WORKING PAPERS

N° 1056

November 2023

"Estimating Social Preferences and Kantian Morality in Strategic Interactions"

Boris Van Leeuwen and Ingela Alger

Toulouse School of Economics

Estimating Social Preferences and Kantian Morality in Strategic Interactions*

Boris van Leeuwen† Ingela Alger‡

15th November 2023

Abstract: Theory suggests that a form of Kantian morality has evolutionary foundations. To investigate the relative importance of Kantian morality and social preferences, we run a laboratory experiment on strategic interaction in social dilemmas. We structurally estimate social preferences and Kantian morality at the individual and aggregate level. We observe considerable heterogeneity in preferences. Finite mixture analyses show that the subject pool is well described as consisting of two or three types: all display a Kantian moral concern, which they combine with aheadness aversion, behindness aversion, or both. The value of adding Kantian morality to well-established preference classes (distributional preferences as well as reciprocity) is also evaluated.

JEL codes: C49, C72, C9, C91, D03, D84.

Keywords: social preferences, other-regarding preferences, distributional preferences, reciprocity, Kantian morality, morality, experiment, structural estimation, finite mixture models.

*We thank Jörgen Weibull for the very many helpful and stimulating discussions in earlier stages of this project. We also thank Gijs van de Kuilen, Wieland Müller, Arthur Schram, and Sigrid Suetens, as well as audiences at Goethe University Frankfurt, EUI Florence, Institute for Advanced Study in Toulouse, Stockholm School of Economics, Universitat de les Illes Balears, University of Amsterdam, University of Copenhagen, University of Göteborg, the conference on Markets, Morality, and Social Responsibility (Toulouse), the ESA 2021 Global Around-the-Clock Virtual Conference, the French Experimental Talks (FETS) workshop, the Virtual Behavioral Economics Seminar (VIBES), and the 2022 AEA/ASSA meetings for helpful suggestions and comments. I.A. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 789111 - ERC EvolvingEconomics) and IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program).

†Department of Economics, Tilburg University. b.vanleeuwen@uvt.nl

‡Toulouse School of Economics, CNRS, University of Toulouse Capitole, Toulouse, France, and Institute for Advanced Study in Toulouse. ingela.alger@tse-fr.eu

1 Introduction

Behavioral and experimental economics has over the past decades provided a host of insights about the motivations that drive human behavior in social dilemmas. Notwithstanding the wealth of preference classes that have been considered—notably, altruism (Becker, 1974), warm glow (Andreoni, 1990), inequity aversion (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000), reciprocity (Rabin, 1993; Charness & Rabin, 2002; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006), guilt aversion (Charness & Dufwenberg, 2006; Battigalli & Dufwenberg, 2007), and image concerns (Bénabou & Tirole, 2006; Ellingsen & Johannesson, 2008)—recent theoretical work has shown that yet another type of preferences should be considered, since it is strongly favored by evolutionary forces. The novel element is a form of Kantian moral concern, so called *Homo moralis* preferences (Alger & Weibull, 2013; Alger, Weibull, & Lehmann, 2020). The Kantian moral concern induces the individual to evaluate each course of action in the light of what material payoff (s)he would achieve, should others choose the same course of action. Theoretical analyses show that—compared to consequentialistic social (or selfish) concerns—this Kantian moral concern leads to qualitatively different behavioral predictions in many situations, such as consumption choices when these entail externalities (Laffont, 1975; Daube & Ulph, 2016), voting for environmental policies (Eichner & Pethig, 2021), voter coordination as well as information aggregation in large electorates (Alger & Laslier, 2022), incentive provision to teams (Sarkisian, 2017), voluntary contributions to public goods (Eichner & Pethig, 2022), and standard finite normal-form games (Alger & Weibull, 2013; Bomze, Schachinger, & Weibull, 2021). The purpose of this paper is to examine the explanatory power of such Kantian moral concerns, when these are assumed to be at work alongside social preferences such as altruism and inequity aversion. We do this by way of conducting an experimental study.

The laboratory experiment consists in letting each subject choose strategies in three classes of two-player social dilemmas: sequential prisoners' dilemmas, mini trust games, and mini ultimatum bargaining games. In such sequential games one subject moves before the other, and it is this feature that allows us to distinguish distributional motives from Kantian morality (à la *Homo moralis*, Alger & Weibull, 2013). Indeed, since each subject is told that he stands an equal chance of being a first- and a second-mover, Kantian morality would make him attach some value to the material payoff he would obtain if his strategy was universalized — as if he played against himself. By contrast, a subject with purely distributional preferences would make the subject attach value solely to the material payoff distribution that he expects to realize, given his beliefs about the opponent's strategy.[1]

In the main, pre-registered, analysis we posit a utility function with three parameters capturing attitudes towards unfavorable inequity, favorable inequity, and the Kantian moral concern, and we use the observed individual choices and reported beliefs in 18 different games (six games in each game class) to structurally estimate the preference parameter values for each individual subject, using a standard random utility model.[2] The use of such structural models has become more commonplace in experimental and behavioral economics, including the estimation of social preferences (DellaVigna, 2018). We also perform aggregate estimations, using a finite mixture approach, the same as that used by Bruhin, Fehr, and Schunk (2019) in their statistical analysis of social preferences.[3]

The estimations at the level of the individual subjects reveal substantial heterogeneity in preferences. While many subjects appear to be averse to unfavorable inequity (behindness aversion) and favorable inequity (aheadness aversion), some appear to be either indifferent or like favorable inequity. Importantly, the behavior of most subjects is compatible

[1]It is well known that the ability to control for subjects' beliefs when trying to identify their preferences is important (Bellemare et al., 2008; Miettinen et al., 2020). This is particularly true here, for Kantian morality reduces the sensitivity to beliefs. In the extreme case of an individual who would be driven entirely by the Kantian moral concern, the beliefs about the opponent's strategy would indeed be irrelevant, for such an individual would simply choose the "right thing to do." Hence, information about subjects' beliefs is crucial to distinguish Kantian moral concerns from consequentialistic ones. Accordingly, instead of hypothesizing subjects' beliefs about the behavior of their opponents (for example by some equilibrium hypothesis), we elicit each subject's belief in each strategic interaction. In further robustness checks, we also impose rational expectations instead.

[2]Social image concerns (Bénabou & Tirole, 2006) are muted because subjects are anonymously and randomly matched.

[3]See also Bardsley and Moffatt (2007), Iriberri and Rey-Biel (2013) and Breitmoser (2013), who use related mixture models to capture heterogeneity in social preferences.

with some concern for Kantian morality. Kantian morality further appears in all the aggregate estimations. The representative agent in the subject pool combines inequity aversion with Kantian morality. Models with two or three types provide a much better fit than the representative agent model. Our finite mixture estimations thus capture the heterogeneity in a tractable way. The two-types model has one type that combines inequity aversion with Kantian morality, while the other type combines behindness aversion with Kantian morality. With three types, all types display a concern for Kantian morality, combined with either behindness aversion, aheadness aversion, or a combination of the two (i.e. inequity aversion). Importantly, allowing for Kantian morality substantially improves the fit of the model.

In a second part of the analysis we add reciprocity parameters to the utility function, as in Charness and Rabin (2002) and Bruhin et al. (2019). These parameters (potentially) modify the subject's attitudes towards being ahead or behind as a second-mover, depending on whether the opponent's action as first-mover is deemed 'nice' or 'not nice'. Hence, reciprocity simply modifies the subject's attitude towards the other actual subject's payoff, i.e., his distributional concerns, depending on the opponent's action as a first mover. The Kantian moral concern is qualitatively different: it instead makes the subject evaluate what material outcome he himself would obtain if his strategy was universalized, without regard to the opponent's actual payoff. This distinction clearly appears in the estimates obtained with the extended model: the estimates of the Kantian moral concern parameter are essentially unaffected, while the estimates of the aheadness and the behindness aversion parameters are affected.

