)$$

where $\hat{\mu}(a_{-i}|a_i)$ is the subject's elicited expectation about the second mover's choice a_{-i} if the subject chooses a_i. Hence, an inequity averse first mover chooses C only if the second mover is expected to be much more likely to cooperate after C than after D. In fact a closer look at (6) reveals that, with correct beliefs about second mover behavior, *Homo economicus* is predicted to have a stronger tendency to cooperate as a first-mover than an inequity averse type. This is precisely the puzzle that was pointed out by Altmann et al. (2008) and further studied by Blanco et al. (2014).

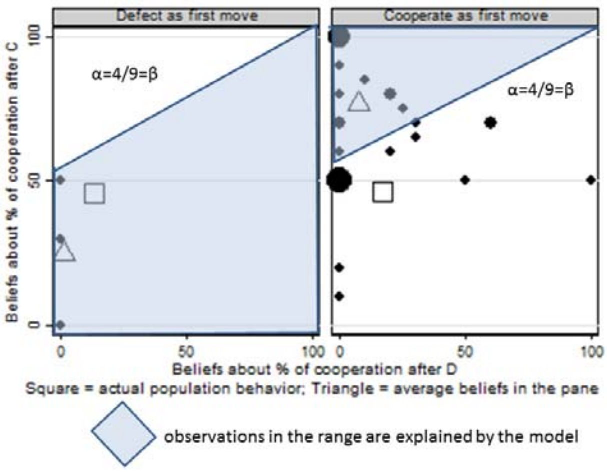

Figure 3: Conditional cooperators. The explanatory power of the inequity aversion model.

In what follows, we estimate best-fitting inequity-aversion parameters. In the estimations, we assume that all subjects with a given action profile $(a_1, a_2) \in \{C, D\} \times \{CC, CD, DC, DD\}$ have common preference parameter estimates, but the estimates may differ for two subjects with different action profiles. Moreover, we impose the theoretical restriction in Fehr and Schmidt (1999) that $\alpha \geq \beta \geq 0$ so that each agent is more averse to disadvantageous inequality than to advantageous inequality.

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 Figure 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.

Let us apply the same approach as we did for the *Homo economicus* model. We can write the condition for the optimality of first-mover cooperation, (6), in terms of the beliefs about second-mover cooperation as

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

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 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 Figure 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 Figure 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, 18 of the 33 observations in the right-hand pane of figure are consistent with these parameter values of the Fehr-Schmidt model (see Figure 3).

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). Thus, a higher α can allow for explaining more observations in the right pane of Figure 3, if its negative effect on the slope is sufficiently strong w.r.t. to the positive effect on the intercept. But 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. With these parameter values, 21 out of 36 first-mover choices of the conditional cooperators can be explained (see Figure 4).

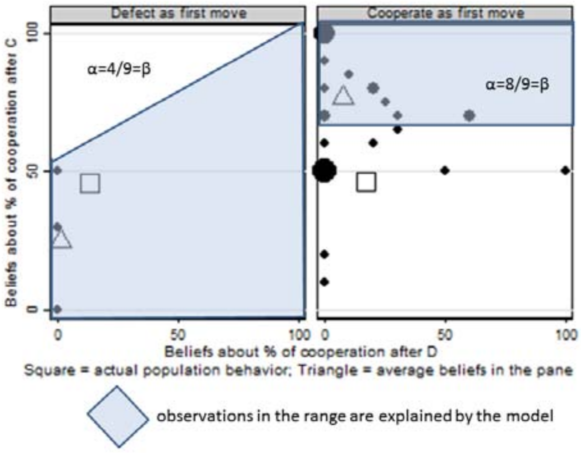

Figure 4: Conditional cooperators. The explanatory power of the *Inequity aversion* model.

Let us next consider the unconditional defectors. According to (4), these reveal themselves having $\beta < 4/9$ (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 Figure 2 can be explained if we impose $\alpha = 4$ (the maximal estimate in Fehr-Schmidt, 1999) and $\beta < 4/9$. This line is depicted in the left-hand pane of Figure 5 below. To maximize the number of observations explained in the right pane, one should set $\alpha = 0 = \beta$.

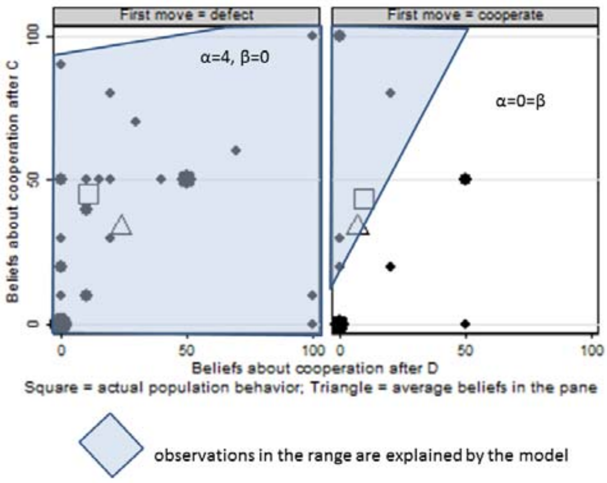

Figure 5: Unconditional defectors. The explanatory power of the *Inequity aversion* model.

Let us finally briefly analyse 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 $\alpha = 4$ and $\beta = 0$. 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: $\alpha = \beta = 8/9$). Since cooperation in response to first-mover defection is inconsistent with this model, the behavior of unconditional cooperators and mismatchers cannot be explained by the model.

4.3 Social welfare

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

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

for $i = 1, 2$ (and $j \neq i$) where ρ and σ are non-negative parameters parameters $\sigma \leq \rho < 1/2$. 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. An equivalent formulation, which more clearly shows the conditional concern for social welfare, is

$$U_i^{(CR)}(a_i, a_{-i}) = \begin{cases} \pi_i(a_i, a_{-i}) + \rho' \cdot [\pi_i(a_i, a_{-i}) + \pi_j(a_{-i}, a_i)] & \text{if } \pi_i(a_i, a_{-i}) \geq \pi_j(a_{-i}, a_i) \\ \pi_i(a_i, a_{-i}) + \sigma' \cdot [\pi_i(a_i, a_{-i}) + \pi_j(a_{-i}, a_i)] & \text{otherwise} \end{cases},$$

where $\rho' = \rho/(1-2\rho)$ and $\sigma' = \sigma/(1-2\sigma)$.

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 $\rho \geq 4/9$ 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:

$$30 \cdot \hat{\mu}(C|C) + [5(1-\sigma) + 50\sigma] \cdot \hat{\mu}(D|C)$$

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

or equivalently

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

In the following diagrams, the straight lines indicates 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 (under the assumption that the first-movers make no mistakes in their second-mover choices).

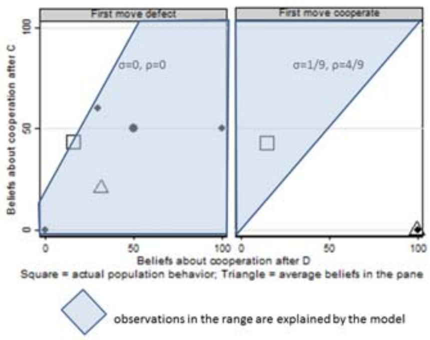

Figure 6: Mismatchers. The explanatory power of the *Efficiency* model.

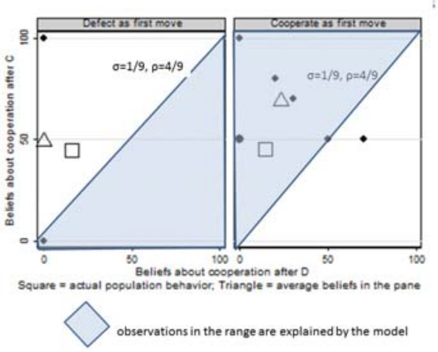

Figure 7: Unconditional cooperators. The explanatory power of the *Efficiency* model

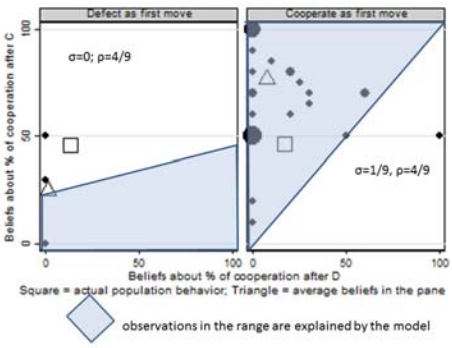

Figure 8: Conditional cooperators. The explanatory power of the *Efficiency* model.

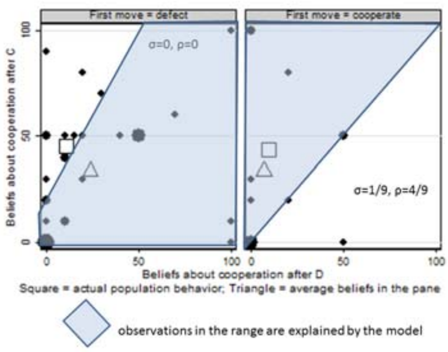

Figure 9: Unconditional defectors. The explanatory power of the *Efficiency* model.

In sum, this two-dimensional class of goal functions—representing a conditional concern for social efficiency—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 a first move (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 the first move (with $\sigma = 1/9$ and $\rho = 4/9$), the behavior of 1 of the 3 conditional cooperators who defect as a first move (with $\sigma = 0$ and

$\rho = 4/9$), the behavior of 32 of the 33 conditional cooperators who cooperate as the first move (with $\sigma = 1/9$ and $\rho = 4/9$), the behavior of 1 of the 2 unconditional cooperators who defect as the first move (with $\sigma = 1/9$ and $\rho = 4/9$), the behavior of 6 of the 7 unconditional cooperators who cooperate as the first move, and the behavior of all 5 mismatchers who defect as the first move (with $\sigma = \rho = 0$).[8]

4.4 Altruism

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

$$U_i^{BB}(a_i, a_{-i}) = \pi_i(a_i, a_{-i}) + \theta \cdot \pi_j(a_{-i}, a_i) \tag{11}$$

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, $U_i^{BB}(a_i, a_{-i}) = (1-\theta) \cdot \pi_i(a_i, a_{-i}) + \theta \cdot [\pi_1(a_{-i}, a_i) + \pi_2(a_{-i}, a_i)]$. Hence, this model is nested by the Charness and Rabin (2002) model, obtained from (8) by requiring $\rho = \sigma$.[9]

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

An altruistic first-mover prefers to cooperate if

$$\begin{aligned} & (30 + 30\theta) \cdot \hat{\mu}(C|C) + (5 + 50\theta) \cdot \hat{\mu}(D|C) \\ \geq \; & (50 + 5\theta) \cdot \hat{\mu}(C|D) + (10 + 10\theta) \cdot \hat{\mu}(D|D). \end{aligned}$$

Substituting $\hat{\mu}(D|D) = 1 - \hat{\mu}(C|D)$ and $\hat{\mu}(D|C) = 1 - \hat{\mu}(C|C)$ yields

$$\begin{aligned} & (30 + 30\theta) \cdot \hat{\mu}(C|C) + (5 + 50\theta) \cdot (1 - \hat{\mu}(C|C)) \\ \geq \; & (50 + 5\theta) \cdot \hat{\mu}(C|D) + (10 + 10\theta) \cdot (1 - \hat{\mu}(C|D)). \end{aligned}$$

and thus we need

$$\hat{\mu}(C|C) \geq \frac{1 - 8\theta}{5 - 4\theta} + \frac{8 - \theta}{5 - 4\theta} \cdot \hat{\mu}(C|D)$$

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 $1/8 \leq \theta \leq 4/5$. Conditional cooperation is never consistent with altruism.