Our paper fits in the large literature that estimates or tests models of social preferences.[4] In relation to this literature, our main contribution is that we allow for the possi-

[4]See, for example, Palfrey and Prisbrey (1997); Andreoni and Miller (2002); Charness and Rabin (2002); Engelmann and Strobel (2004); Bardsley and Moffatt (2007); Fisman, Kariv, and Markovits (2007); Bellemare et al. (2008); Blanco, Engelmann, and Normann (2011); DellaVigna, List, and Malmendier (2012); Breitmoser (2013); Iriberri and Rey-Biel (2013); Ottoni-Wilhelm, Vesterlund, and Xie (2017) and, for recent surveys, see Cooper and Kagel (2015) and Nunnari and Pozzi (2022). Closest to our work in terms of empirical strategy is the recent study by Bruhin et al. (2019), who use the same finite mixture approach as we do, but who do

bility of Kantian morality as part of the motivation behind subjects' choices, in addition to social preferences. Closest to our work is the paper by Miettinen et al. (2020), who also allow for this possibility.[5] Our study is similar to theirs in two respects. First, both experiments rely on sequential games (our experimental design was indeed inspired by theirs in this respect). Second, in both experiments the subjects' beliefs about opponents' choices are elicited and used as controls in the empirical estimations. The key difference between ours and their study is that our data set is much richer: we collect data on individual choices in 18 strategic interactions while in their study each subject faces one single sequential prisoners' dilemma. Our data set gives us access to a rich set of empirical tools. In particular, while Miettinen et al. (2020) compare the explanatory power of six alternative utility functions, which involve either a consequentialistic, a reciprocity, or a Kantian concern, our data set enables us to estimate preference parameters at the individual level, and to apply finite mixture methods in order to detect the presence of common preference types that *combine* social preferences, Kantian morality, and reciprocity. As indicated by our results, most subjects indeed appear to have such complex preferences. Furthermore, our data allows us to conduct out-of-sample predictions to evaluate the explanatory power of the estimated preference types.

The remainder of this paper is organized as follows. Section 2 describes the experimental design and introduces the class of preferences we estimate, and Section 3 presents our econometric approach. The results of the pre-registered analysis (no reciprocity, subjective beliefs, and risk neutrality) are presented in Section 4, and we check the robustness of these results to allowing for risk aversion and rational expectations in Section 4.3. In Section 5 we incorporate reciprocity, and we also report several measures of the value added of Kantian morality in our experiment. Section 6 concludes.

not consider Kantian morality.

[5]See also Capraro and Rand (2018), who evaluate the explanatory power of *Homo moralis* preferences in standard games; however, and by contrast to our experiment and that by Miettinen et al. (2020), they rely on framing. More generally, economists are increasingly seeking to evaluate the explanatory power of non-consequentialistic motives; see, e.g., Bénabou, Falk, Henkel, and Tirole (2020).

2 The experiment: game protocols, preferences, and procedures

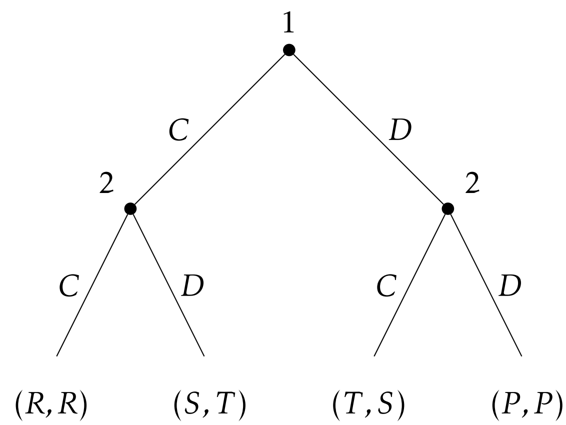
2.1 Game protocols

In the experiment, subjects play three types of well-known game protocols, illustrated in Figure 1: the Sequential Prisoner's Dilemma protocol (SPD), shown in Figure 1a, the mini Trust Game protocol (TG), shown in Figure 1b, and the mini Ultimatum Game protocol (UG), shown in Figure 1c.[6] We use the standard notation for prisoners' dilemmas, where R stands for "reward", S for "sucker's payoff", T for "temptation", and P for "punishment", and we throughout assume $T > R > P > S$.

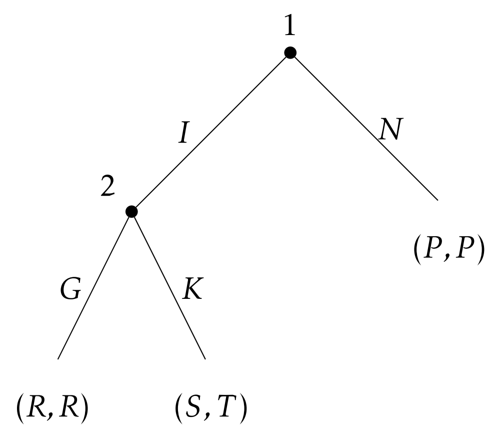
The objective of the experiment is to test whether Kantian morality (à la *Homo moralis*, Alger & Weibull, 2013) can help explain the choices subjects make in these game protocols. A subject with such Kantian morality evaluates each strategy in the light of what his/her material payoff would be if, hypothetically, the opponent were to choose the same strategy. This requires that the interaction is symmetric. To symmetrize the game protocols in Figure 1—which are asymmetric with one first-mover and one second-mover—we make it clear to the subjects that they are equally likely to be drawn to play in each player role. This defines a symmetric (meta) game protocol, in which "nature" first draws the role assignment, with equal probability for both assignments, and then the players learn their respective roles. The game tree corresponding to this game protocol for the SPD is shown in Figure 2. A behavior strategy consists of specifying (potentially randomized) choices at *all* decision nodes in this game protocol. Let $x = (x_1, x_2, x_3)$ denote the behavior strategy of subject i in this game tree: x_1 is the probability that i plays C as a first mover, x_2 the probability that i plays C as a second mover following play C by the opponent, and x_3 the probability that i plays C as a second mover following play D by the opponent. Likewise, let $y = (y_1, y_2, y_3)$ denote the behavior strategy used by the opponent (subject

[6] By a "game protocol", we mean a game tree and associated monetary payoffs.