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 maximum likelihood θ parameter is one that induces a high intercept and slope. By choosing

[8] The model cannot explain the behavior of the single mismatcher who cooperates as the first move.

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

$\theta = 0$ 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 θ inducing a small intercept and small slope coefficient maximize the likelihood. With $\theta = 1/8$ the intercept equals zero and the slope equals $(7/8) \cdot (2/9) = 7/36$ so that all but one of the observations in the right pane of Figure 10, below, can be accounted for.

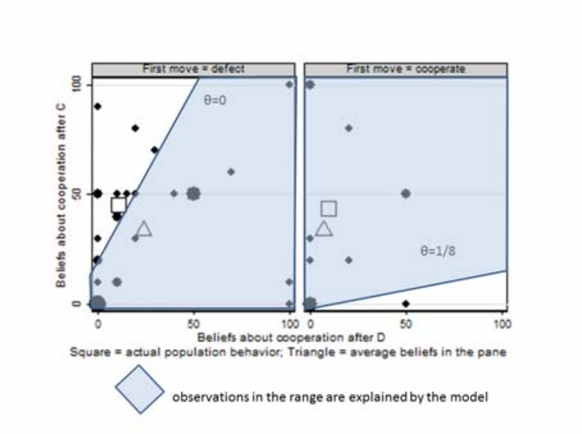

Figure 10: Unconditional defectors. The explanatory power of the *Altruism* model.

The unconditional cooperators have $\theta \geq 4/5$. All of the first-mover cooperators' choices in the right pane of Figure 11, below, can be explained if we choose $\theta = 1$. Yet, for unconditional cooperators who defect as a first mover, choosing $\theta = 4/5$ maximizes the prospects of explaining the observations. Yet, the implied intercept of -3 and the slope of 4 do not allow to explain neither of the 2 observations in the left pane of Figure 11, below. For the mismatchers with $\theta = 1/8$ the intercept equals zero and the slope equals $(7/8) \cdot (2/9) = 7/36$, so that one of the observations among the mismatchers who defect in the first-mover role can be accounted for. There is no parameter value that allows explaining the unique mismatcher who cooperates as the first move.

In sum, 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 cooperators who defect as the first move (parameter estimate $\theta = 0$ for this behavioral profile), the behavior of 12 of the 13 unconditional defectors who cooperate as a first move (with $\theta = 1/8$), and the behavior of all 7 unconditional cooperators who cooperate as a first move (with $\theta = 1$). The behavior of the conditional cooperators and the mismatchers is inconsistent with the model. This class of goal functions thus comes in third. That it beats the Homo economicus model is no surprise since it has one more parameter.

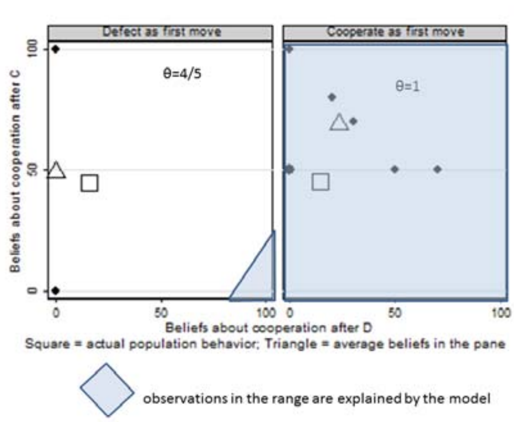

Figure 11: Unconditional cooperators. The explanatory power of the *Altruism* model.

4.5 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 and material payoffs 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.[10] In such a setting, a behavior strategy for a subject 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. In our experiment, we found the empirical mean-values of each of the three local strategies x_i to be $\bar{x}_1 \approx 0.56$, $\bar{x}_2 \approx 0.47$ and $\bar{x}_3 \approx 0.16$.

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 subject who uses strategy $x \in X$ against an opponent who uses strategy $y \in X$ is

$$\begin{aligned}\pi(x,y) &= \frac{1}{2}\left[(25y_2+5)\,x_1+(40y_3+10)\,(1-x_1)\right]\\&\quad+\frac{1}{2}\left[(50-20x_2)\,y_1+(10-5x_3)\,(1-y_1)\right]\\&= \frac{1}{2}\left[10+40y_3+(25y_2-40y_3-5)\,x_1+(50-20x_2)\,y_1+(10-5x_3)\,(1-y_1)\right]\end{aligned}$$

[10] 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 Figure 1, with player labels reversed in one of the copies.

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

$$u(x,y) = (1-\kappa)\,\pi(x,y) + \kappa\pi(x,x).$$

Hence, up to a positive monotone transformation, we may write

$$\begin{aligned} u(x,y) &= [(25y_2 - 40y_3 - 5)\,x_1 - 20y_1x_2 - 5(1-y_1)\,x_3]\cdot(1-\kappa) \\ &\quad + [35x_1 + 35x_3 + 5x_1x_2 - 35x_1x_3]\cdot\kappa \end{aligned}$$

Hence, if we know the subjective probabilities $\hat{y} = (\hat{y}_1, \hat{y}_2, \hat{y}_3)$ that a subject attaches to the different moves of his or her opponent, then we can calculate the best reply for *Homo moralis*. It follows that cooperating as a first mover, $x_1 = 1$, is optimal if and only if[11]

$$(8 - 5\hat{y}_2 + 8\hat{y}_3 + x_2 - 7x_3)\kappa \geq 1 - 5\hat{y}_2 + 8\hat{y}_3$$

or

$$\hat{y}_2 \geq \frac{1 - 8\cdot\kappa - x_2\cdot\kappa + 7x_3\cdot\kappa}{5\cdot(1-\kappa)} + \frac{8}{5}\hat{y}_3, \tag{12}$$

or, in the notation used for the other goal functions,

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

Likewise, cooperation in reaction to cooperation, $x_2 = 1$, is optimal iff

$$(4\hat{y}_1 + x_1)\kappa \geq 4\hat{y}_1, \tag{13}$$

and cooperation in reaction to defection, $x_3 = 1$, is optimal iff

$$(8 - 7x_1 - \hat{y}_1)\kappa \geq 1 - \hat{y}_1. \tag{14}$$

Unfortunately we do not have data on $\hat{y}_1$, the expected initial cooperation rate.[12] We analyze the predictive power of this class of goal functions by using two alternative approaches to fill in the missing data, and by then rating the goal function according to its explanatory power under the approach that gives it the lowest predictive power. One approach is to assume that the beliefs about first-mover choices is empirically correct, that is set $\hat{y}_1 = \bar{x}_1 = 0.5625$. The other approach is to assume a maximal consensus effect, where all subjects in the same behavior class believe that all subjects in the experiment behave like they do, in which case $\hat{y}_1 = x_1$. It turns out that the correct beliefs specification is the more demanding one in this context (leads to a slightly lower explanatory power), but for the sake of completeness, we will present both approaches here.

[11] See the first order conditions in the appendix.

[12] The experiments were run before *Homo moralis* preferences were discovered.

So let's consider the explanatory power of this one-parameter model. We begin with the conditional cooperators. Some of these cooperate as first movers, and others defect as first movers. Consider, first, a conditional cooperator who cooperates as a first move. Cooperation as second mover after cooperation is optimal iff

$$\kappa \geq \frac{4\hat{y}_1}{4\hat{y}_1 + 1}.$$

We do not know the conditional cooperator's belief about y_1, first movers' cooperation rate. If we suppose consensus beliefs, then $\hat{y}_1 = 1$, and the condition states that $\kappa \geq 4/5$. If we plug in $\kappa = 4/5$ and pure-strategy behavior into (12), we get

$$\hat{\mu}(C|C) = \hat{y}_2 \geq -\frac{31}{5} + \frac{8}{5}\hat{y}_3 = -\frac{31}{5} + \frac{8}{5}\hat{\mu}(C|D).$$

This line is below the horizontal axis in the right-hand pane of Figure 12, below, and all the points above this line (or above the horizontal axis) are consistent with the prediction that cooperation should be observed by the first mover. Therefore, all the points in the right-hand pane of the figure are consistent with the *Homo moralis* model. This is true both for consensus beliefs and empirically correct beliefs.

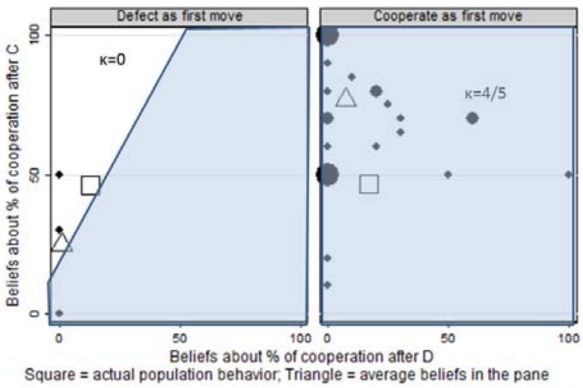

Figure 12: Conditional cooperators. The explanatory power of the *Homo moralis* model.

If a conditional cooperator defects as a first move and we apply consensus beliefs for $\hat{y}_1$, then condition (13) imposes no further limitations on κ but from (14) we get that $\kappa \leq 1/8$ is required. Subject to this restriction, the κ-value that maximizes the number of observations that can be explained equals 0. We are interested in conditional cooperators who defect as a first move. Defection is optimal if and only if (12) with the reversed inequality hold. With $\kappa = 0$, this coincides with the first-mover defection condition of *Homo economicus*, i.e.

$$\hat{\mu}(C|C) = \hat{y}_2 \leq \frac{1}{5} + \frac{8}{5}\hat{y}_3,$$

and thus only one of the three observations in the left pane of Figure 12 is consistent with the model. Assuming correct beliefs (instead of consensus beliefs), (13) requires that $\kappa \geq 1$, and thus none of the observations in the left pane of Figure 12 can be explained.

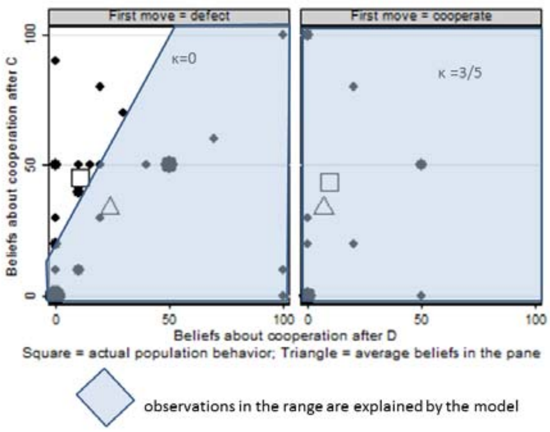

Figure 13: Unconditional defectors. The explanatory power of the *Homo moralis* model.