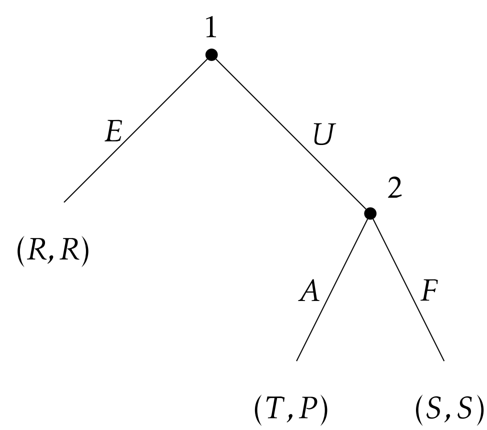
Figure 1: Game protocols



(a) Sequential Prisoner's Dilemma game protocol

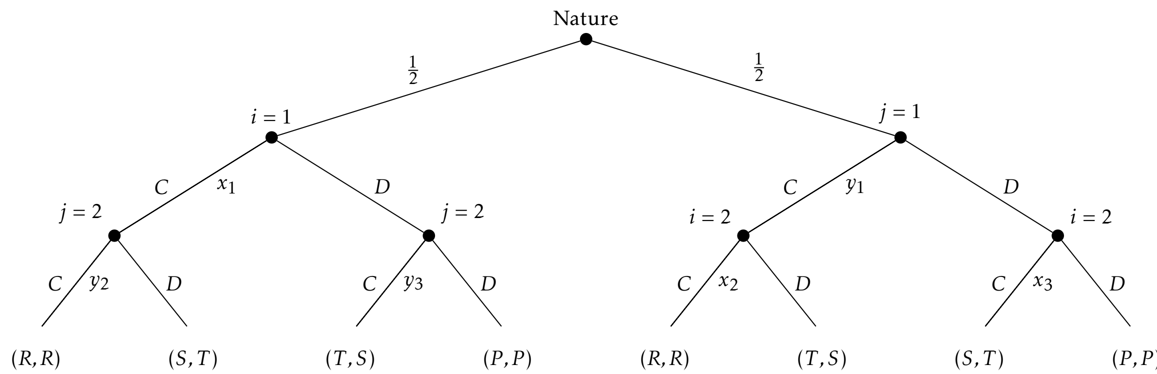


(b) Trust Game protocol



(c) Ultimatum Game protocol

Figure 2: Meta-game protocol for the SPD



j). Each strategy pair (x, y) determines the realization probability $\eta_{(x,y)}(\zeta)$ of each play ζ of the game protocol, where a *play* is a sequence of moves through the game tree, from its "root" to one of its end nodes (see Figure 2). For example: $\eta_{(x,y)}((1, C, C)) = \frac{x_1 \cdot y_2}{2}$ and $\eta_{(x,y)}((2, D, C)) = \frac{(1-y_1) \cdot x_3}{2}$.

Turning to the two other game protocols, when the trust game protocol is symmetrically randomized, a behavior strategy is a vector, $x = (x_1, x_2) \in [0,1]^2$, where x_1 is the probability with which i invests (selects I) and x_2 the probability with which i gives back something (selects G) if the first-mover invested. When the ultimatum game protocol is symmetrically randomized, a behavior strategy is a vector, $x = (x_1, x_2) \in [0,1]^2$, where x_1 is the probability with which i proposes an equal sharing (selects E), and x_2 the probability with which i accepts an unequal sharing (selects A). Like in the SPD game protocol, for both the TG and the UG protocols we denote by $y = (y_1, y_2)$ the strategy of i's opponent j, and write $\eta_{(x,y)}(\zeta)$ to denote the probability of each play ζ of the game protocol at hand.

Having formally defined the game protocols, we are in a position to define the utility function that we posit.

2.2 Preferences

In our empirical analysis we posit preferences that combine material self-interest, attitudes towards being ahead as well as behind (Fehr & Schmidt, 1999), reciprocity (Charness & Rabin, 2002), and a Kantian moral concern (Alger & Weibull, 2013). Thus, let the expected utility of a subject i playing against a subject j be

$$\begin{aligned} u_i(x,y) &= (1-\kappa_i)\cdot\sum_{\zeta}\eta_{(x,y)}(\zeta)\cdot\pi_i(\zeta) \\ &\quad -(\alpha_i+q\delta_i)\cdot\sum_{\zeta}\eta_{(x,y)}(\zeta)\cdot\max\left\{0,\pi_j(\zeta)-\pi_i(\zeta)\right\} \\ &\quad -(\beta_i+p\gamma_i)\cdot\sum_{\zeta}\eta_{(x,y)}(\zeta)\cdot\max\left\{0,\pi_i(\zeta)-\pi_j(\zeta)\right\} \\ &\quad +\kappa_i\cdot\sum_{\zeta}\eta_{(x,x)}(\zeta)\cdot\pi_i(\zeta), \end{aligned} \tag{1}$$

where x and y are i's and j's behavior strategy, respectively, $\pi_i(\zeta)$ is i's material payoff following play ζ and $\pi_j(\zeta)$ that of j. The dummy variable q takes the value 1 if j 'misbehaved' and 0 otherwise, while the dummy variable p takes the value 1 if j 'behaved nicely' and 0 otherwise.[7] We follow Charness and Rabin (2002) by labeling a first-mover action as misbehavior if it excludes an outcome that has maximal joint monetary payoffs and as nice behavior if it includes an outcome that has maximal joint monetary payoffs.[8]

This utility function has five parameters. Two of them are the familiar measures of inequity aversion. The parameter α_i captures i's disutility (if $\alpha_i > 0$) or utility (if $\alpha_i < 0$)

[7] Note that we assume "ex-post" inequity aversion. For a discussion of "ex-post" and "ex-ante" inequity aversion, see for example Krawczyk and Le Lec (2010), Cappelen, Konow, Sørensen, and Tungodden (2013), Brock, Lange, and Ozbay (2013) and Krawczyk and Le Lec (2016).

[8] For our case this means that defecting (resp. cooperating) as a first mover in a SPD protocol (if $2R > T+S$) constitutes misbehavior (resp. nice behavior). Furthermore, not investing in a TG protocol constitutes misbehavior (note, however, that the δ_i term cancels in the latter case, as not investing will lead to equal payoffs for both players). In addition, we also label not proposing an equal split in the UGs as misbehavior, while proposing the equal split is nice behavior (although the γ_i term cancels in the latter case, as proposing the equal split leads to equal payoffs for both players). The astute reader will have noticed that in Charness and Rabin (2002) negative (resp. positive) reciprocity is at work independently of whether the individual's material payoff is smaller (resp. larger) than that of the opponent. In our experimental setting the two formalizations would lead to the same behavioral predictions, since negative (resp. positive) reciprocity is relevant only when the individual is behind (resp. ahead) materially. The specification in (1) makes it clear that δ_i and γ_i act as shifters of α_i and β_i respectively, and are thus qualitatively different from κ_i.

from disadvantageous inequity, i.e., from falling short in terms of material payoff in the interaction. Likewise, the parameter β_i captures i's disutility (if $\beta_i > 0$) or utility (if $\beta_i < 0$) from advantageous inequity, i.e., from being ahead in terms of material payoff. Then, reciprocity is captured by the parameters δ_i and γ_i. Finally, κ_i captures the Kantian moral concern (à la *Homo moralis*, Alger & Weibull, 2013). It places weight on the expected material payoff that the subject would obtain if, hypothetically, both individuals were to use the subject's strategy x. Under this hypothesis, the probability that a play ζ would occur is $\eta_{(x,x)}(\zeta)$. A κ_i-value strictly between zero and one represents a partly deontological motivation, an individual who, in addition to the social concern that consists in caring about his or her own material payoff and that to the other individual in the interaction, is also motivated by what is the "right thing to do", what strategy to use if it were also used by the opponent. To choose a strategy x in order to maximize the last term in (1) is to choose a strategy that maximizes material payoff if used by both subjects (see Alger & Weibull, 2013, for a discussion).[9]

The utility function in (1) nests many familiar utility functions in the literature. Clearly, setting all five parameters to zero, $\alpha_i = \beta_i = \kappa_i = \delta_i = \gamma_i = 0$, represents pure self-interest and thus amounts to the classical *Homo oeconomicus*. The Fehr and Schmidt (1999) model of inequity aversion is obtained by setting $\alpha_i \geq \beta_i > 0$ and $\kappa_i = \delta_i = \gamma_i = 0$. One obtains Becker's (1974) model of pure altruism by setting $\kappa_i = \delta_i = \gamma_i = 0$ and $\alpha_i = -\beta_i$, for some $\beta_i \in (0, 1/2]$.[10] Here β_i is the individual's "degree of altruism", the weight placed on the other subject's material payoff, while the weight $1 - \beta_i$ is placed on own material payoff. Pure *Homo moralis* preferences are obtained by setting $\alpha_i = \beta_i = \delta_i = \gamma_i = 0$ and $\kappa_i \in (0, 1]$. Here κ_i is the individual's "degree of Kantian morality", the weight placed on the material payoff that would be obtained if both subjects in the interaction at hand played x, the strategy used by individual i, while the weight $1 - \kappa_i$ is placed on own material payoff,

[9]Note that the *Homo moralis* motivation is clearly distinct from behavioral motivations based on biased beliefs, such as the false consensus effect (Ross, Greene, & House, 1977) or magical thinking (Daley & Sadowski, 2017), whereby an individual overestimates the likelihood that the opponent plays the same strategy as him/her. Any such biased beliefs would indeed appear in the first term in the utility function in (1).

[10]See also the note by Engelmann (2012) on extending inequity aversion models to incorporate altruism.

given the strategy profile (x, y) effectively played. Finally, the utility function in (1) also nests models with reciprocity like in Charness and Rabin (2002) and Bruhin et al. (2019) when δ_i, γ_i, α_i, and β_i are non-nil and $\kappa_i = 0$.

A detailed account of how our experimental design allows us to disentangle the motivations is provided in subsection 2.4. However, the qualitative distinction between reciprocity and a Kantian moral concern is already clear from (1): while the reciprocity parameters δ_i and γ_i are akin to modifications of the parameters α_i and β_i that capture the subject's attitude towards the other actual subject's payoff, the Kantian moral concern instead makes the subject evaluate what material outcome that he himself would obtain if his strategy was universalized.

2.3 Experimental procedures

In total, 136 subjects (69 men, 67 women) participated in the experiment. We conducted 8 sessions at the CentERlab of Tilburg University, with between 12 and 22 subjects per session. Using the strategy method, each subject made decisions both as a first mover and a second mover for 18 game protocols (6 SPDs, 6 TGs and 6 UGs),[11] for different monetary payoff assignments T, R, P and S, listed in Table 1.[12]

All payoffs are denoted in 'points', where one point is equivalent to 17 eurocents. At the beginning of each session, the order of the 18 game protocols was fully randomized, meaning that participants could for example play an UG protocol first, then a TG protocol, followed by an SPD, and then another TG. For each game protocol, subjects first indicated what they would do at each decision node and second what they believed others would

[11] Iriberri and Rey-Biel (2011) find that "role uncertainty" increases social welfare maximizing behavior and decreases self-interested behavior in dictator games. Note that to estimate Kantian morality concerns, we require symmetric games and hence need a form of role uncertainty in our design (see subsection 2.1). Possibly, this means that with our design we estimate an upper bound on the importance of social preferences.

[12] In the process of selecting the number of game protocols and the monetary payoffs, we conducted simulations to verify if we could retrieve the simulated parameters, see also Appendix A5 for examples of these simulations.

Table 1: Game protocols: monetary payoffs, actions and beliefs

No.	T	R	P	S	x_1	x_2	x_3	y_1	y_2	y_3
Sequential Prisoner's Dilemmas										
1	90	45	15	10	0.18	0.15	0.10	0.33	0.20	0.13
2	90	55	20	10	0.24	0.20	0.06	0.30	0.21	0.07
3	80	65	25	20	0.35	0.29	0.13	0.32	0.30	0.16
4	90	65	25	10	0.29	0.31	0.03	0.31	0.25	0.08
5	80	75	30	20	0.43	0.50	0.04	0.40	0.41	0.11
6	90	75	30	10	0.30	0.40	0.01	0.33	0.33	0.08
All SPDs					0.30	0.31	0.06	0.33	0.28	0.11
Trust Games										
7	80	50	30	20	0.44	0.27	.	0.41	0.23	.
8	90	50	30	10	0.18	0.18	.	0.33	0.19	.
9	80	60	30	20	0.56	0.35	.	0.47	0.30	.
10	90	60	30	10	0.35	0.25	.	0.37	0.24	.
11	80	70	30	20	0.62	0.51	.	0.54	0.42	.
12	90	70	30	10	0.46	0.40	.	0.42	0.31	.
All TGs					0.44	0.33	.	0.42	0.28	.
Ultimatum Games										
13	60	50	40	10	0.49	0.96	.	0.48	0.91	.
14	65	50	35	10	0.52	0.96	.	0.49	0.88	.
15	70	50	30	10	0.46	0.96	.	0.47	0.87	.
16	75	50	25	10	0.43	0.90	.	0.47	0.83	.
17	80	50	20	10	0.60	0.88	.	0.51	0.79	.
18	85	50	15	10	0.60	0.81	.	0.55	0.72	.
All UGs					0.51	0.91	.	0.50	0.83	.

Notes: Here x_1, x_2 and x_3 denote action frequencies. In the SPDs, x_1 is the frequency by which the first mover plays C, x_2 the frequency by which the second mover plays C after C, and x_3 the frequency by which she plays C after D. In the TGs, x_1 is the frequency by which the first mover plays I, and x_2 the frequency by which the second mover plays G after I. For the UGs, x_1 is the frequency by which the first mover plays E, and x_2 the frequency by which the second mover plays A after U. Likewise, y_1, y_2 and y_3 are the mean values of the stated beliefs about x_1, x_2 and x_3. Table based on all 136 subjects.

do at each decision node.[13] In all game protocols, we used neutral labels. Two of the 18 game protocols were randomly selected for payment. To minimize the possibility to hedge, for one game protocol subjects were paid based on their actions and for the second game protocol they were paid based on the accuracy of their beliefs. For the payment based on actions, subjects were randomly matched in pairs and randomly assigned the role of first-mover or second-mover. Based on the actions in a pair, earnings for both subjects in the pair were calculated. For the payment based on beliefs, one decision node was randomly selected and subjects were paid using a quadratic scoring rule.

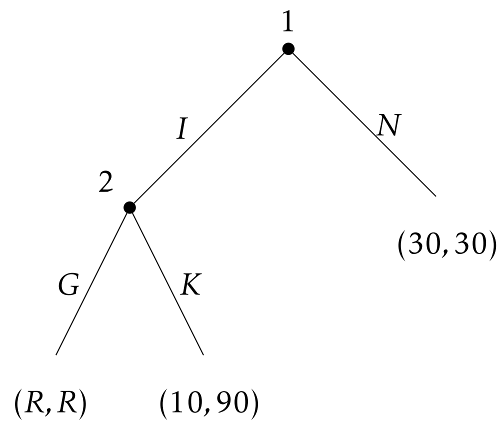
At the beginning of each session, subjects were randomly assigned a cubicle and read the instructions on-screen at their own pace. Subjects also received a printed summary of the instructions. At the end of the instructions subjects had to successfully complete a quiz to test their understanding of the instructions before they could continue. After completing the game protocols, we elicited risk attitudes using an incentivized method similar to the method of Eckel and Grossman (2002). Self-reported demographic data was gathered by way of asking the subjects to complete a short questionnaire at the end of the session. The instructions, quiz questions and risk elicitation task are reproduced in Appendix A7. Sessions took around 1 hour and subjects earned between €10.50 and €26.90 with an average of €18.80. Key features of the experimental design and main analyses were pre-registered.[14]

In Table 1, we present an overview of the average actions and beliefs for each game protocol. On average, observed behavior follows patterns that accord well with other experiments. For example, in the SPDs, on average subjects display conditional cooperation ($x_2 > x_3$). In the TGs, increasing the temptation payoff T and decreasing the sucker payoff S (compare game protocols 7 vs 8, 9 vs 10, 11 vs 12) reduces both trust (x_1) and trustwor-

[13]The literature on whether and how eliciting beliefs affects decisions provides mixed evidence. In Public Goods games for example, Croson (2000) finds that eliciting beliefs decreases contributions, while Gächter and Renner (2010) find that eliciting beliefs *increases* contributions and Wilcox and Feltovich (2000) find no effect of eliciting beliefs.

[14]See https://aspredicted.org/blind.php?x=4u5nu8 and Appendix A6. We pre-registered the type of game protocols (SPDs, TGs, UGs), the sample size, the main parameters of interest (α, β, κ), and using a logit model to estimate these parameters.