Suppose then that the unconditional defector defects as a first move. When it comes to the second mover choices, defection is optimal iff the reverse of (13) and (14), respectively, is satisfied while imposing $x_1 = 0$. If we suppose consensus beliefs, then $x_1 = 0 = \hat{y}_1$ and the reverse of (13) is satisfied for all parameter values. The reverse of (14) is satisfied iff $\kappa \leq 1/8$. Again the best fit in the left hand pane of Figure 13 is provided by $\kappa = 0$ (i.e. *Homo economicus*) in which case we get

$$\hat{\mu}(C|C) = \hat{y}_2 \leq \frac{1}{5} + \frac{8}{5}\hat{\mu}(C|D).$$

The explanatory power coincides with that of the *Homo economicus* model in this pane, i.e. both with consensus and correct beliefs, 22 out of the 32 observations are consistent with the model.

If an unconditional defector cooperates as a first move, then consensus beliefs would require $\kappa \leq 4/5$. Correct beliefs would require

$$\kappa \leq \frac{2.25}{2.25 + 1} = 0.6923.$$

Moreover, the reverse of (14) must be satisfied. If we assume consensus beliefs, this restriction does not impose any further constraints on allowable κ values. If we assume correct beliefs, then $\kappa \leq \frac{1-0.5625}{1-0.5625} = 1$. So again no further constraints are imposed. By picking $\kappa = 3/5 < 0.6923 < 4/5$, we get

$$\hat{\mu}(C|C) = \hat{y}_2 \geq -\frac{19}{10} + \frac{8}{5}\hat{y}_3,$$

and thus all observations of the right-hand pane of Figure 13 are consistent with the model (both with consensus and correct beliefs).

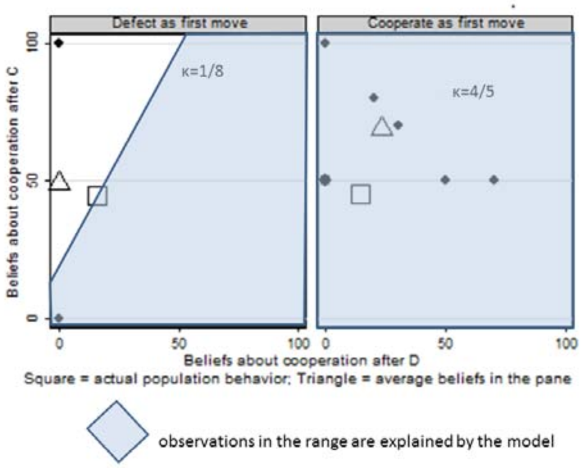

Figure 14: Unconditional cooperators. The explanatory power of the *Homo moralis* model.

Consider an unconditional cooperator who cooperates as a first move (and thus $x_1 = x_2 = x_3 = 1$). If we apply consensus beliefs, then condition (14) imposes no limitations on κ but from (13) we get that $\kappa \geq 4/5$ is required. Again, the implied condition for the optimality of first-mover cooperation is

$$\hat{\mu}(C|C) = \hat{y}_2 \geq -\frac{31}{5} + \frac{8}{5}\hat{y}_3 = -\frac{31}{5} + \frac{8}{5}\hat{\mu}(C|D)$$

and all the observations in the right pane of Figure 14 can be explained.

If we apply correct beliefs to an unconditional cooperator who cooperates as a first move, then (13) requires that

$$\kappa \geq \frac{2.25}{2.25+1} = 0.6923$$

and (14) requires that $\kappa = 1$, leading to

$$\hat{y}_2 \geq \frac{1-2\kappa}{5(1-\kappa)} + \frac{8}{5}\hat{y}_3$$

where the right-hand side approaches $-\infty$ as κ tends to one, so that all observations in the right pane of Figure 14 can be explained.

Consider then an unconditional cooperator who defects as a first move. With correct beliefs, based on the preceding argument but now with the requirement that

$$\hat{y}_2 \leq \frac{1-2\kappa}{5(1-\kappa)} + \frac{8}{5}\hat{y}_3$$

and $\kappa \gtrsim 0.69$ must hold, none of the observations in the left pane of Figure 14 can be explained.

Applying consensus beliefs, condition (13) imposes no limitations on κ but from (14) we get that $\kappa \geq 1/8$ is required leading to

$$\hat{y}_2 \leq \frac{1}{5} - \frac{\kappa}{5 \cdot (1-\kappa)} + \frac{8}{5}\hat{y}_3$$

and thus $\kappa = 1/8$ provides the highest explanatory power. Indeed, the inequality above thus reduces to

$$\hat{y}_2 \leq \frac{6}{35} + \frac{8}{5}\hat{y}_3,$$

and thus one of the two observations in the left pane of Figure 14 can be explained.

Finally, consider mismatchers, and start with those mismatchers who cooperate as a first move. Applying consensus beliefs, condition (14) imposes no limitations on κ but from (13) we get that $\kappa \leq 4/5$ is required. The κ with the greatest explanatory power equals 4/5 and again all the observations in the right pane of Figure 15 can be explained.

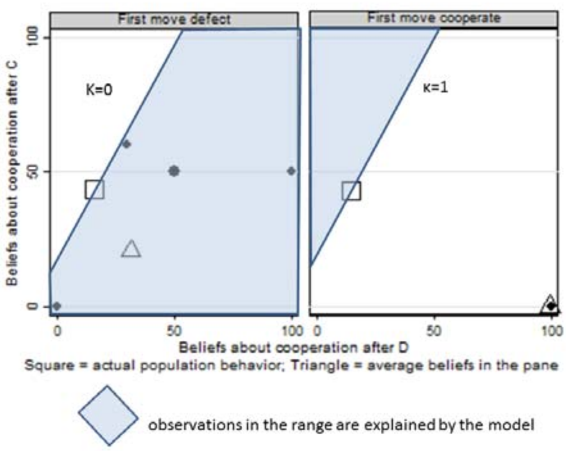

Figure 15: Mismatchers. The explanatory power of the *Homo moralis* model.

With correct beliefs, (13) imposes

$$\kappa \leq \frac{2.25}{2.25+1} = 0.6923$$

and (14) imposes $\kappa \leq 1$. Thus none of the observations can be explained.

For mismatchers who defect as a first move, the separating straight line

$$\hat{y}_2 = \frac{1}{5} + \frac{8}{5} \cdot \hat{y}_3$$

is independent of κ. Thus the explanatory power is independent of κ, and all of the observations in the left pane of Figure 15 can be explained.

In sum, this one-dimensional class of goal functions—representing Kantian morality—can explain the behavior of 80 out of the 96 subjects when missing belief data is replaced by empirically accurate beliefs, and it can explain the behavior of 83 of the 96 subject when missing belief data is replaced by maximal (false) consensus. All 33 conditional cooperators who cooperate as their first move (with an estimated $\kappa = 4/5$); 1 of the three conditional cooperators who defect as the first-move (with an estimated $\kappa = 1/8$); all 7 unconditional cooperators who cooperate as the first-move (with an estimated $\kappa = 4/5$); 1 of the two unconditional cooperators who defect as the first-move (with an estimated $\kappa = 1/8$); all 5 of the mismatchers who defect as the first-move (with an estimated $\kappa = 0$) ; 22 of the unconditional defectors who defect as the first move (with an estimated $\kappa = 0$) and all of the 13 unconditional defectors who cooperate as the first move (with an estimated kappa $\kappa = 3/5$). In the first approach the hit rate is 83% and in the second 87%. We use the lower hit rate as this model's prediction power.

4.6 Model comparison

The explanatory power of the five models, expressed as their "hit rate", the percentage of the subject pool whose behavior can be explained, are given in the table below.

TABLE 2

Model	**Hit rate**
Homo moralis	0.83
Social welfare	0.82
Inequity aversion	0.60
Altruism	0.44
Homo economicus	0.28

This suggests that *Homo moralis* and *Social welfare* models are most capable of explaining our data.[13] This rough comparison should of course be taken with a grain of salt. However, it

[13] Notice, however, that without the inequality restrictions on ρ and σ, the *Conditional efficiency* 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 87%.

suggests that prediction scores can be obtained on the basis of experiments of the kind we have made, and the explanatory power of alternative preference models can then be compared.

The above scores are rough, however, and they are not "fair" in that they disregard the number of free parameters and constraints imposed on these. For instance, one of the "winners", the *Social welfare* model, has the same number of parameters (two) as the *Inequity aversion* model, but the latter imposes two inequality constraints on its parameters. If these constraints and the non-negativity constraints of the *Social welfare* model parameters were relaxed, then the *Inequity aversion* model would obtain exactly the same hit rate as the *Social welfare* model. Indeed, the two models would then be mathematically identical. We also note that while the *Homo economicus* model has no free parameter, the *Altruism* and *Homo moralis* models have one each. Hence, the playing field is not "horizontal". For a "fair" comparison, the models should be given some "handicap" depending on their numbers of free parameters. To do this in a serious way is beyond the scope of this note. However, we conclude by making some observations and preliminary suggestions for how this could be done.[14]

The current literature in econometrics suggests two "handicap rules", the BIC (the Bayesian information criterion, Schwartz, 1978) and the AIC (the Akaike information criterion, Akaike, 1985), intended to enable comparison of the explanatory power of models with different numbers of parameters. Let $L(U,X)$ be the maximum likelihood of a model U given data X (that is, with the maximum likelihood estimates of the parameters of the model on the basis of X), let n be the number of observations in the data set X, and let k be the number of parameters in the model. Then

$$BIC(U,X) \;=\; k \cdot \ln n \;-\; 2 \cdot \ln L(U,X)$$

and

$$AIC(U,X) \;=\; 2k \;-\; 2 \cdot \ln L(U,X),$$

where a *lower* score represents higher explanatory power. If we use as our data set each subject's first-mover choice, then $n = 96$, and thus $\ln 96 \approx 4.56$. Consequently, the BIC will penalize models harder for their number of free parameters than the AIC. To give an indication of the orders of magnitude, we briefly sketch the simplest possible error specification in our five models and calculate their AIC and BIC scores on the basis of this.

Take any one of our five preference models U and suppose that it prescribes the first-mover choice $a_U \in \{C, D\}$. Assume that each subject obeys this utility function with probability $1-\varepsilon$ and otherwise randomizes uniformly over the two choices. Assuming statistical independence between subjects and letting X_i be the random variable describing the choice of subject i, we then have $\Pr(X_i = a_U) = 1 - \varepsilon/2$ for all subjects i. The number of choices according to the model U, the random variable $Y = X_1 + .. + X_n$, is binomially distributed, $Y \sim Bin(n,p)$ for

[14] Not only is the error specification simplistic; we also neglect the inequality constraints imposed by some of the models, which reduce the degree of freedom for some parameters. We are unaware of a methodology that would account for such restrictions.

$p = 1 - \varepsilon/2$. We consider three alternative values of ε: 5%, 10%, and 20% (see Tables 3-5 in the appendix). These rough calculations (based on unrestricted maximum-likelihood estimates) suggest the *Homo moralis* model as the winner. A more complete data set and a more carefully modelled error specification is needed if the explanatory power of the models are to be seriously compared.