Figure 3: Trust Game protocol example



thiness (x_2). In the UGs, lower offers (P) are accepted less frequently (x_2). Moreover, on average actions (x) and beliefs (y) are highly correlated (see also Figure A.1 in Appendix A1). Table A.1 in Appendix A1 presents all decisions in the risk elicitation task. Based on their lottery choice, most subjects (83%) are classified as being risk-averse.

2.4 Distinguishing Kantian morality from social preferences

Many experimental studies use dictator game protocols to estimate social preferences. An advantage of such protocols is that they contain no strategic element, and hence there is no need to elicit subjects' beliefs about other subjects' behaviors. However, this class of game protocols would not allow us to distinguish between social preferences and Kantian morality à la *Homo moralis*, as shown in detail in Appendix A2. By instead using game protocols that contain strategic elements and collecting data on decisions at all nodes in the game tree as well as beliefs about opponent's play, our experimental design allows us to discriminate between social and Kantian moral preferences. Here we explain why.

The key effect is that an individual with a Kantian moral concern is not only influenced by his belief about the opponent's actual play, but also by what he would himself have done had the player roles been reversed (information that we collect in the experiment by using the strategy method). Put differently, an important consequence of Kantian morality is that a subject's preferences over moves off the equilibrium path associated with a strategy pair (x,y) may influence his or her decisions on its path. This differs sharply from distributional concerns such as altruism, inequity aversion or spite (whether or not augmented by reciprocity). We illustrate this with the Trust game protocol shown in Figure 3, for two different values of R: 50 and 70. Consider a subject who, for both values of R, believes that the opponent will keep (K) as a second-mover. If driven by purely distributional concerns (or reciprocity), such a subject should choose the same strategy for both values of R, since G is off the equilibrium path. By contrast, if the subject has a Kantian moral concern, and would himself choose G as a second-mover, the value of R does matter in his evaluation of all the behavior strategies. In particular, if the subject selects N for $R = 50$, a sufficiently high κ can make the subject switch to I for $R = 70$.

More generally and more formally, consider a (symmetrically randomized) Trust Game protocol (see Figure 1b) with $2R > T + S$. Suppose that an individual i believes that the opponent will play K ("keep") as second-mover and I ("invest") as a first-mover. The conditions for i to choose I as first-mover and G ("give back") as second-mover, respectively, are then:[15]

$$(1-\kappa_i)(S-P)-\alpha_i(T-S)+\kappa_i 2(R-P)\geq 0 \tag{2}$$

$$(1-\kappa_i)(R-T)+(\beta_i+p\gamma_i)(T-S)+\kappa_i(2R-S-T)\geq 0. \tag{3}$$

The first condition, which pertains to the choice as first-mover, shows formally the observation made above: the value of R, which given the subject's posited belief is off-the-equilibrium path, matters if and only if $\kappa_i > 0$. Indeed, Kantian morality makes this individual evaluate the improvement in the material payoff he would obtain from selecting I instead of N, under the hypothetical scenario that the opponent would also pick G as second-mover: collecting the terms multiplying κ, this improvement equals $R-P/2-S/2$ (the probability 1/2 has been omitted in (2)). Turning now to the choice as second-mover,

[15]These conditions are implied by the expressions (21) and (22) in Appendix A2. Note that even if the term that multiplies κ_i in (3), i.e., $2R-S-T$, is nil, these payoffs would still have an effect on the decision to choose I as first-mover, as seen in (2).

a positive κ_i makes the individual evaluate the increase in expected material payoff (the expectation being taken over the two player roles) he would obtain if he as well as the opponent (hypothetically) were to choose G rather than K as second-mover, given that he himself picks I as first-mover: collecting the terms multiplying κ, this equals $\frac{1}{2}(R-S)$ (the probability 1/2 has been omitted in (3)).

Two important implications appear from conditions (2) and (3). First, payoffs off-the-equilibrium path may matter: for example, condition (2) shows that a change in the payoff R (which is off the equilibrium path if the individual at hand moves first and his beliefs about his opponent are correct) can make the individual switch from N to I. Second, condition (3) reveals that in a model where the Kantian moral concern is omitted, an individual must be averse to being ahead ($\beta_i > 0$) for him to choose G. By contrast, an individual with a positive degree of morality $\kappa_i > 0$ may choose G even if $\beta_i = 0$. In fact, if κ_i is large enough, he can even be spiteful ($\beta_i < 0$) and still choose G.

We provide a detailed analysis of the first-order conditions for the three game protocols in Appendix A2.2. Furthermore, as an illustration of how Kantian morality may lead to different behavioral predictions than the social concerns included in the posited utility function (1), we compare the predicted behavior for five different preference types for all the payoffs used in the experiment in Table 2. The types are: pure self-interest ($\alpha_i = \beta_i = \gamma_i = \delta_i = \kappa_i = 0$), behindness aversion ($\alpha_i = 0.4$, $\beta_i = \gamma_i = \delta_i = \kappa_i = 0$), altruism ($\alpha_i = -0.2, \beta_i = 0.5$, $\gamma_i = \delta_i = \kappa_i = 0$), a combination of altruism and reciprocity ($\alpha_i = -0.2, \beta_i = 0.5$, $\gamma_i = \delta_i = 0.1, \kappa_i = 0$), or Kantian morality ($\alpha_i = \beta_i = \gamma_i = \delta_i = 0$, $\kappa_i = 0.2$). The behindness-averse type and the type combining altruism and reciprocity qualitatively resemble the behindness-averse and strongly altruistic type estimated by Bruhin et al. (2019).

All types display different behavior. In the Sequential Prisoner's Dilemma protocols, both self-interest and behindness aversion lead to unconditional defection as a second mover, but self-interested types will more frequently (opportunistically) cooperate as a first mover. An altruist will frequently unconditionally cooperate as a second mover, unless defection after cooperation leads to higher joint payoffs (SPD 1), or when punishment becomes sufficiently attractive (SPD 6). When enriching altruism with reciprocity, conditional cooperation emerges. Likewise, an individual motivated by Kantian morality ("*Homo moralis*") will typically conditionally cooperate, unless the benefits to joint cooperation become too small (SPDs 1 and 2). In particular, note that if $S + T > 2R$ (as in SPD 1), the convex combination of self-interest and Kantian morality entails a behavior not seen in any of the other types. By contrast to self-interest and behindness aversion, the Kantian moral concern entails a second-mover behavior that maximizes the expected material payoff from an *ex ante* perspective (i.e., cooperate following defection and *vice versa*). However, by contrast to the altruistic type in Table 2, which also selects this second-mover behavior, the type that combines self-interest and Kantian morality defects as a first mover: given that such an individual would cooperate as a second-mover following defection, both the self-interest part and the Kantian part of the utility function indeed entails a wish to defect.

The behavior of those motivated by Kantian morality differs even more strongly from those exhibiting a combination of altruism and reciprocity in the Trust Game and Ultimatum Game protocols. In the Trust Game protocols, (strong) altruists will always invest (I) as first mover and "give back" (G) as a second mover, while individuals motivated by Kantian morality will play "keep" (K) when R is relatively low.[16] In the Ultimatum Game, those motivated by Kantian morality will make unequal offers (U) and accept any offer (A), while those motivated by altruism and negative reciprocity will propose equal splits (E) and refuse low offers (F, UGs 17 and 18).

[16] Of course, these predictions depend on the strength of the degree of altruism or Kantian morality. However, only when $\kappa = 1$, an individual motivated by Kantian morality would always choose (I, G). In Table A.2 in Appendix A1, we show some more behavioral predictions for types motivated by different degrees of altruism and Kantian morality, illustrating that the qualitative differences between altruism and Kantian morality are not driven by a different weight on self-interest.

Table 2: Behavioral predictions

No.	T	R	P	S	self-interest $\alpha = 0$ $\beta = 0$ $\delta = 0$ $\gamma = 0$ $\kappa = 0$	behindness aversion $\alpha = 0.4$ $\beta = 0$ $\delta = 0$ $\gamma = 0$ $\kappa = 0$	altruism $\alpha = -0.2$ $\beta = 0.5$ $\delta = 0$ $\gamma = 0$ $\kappa = 0$	altruism + reciprocity $\alpha = -0.2$ $\beta = 0.5$ $\delta = 0.1$ $\gamma = 0.1$ $\kappa = 0$	homo moralis $\alpha = 0$ $\beta = 0$ $\delta = 0$ $\gamma = 0$ $\kappa = 0.2$
Sequential Prisoner's Dilemmas									
1	90	45	15	10	(D,D,D)	(D,D,D)	(C,D,C)	(C,D,C)	(D,D,C)
2	90	55	20	10	(D,D,D)	(D,D,D)	(C,C,C)	(C,C,D)	(C,D,D)
3	80	65	25	20	(C,D,D)	(D,D,D)	(C,C,C)	(C,C,C)	(C,C,D)
4	90	65	25	10	(C,D,D)	(D,D,D)	(C,C,C)	(C,C,D)	(C,C,D)
5	80	75	30	20	(C,D,D)	(C,D,D)	(C,C,C)	(C,C,D)	(C,C,D)
6	90	75	30	10	(C,D,D)	(D,D,D)	(C,C,D)	(C,C,D)	(C,C,D)
Trust Games									
7	80	50	30	20	(N,K)	(N,K)	(I,G)	(I,G)	(I,K)
8	90	50	30	10	(N,K)	(N,K)	(I,G)	(I,G)	(N,K)
9	80	60	30	20	(I,K)	(N,K)	(I,G)	(I,G)	(I,K)
10	90	60	30	10	(N,K)	(N,K)	(I,G)	(I,G)	(I,K)
11	80	70	30	20	(I,K)	(I,K)	(I,G)	(I,G)	(I,G)
12	90	70	30	10	(I,K)	(N,K)	(I,G)	(I,G)	(I,G)
Ultimatum Games									
13	60	50	40	10	(U,A)	(U,A)	(E,A)	(E,A)	(U,A)
14	65	50	35	10	(U,A)	(U,A)	(E,A)	(E,A)	(U,A)
15	70	50	30	10	(U,A)	(U,A)	(E,A)	(E,A)	(U,A)
16	75	50	25	10	(U,A)	(U,F)	(E,A)	(E,A)	(U,A)
17	80	50	20	10	(U,A)	(U,F)	(E,A)	(E,A)	(U,A)
18	85	50	15	10	(U,A)	(U,F)	(E,A)	(E,A)	(U,A)

Notes: Predicted behavioral strategies, assuming rational expectations (see Table 1 for average play in each game protocol).

3 Statistical analysis

The econometric strategy consists in producing both individual and aggregate estimates of the parameters in the utility function specified in (1) using a random utility model. In the main specification we employ subjects' stated beliefs (note that this implies that no equilibrium assumption is needed). We will then conduct several robustness checks and propose ways to evaluate the value-added of including Kantian morality.

3.1 Individual preferences

For each subject i, we estimate the individual's social and moral preference parameters α_i, β_i, δ_i, γ_i, and κ_i as specified in (1), using a standard additive error specification. We refer to these preference parameters using the vector $\theta_i = (\alpha_i, \beta_i, \delta_i, \gamma_i, \kappa_i)$. We consider pure strategies (that is, assigning a unique action at each decision node), and assume that subject i's true (expected) utility from using pure strategy x_i when $\hat{y}_i$ is i's expectation about his opponents behavior, is a random variable of the additive form

$$\tilde{u}_i(x_i, \hat{y}_i, \theta_i) = u_i(x_i, \hat{y}_i, \theta_i) + \varepsilon_{ix_i},$$

where $u_i(x_i, \hat{y}_i, \theta_i)$ is the expected utility of using strategy x_i given beliefs $\hat{y}_i$ following from the utility function in (1), and ε_{ix_i} is a random variable representing idiosyncratic tastes not picked up by the hypothesized utility $u_i(x_i, \hat{y}_i, \theta_i)$. Such a random utility specification sometimes induces choice of actions that do not maximize the deterministic component $u_i(x_i, \hat{y}_i, \theta_i)$. Assuming that the noise terms ε_{ix_i} are statistically independent (between subjects and across pure behavior strategies x_i for each subject) and Gumbel distributed with the same variance, the probability that subject i will use strategy x_i, given his probabilistic belief $\hat{y}_i$ about the opponent's play is given by the familiar logit formula (McFadden, 1974):

$$p_i(x_i, \hat{y}_i, \theta_i, \lambda_i) = \frac{\exp\left[(u_i(x_i, \hat{y}_i, \theta_i))/\lambda_i\right]}{\sum_{x' \in X_g} \exp\left[(u_i(x', \hat{y}_i, \theta_i))/\lambda_i\right]}, \tag{4}$$

where $\lambda_i > 0$ is a "noise" parameter, which is estimated alongside the preference parameters in θ_i, and X_g denotes the set of pure strategies in game protocol $g \in G$, where G is the set of game protocols. The smaller the parameter λ_i is, the higher is the probability that individual i makes his or her choices according to the hypothesized utility function $u_i(x_i, \hat{y}_i, \theta_i)$. We use maximum likelihood to estimate the preference parameter vector $\theta_i = (\alpha_i, \beta_i, \delta_i, \kappa_i)$ and the "noise" parameter λ_i for each individual i.[17] Then, the probability density function can be written as:

$$f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_i, \lambda_i) = \prod_{g \in G} \prod_{x \in X_g} p_i\,(x, \hat{y}_i, \theta_i, \lambda_i)^{I(i,g,x)}, \tag{5}$$

where $\mathbf{x}_i$ is the vector of the observed pure strategies of individual i, $\hat{\mathbf{y}}_i$ is the vector of stated beliefs of individual i about opponent's strategy in all the game protocols, and $I(i,g,x)$ is an indicator function that equals 1 if i played strategy x in game protocol g and 0 otherwise.

3.2 Aggregate estimations

We estimate preference parameters both for a representative agent and a given number of "preference types". For the representative agent, we simply aggregate all individual decisions and treat them as if they come from a single decision-maker. For the types estimations, we use finite mixture models, similar to the approach used by Bruhin et al. (2019). The finite mixture estimations allow us to capture heterogeneity in the population in a tractable way. For these estimations, we assume that there is a given number of types K in the population. For each type $k = \{1, ..., K\}$, we estimate the parameter vector

[17] In the maximum likelihood estimations, we use at least 4 different starting values for each parameter, so for the model with all six parameters $(\alpha_i, \beta_i, \delta_i, \gamma_i, \kappa_i, \lambda_i)$, we use $4^6 = 4{,}096$ starting values per individual i. For models with fewer than 5 parameters, we use 6 starting values per parameter.

$\theta_k = (\alpha_k, \beta_k, \delta_k, \kappa_k)$ and the noise parameter λ_k. The log-likelihood is then given by:

$$\ln L = \sum_{i=1}^{N} \ln\left(\sum_{k=1}^{K} \phi_k \cdot f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_k, \lambda_k)\right), \tag{6}$$

where ϕ_k is the population share of type k in the population. To maximize the log-likelihood in (6), we use an Expectation-Maximization (EM) algorithm (see for instance McLachlan, Lee, & Rathnayake, 2019).[18] As part of the EM algorithm, we estimate the posterior probabilities $\tau_{i,k}$ that individual i belongs to type k by:

$$\tau_{i,k} = \frac{\phi_k \cdot f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_k, \lambda_k)}{\sum_{m=1}^{K} \phi_m \cdot f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_m, \lambda_m)}. \tag{7}$$

4 Results of pre-registered analyses

The main analyses that we pre-registered were to estimate α_i, β_i, and κ_i, and to compare the predictive value of this model to restricted versions of the model (the pre-registration is reproduced in Appendix A6). In the following section, we present the results of these analyses assuming that subjects act on their subjective beliefs and are risk neutral.[19] In section 4.3, we perform several robustness analyses by allowing for risk aversion, rational expectations, or game protocol type specific noise parameters. Finally, in section 5, we will extend the pre-registered model to allow for reciprocity (δ and γ) and compare the added value of α, β, and κ, as well as δ and γ.