4.7 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 cooperation rate conditional on defection than conditional cooperators.

In this subsection, we further analyze the observed differences in beliefs. We suggest a model which permits that subjective beliefs may deviate from actual population behavior and that the biases are influenced both by the consensus effect and optimism. The best-fitting such model, the one that minimizes the errors in predicting the stated beliefs about second-mover cooperation rates shows a 33% consensus bias and 15% optimism bias in the elicited beliefs.

To illustrate the calculations, consider a first-mover's beliefs about reactions to first-mover cooperation. We hypothesize that i's subjective belief about the second mover's cooperation rate is a combination of the actual second-mover cooperation rate, $\bar{x}_2$, i's own second-mover behavior (the consensus effect), x_2, and i's optimism bias, as follows:

$$\tilde{\mu}_i(C|C) = (1-\gamma) \cdot \bar{x}_2 + \gamma \cdot x_2 + \omega \cdot \frac{u_i(C,C) - u_i(C,D)}{u_i(C,C) + u_i(C,D)} \cdot \bar{x}_2. \tag{15}$$

Here γ is the weight on the consensus effect, the *consensus bias*, and ω the weight on the optimism effect, the *optimism bias.* In the third term on the right-hand side, $u_i(C,C)$ is i's utility if the second mover reacts cooperatively to i's cooperation, and $u_i(C,D)$ is $i's$ utility if the second-mover instead defects. We estimate these utilities parametrically according to the *Social welfare* model.[15] Note that the optimism effect vanishes if $u_i(C,C) = u_i(C,D)$; then $\tilde{\mu}_i(C|C) = (1-\gamma)\bar{x}_2 + \gamma x_2$. This illustrates that optimism acts through the subject's utility evaluations of outcomes. Since different behavior classes in our subject pool have different parameter estimates, the optimism effect may differ between these classes.

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 the left pane of Figure 9). 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 (15) accordingly becomes $\omega \cdot \frac{25}{35} \cdot 0.47 \approx 0.34\omega$, and since this term is positive there is a strong upward-biasing

[15] 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.

effect of optimism on the predicted conditional cooperation rate. The same calculation can be carried out for the beliefs about cooperation rates conditional upon first-mover defection. Similar exercises for all behavioral classes yields two prediction errors, defined as absolute deviations between subjective and objective cooperation rates, for each subject; $|\tilde{\mu}_i(C|C) - \hat{\mu}_i(C|C)|$ and $|\tilde{\mu}_i(C|D) - \hat{\mu}_i(C|D)|$, respectively. Minimization of the sum of prediction errors for all subjects results in the population estimates $\gamma \approx 0.33$ and $\omega \approx 0.15$. The consensus bias is thus about twice as strong as the optimism bias.

5 Conclusion

We here compare the explanatory power of five established preference models and identify subjects' belief biases. We do this by way of a novel approach that is, arguably, both intuitive and visually appealing, applied to a simple two-stage games, a sequential prisoners' dilemma. The five preference models are Homo economicus, altruism (Becker, 1976), inequity aversion (Fehr and Schmidt, 1999), a concern for social welfare (Charness and Rabin, 2002), and a concern for morality (Alger and Weibull, 2013). Using maximum likelihood techniques that use subjects' elicited beliefs about each others' behavior, we find that a concern for social welfare and a concern for morality share the "first prize" in this model horse-race. If account is taken for the number of free parameters in each of the five preference models, the latter comes out as the winner. 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 estimated preferences). We find that in our subject pool the consensus bias is about twice as strong as the optimism bias.

All five preference models are more complex than the *Homo economicus* 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 raise the explanatory power by adding just one free parameter from about 28% to about 83%. The psychological realism of the five preference models appears harder to evaluate, and on this we have no data, but arguably, each has a fairly clear intuitive and distinctive psychological appeal. Self-interest, altruism, inequity concern, a concern for social welfare and/or 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. Three limitations are particularly important. First, we throughout assume risk-neutrality, although recent evidence suggests that other-regarding preferences in the face of risk are more complicated than suggested by simple risk-neutrality (Trautmann and Vieider, 2011, Fudenberg and Levine, 2012, Miettinen et al. 2017). Secondly, and even more importantly, we make very strong simplifying error specifications. Third, our data set is very limited in that each subject plays a single game (protocol) only once.

In future research, it would thus be interesting to collect a more comprehensive data sets where several games (game protocols) are played by each participant (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; Charness and Rabin, 2002, Dufwenberg and Kirchsteiger, 2004, Weibull, 2004, Falk and Fishbacher, 2006, Cox et al. 2008, Gul and Pesendorfer, 2008), or even models of concerns for social image or self-image (Benabou and Tirole, 2006, Ellingsen et al. 2012; Malmendier et al. 2014).

Appendix

A1. Homo moralis

First-order conditions for the maximization problem are

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

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

$$\frac{\partial u(x,y)}{\partial x_3} = -5(1-y_1)(1-\kappa) + 35(1-x_1)\kappa.$$

A2. Model comparison

TABLE 3: 5% error.

model	parameters	BIC	AIC
Homo economicus	0	307	307
Inequity aversion	2	119	114
Social welfare	2	34	29
Homo moralis	1	27	24
Altruism	1	206	203

TABLE 4: 10% error.

model	parameters	BIC	AIC
Homo economicus	0	214	214
Inequity aversion	2	72	67
Social welfare	2	19	14
Homo moralis	1	13	10
Altruism	1	136	133

TABLE 5: 20% error.

model	parameters	BIC	AIC
Homo economicus	0	125	125
Inequity aversion	2	33	28
Social welfare	2	14	9
Homo moralis	1	10	7
Altruism	1	71	68

References

[1] Akaike, H. (1985): "Prediction and entropy". In *Selected Papers of Hirotugu Akaike*, pp. 387-410. Springer, New York.