4.1 Individual preferences

Figure 4 shows the marginal distributions of the estimated individual preference parameters α_i, β_i, and κ_i for our core sample of 112 subjects.[20] For all three parameters, we

[18] We use 24 sets of starting values.

[19] In the estimations, we use the CRRA functions in equations (27) and (28) that we will discuss in section 4.3, and impose $r = 0$.

[20] In the estimations, we do not restrict the size or the sign of the parameter estimates. For most subjects, the parameter estimates are of reasonable size. However, for some subjects we obtain very large estimates of

Table 3: Individual parameter estimates

Parameter	Median	Mean	S.D.	Min	Max
α_i	0.11	0.16	0.20	−0.19	1.06
β_i	0.18	0.15	0.38	−1.55	1.08
κ_i	0.10	0.13	0.14	−0.16	0.72

Notes: Table based on the 112 subjects for whom the α_i, β_i and κ_i estimates have absolute value below 2. Table A.3 in Appendix A1 shows a similar table based on all 136 subjects.

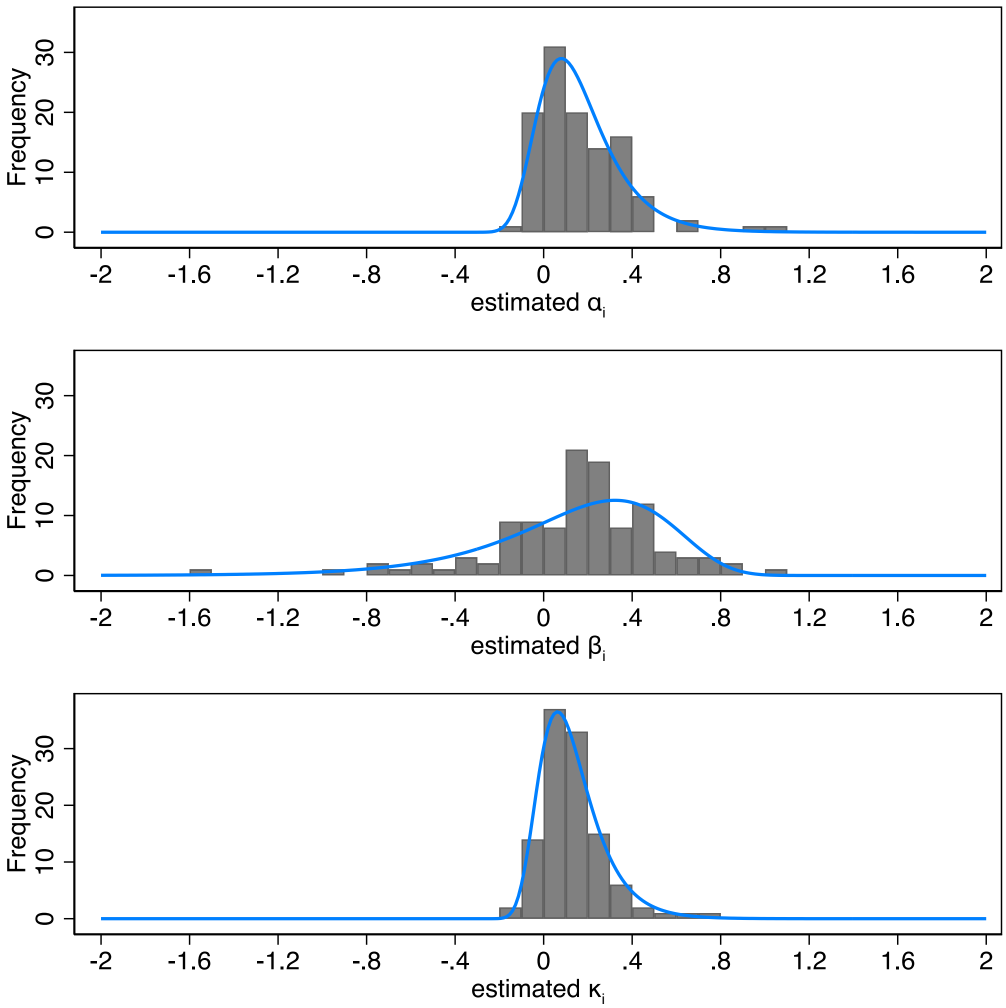
observe considerable heterogeneity. Most estimates of α_i, β_i, and κ_i are positive and signed-ranks tests confirm that the parameter distributions are located to the right of zero ($p < 0.001$ for either α_i, β_i, and κ_i estimates).

Table 3, which shows summary statistics for the parameter estimates, provides further support for the pattern observed in Figure 4. Median and mean estimates are positive for α_i, β_i and κ_i. Moreover, the relatively large standard deviations indicate that there is considerable heterogeneity in social preferences and Kantian morality.

Figure 5 illustrates the pairwise correlations between the three preference parameter estimates. The left panel of Figure 5 shows that the estimates for α_i and β_i are negatively correlated (Spearman's $\rho = -0.235$, $p = 0.013$, $n = 112$). For many individuals we observe a combination of $\alpha_i > 0$ and $\beta_i > 0$, in line with inequality averse preferences. However, we also observe a number of individuals for whom $\alpha_i > 0$ and $\beta_i < 0$, in line with spiteful or competitive preferences. The middle panel of Figure 5 reveals a strong and positive correlation between α_i and κ_i estimates (Spearman's $\rho = 0.427$, $p < 0.001$, $n = 112$). This means that many individuals combine a distaste for behindness aversion with Kantian morality. For the estimates of β_i and κ_i we find a negative correlation (Spearman's $\rho = -0.217$, $p = 0.022$, $n = 112$). We also use copula methods to describe the joint parameter

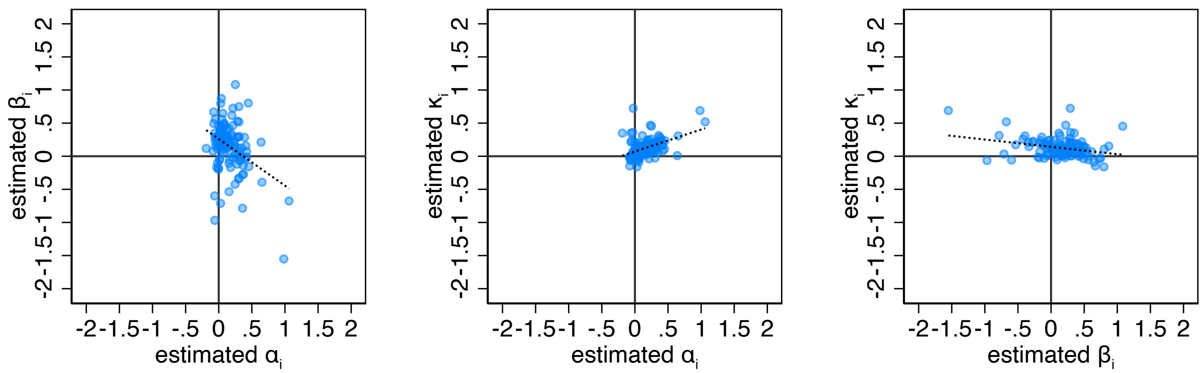
α_i, β_i, and/or κ_i (in absolute value), suggesting that our utility function (1) does not explain the decisions of these subjects well, either because they use a decision rule not nested in (1), or because their decisions are simply too noisy to be generated by any utility function. In the remainder of this section, we report results for our 'core sample', which consists of the 112 subjects for whom all three preference parameter estimates lie between -2 and 2. The fraction that we leave out in the main text (17.6%) is comparable in size to the fraction of 26.3% for whom Fisman et al. (2007) conclude that their decisions are too noisy to be utility-generated. In Appendix A1 we report results based on data for all 136 subjects. While the latter results are more noisy, they are qualitatively quite similar to those for the core sample.