[2] Alger, I., and J. Weibull (2013): "Homo moralis – preference evolution under incomplete information and assortativity", *Econometrica* 81, 2269-2302.

[3] Altmann, S., T. Dohmen, and M. Wibral (2008): "Do the reciprocal trust less?" *Economics Letters* 99, 454-457.

[4] Becker, G. (1976): "Altruism, egoism, and genetic fitness: economics and sociobiology", *Journal of Economic Literature* 14, 817-826.

[5] Bellemare, C., S. Kroger, and A. Van Soest (2008): "Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities", *Econometrica* 76, 815-839.

[6] Bénabou, R., and J. Tirole (2006): Incentives and pro-social behavior", *American Economic Review* 96, 1652-1678.

[7] Blanco, M., D. Engelmann, and H. Normann (2011): "A within-subject analysis of other-regarding preferences", *Games and Economic Behavior* 72, 321-338.

[8] Blanco, M., D. Engelmann, A. Koch, and H. Normann (2014): "Preferences and beliefs in a sequential social dilemma: a within-subjects analysis", *Games and Economic Behavior* 87, 122-135.

[9] Boschini, A., A. Muren, and M. Persson (2013): "The social egoist", WP 2013: 14, Department of Economics, Stockholm University.

[10] Charness, G., and M. Rabin (2002): "Understanding social preferences with simple tests", *Quarterly Journal of Economics* 117, 817-869.

[11] Cox, J., D. Friedman, and V. Sadiraj (2008): "Revealed Altruism", *Econometrica* 76, 31-69.

[12] Dillenberger, D., A. Postlewaite, and K. Rozen (2017): "Optimism and pessimism with expected utility", *Journal of the European Economic Association*, forthcoming.

[13] Dreber, A., D. Fudenberg, and D. Rand (2014): "Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics", *Journal of Economic Behavior and Organization* 98, 41-55.

[14] Dufwenberg, M., and G. Kirchsteiger (2004): "A theory of sequential reciprocity", *Games and Economic Behavior* 47, 268-298.

[15] Ellingsen, T., M. Johannesson, J. Mollerstrom, and S. Munkhammar (2012): "Social framing effects: Preferences or beliefs?", *Games and Economic Behavior* 76, 117–130.

[16] Falk, A., and U. Fischbacher (2006): "A theory of reciprocity", *Games and Economic Behavior* 54, 293-315.

[17] Fehr, E., and K. Schmidt (1999): "A theory of fairness, competition, and cooperation", *Quarterly Journal of Economics* 114, 817-868.

[18] Fischbacher, U. (2007):"z-Tree: Zurich toolbox for ready-made economic experiments", *Experimental Economics* 10, 171-178.

[19] Fudenberg, D., and D. Levine (2012): "Fairness, risk preferences and independence: Impossibility theorems", *Journal of Economic Behavior and Organization* 81, 606-612.

[20] Gächter, S., D. Nosenzo, E. Renner, and M. Sefton (2012): "Who makes a good leader? Cooperativeness, optimism and leading-by-example", *Economic Inquiry* 50, 867-879.

[21] Gul, F., and W. Pesendorfer (2016): "Interdependent preference models as a theory of intentions", *Journal of Economic Theory* 165, 179-208.

[22] Hey, J. D. (1984): "The economics of optimism and pessimism", *Kyklos* 37, 181-205.

[23] Kant, I. (1785): *Grundlegung zur Metaphysik der Sitten*. In English: *Groundwork of the Metaphysics of Morals*, 1964. New York: Harper Torch books.

[24] Levine, D. (1998): "Modeling altruism and spitefulness in experiments", *Review of Economic Dynamics* 1, 593-622.

[25] Malmendier, U., V. de Velde, and R. Weber (2014): "Rethinking reciprocity", *Annual Review of Economics* 6, 849-874.

[26] Miettinen, T., O. Ropponen, and P. Sääskilahti (2017): "Prospect theory, fairness, and the escalation of conflict at negotiation impasses", mimeo., Hanken School of Economics.

[27] Muren, A. (2012): "Optimistic behavior when a decision bias is costly: an experimental test", *Economic Inquiry* 50, 463-469.

[28] Puri, M., and D. Robinson (2007): "Optimism and economic choice", *Journal of Financial Economics* 86, 71-99.

[29] Ross, L., D. Greene, and P. House (1977): "The 'false consensus effect': An egocentric bias in social perception and attribution processes", *Journal of Experimental Social Psychology* 13, 279-301.

[30] Schwarz, G. (1978): "Estimating the dimension of a model", *Annals of Statistics* 6, 461-464.

[31] Spinnewijn, J. (2015): "Unemployed but optimistic: Optimal insurance design with biased beliefs", *Journal of the European Economic Association* 13, 130-167.

[32] Trautmann, S., and F. Vieider (2011): "Social influences on risk attitudes: Applications in economics", Chapter 29 in S. Roeser et al. (eds.), *Handbook of Risk Theory.* Springer, New York.

[33] Weibull, J. (2004): "Testing game theory", Chapter 6 in S. Huck (ed.), *Advances in Understanding Strategic Behaviour: Game Theory, Experiments, and Bounded Rationality - Essays in Honor of Werner Güth.* Palgrave MacMillan.

[34] Weinstein, N. D. (1980): "Unrealistic optimism about future life events", *Journal of Personality and Social Psychology* 39, 806-820.