Figure 4: Distributions of individual parameter estimates



Note: Figure based on the 112 subjects for whom the α_i, β_i and κ_i estimates have absolute value below 2. The (blue) lines indicate fitted Gumbel distributions (see Appendix A3 for details). Figure A.2 in Appendix A1 shows a similar figure based on all 136 subjects.

Figure 5: Correlations between estimated preference parameters



Notes: Each dot represents one subject. Dotted lines indicate linear predictions (intercept+slope). Specifically, we estimate $\beta_i = 0.26 - 0.70\alpha_i$, $\kappa_i = 0.07 + 0.33\alpha_i$, $\kappa_i = 0.14 - 0.11\beta_i$. Figure based on the 112 subjects for whom the α_i, β_i and κ_i estimates have absolute value below 2.

distributions for the individual estimates of α_i, β_i and κ_i. As for the pairwise correlations reported above, we observe that the individual estimates of α_i, β_i and κ_i are not statistically independent. Appendix A3 provides more details.

4.2 Aggregate estimations

We now turn to estimation of preferences at the aggregate level (see Section 3.2 for details). To distinguish these estimates from the individual ones, we use an index k to designate the type. Table 4 presents the estimates of the finite mixture models for one, two and three types.

4.2.1 The representative agent

When assuming only one type, that is, a representative agent, we obtain the estimates $\alpha_0 = 0.16$, $\beta_0 = 0.24$, and $\kappa_0 = 0.10$, where the index 0 stands for the representative agent. In other words, the representative agent dislikes both disadvantageous and advantageous inequity, and has a positive degree of Kantian morality. The representative agent thus exhibits Kantian morality and inequity aversion.

Table 4: Estimates at the aggregate level

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.16	0.12	0.18	0.18	0.01	0.17
	(0.01)	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)
β_k	0.24	0.35	0.00	0.27	0.47	0.00
	(0.03)	(0.04)	(0.04)	(0.06)	(0.06)	(0.04)
κ_k	0.10	0.10	0.10	0.11	0.15	0.09
	(0.01)	(0.02)	(0.01)	(0.02)	(0.04)	(0.01)
λ_k	7.19	8.44	3.96	8.83	6.47	3.68
	(0.47)	(0.66)	(0.54)	(0.94)	(1.00)	(0.25)
ϕ_k	1.00	0.62	0.38	0.48	0.17	0.36
	(-)	(0.07)	(0.07)	(0.07)	(0.06)	(0.05)
$\ln L$	-2441.1	-2254.4		-2225.3		
$EN(\tau)$	0.00	6.06		15.16		
ICL	4901.1	4557.3		4531.9		
NEC	-	0.032		0.070		

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 112 subjects. Table A.4 in Appendix A1 shows estimates based on the full sample. Table A.5 in Appendix A1 shows the estimates of a 4-type model.

4.2.2 The two- and three-type models

As can be seen in Table 4, in both multi-type models all types exhibit Kantian morality ($\kappa_k > 0$), roughly of the same order of magnitude as the representative agent. There is stronger heterogeneity in terms of the inequity aversion parameters α_k and β_k: in particular, some types exhibit only behindness aversion ($\alpha_k > 0$), some exhibit only aheadness aversion ($\beta_k > 0$), while some exhibit a combination of the two.

More specifically, when assuming two types, the most common type (Type 1) exhibits inequity aversion, with parameter estimates $\alpha_1 = 0.12$ and $\beta_1 = 0.35$, combined with a degree of Kantian morality $\kappa_1 = 0.10$. This type represents about 62% of the subjects. The other type, Type 2, exhibits a combination of behindness aversion and Kantian morality, with $\alpha_2 = 0.18$, $\beta_2 = 0.00$, and $\kappa_2 = 0.10$.

When assuming three types, for all types we again estimate a positive Kantian morality parameter κ_k. In comparison with the results under the two-types approach, Type 3 is very close to the previous Type 2. This type is again characterized as combining behindness aversion with Kantian morality, and represents a similar fraction of the population (36%).[21] The new Type 2 combines (strong) aheadness aversion with Kantian morality. It represents around 17% of the population. As in the two-types model, Type 1 in the three-types model combines inequity aversion with Kantian morality. This type represents 48% of the population. In sum: under the three-types approach, Type 1 displays a combination of inequity aversion and Kantian morality, Type 2 is aheadness averse and moral, and Type 3 is behindness averse and moral.

How do the estimated types behave? Table A.7 in Appendix A1 lists the chosen strategies for each of the three game protocols. For the multi-type models, we classify each subject i into a specific type by estimating the posterior probability $\tau_{i,k}$ that i belongs to type k (as defined in (7)). By taking the largest value $\tau_{i,k}$ for each subject i, we can assign each of the subjects to one of the types. Table A.7 unveils the following patterns.

First, in the two-type model, the "Type 2 subjects", who combine behindness aversion and Kantian morality, mostly choose to always defect (D, D, D) in the SPDs (in 84% of the cases), while "Type 1 subjects", who combine inequity aversion and Kantian morality, choose (D, D, D) less frequently (40%) and often conditionally cooperate (C, C, D) instead (30%). Similarly, in the TGs, Type 2 subjects most frequently choose not to invest as first mover and to "keep" as second mover (N, K) (83%), while Type 1 subjects most frequently invest as first mover and "give" as a second mover (I, G) (43%). In the UGs, Type 2 subjects mostly choose the unequal option as a first mover (75%) and accept unfair offers as a second mover (97%). Instead, Type 1 subjects most frequently propose an equal payoff (66%) and accept fewer unequal offers (90%).

Second, in the three-type model, Type 3 behaves almost identical as Type 2 in the two-types model. The new Type 1 and Type 2 differ in some respects. In the SPDs, the new

[21] In panel A of Table A.6 (see Appendix A1), we show a transition matrix for the two-types and three-types models. All but three subjects who are classified as Type 2 in the two-types model, are classified as Type 3 in the three-types model. All subjects who were classified as Type 1 in the two-types model are now distributed across the new Types 1 and 2.

Type 2 acts conditionally cooperative more often than Type 1. Similarly, Type 2 chooses to "give" more often than Type 1 in the TGs. In the UGs, Types 1 and Type 2 behave quite similarly.

In sum, the aggregate estimates lead to two observations. First, we observe relatively little heterogeneity in estimates of the morality parameter κ_k. In most cases, κ_k is around 0.1, showing that most people are well described by having Kantian morality concerns. Second, we note that in both multi-type models, we do not observe types who are best described by pure self-interest ($\alpha_k = \beta_k = \kappa_k = 0$). This is in line with the findings by Bruhin et al. (2019). Nonetheless, self-interest is still an important driver for all the types.

4.2.3 Comparing the one-, two-, and three-types models

Clearly, adding more types improves the fit of the model, but this comes at the cost of parsimony as well as precision of allocating individuals to types. Information criteria like the Bayesian information criterion (BIC) are not well suited to select the number of clusters (or in our case, 'types') in finite mixture models. In a recent overview paper on the use of finite mixture models, McLachlan et al. (2019) recommend using the 'integrated completed likelihood' (or 'integrated classification', ICL, Biernacki, Celeux, & Govaert, 2000). This criterion is approximated by

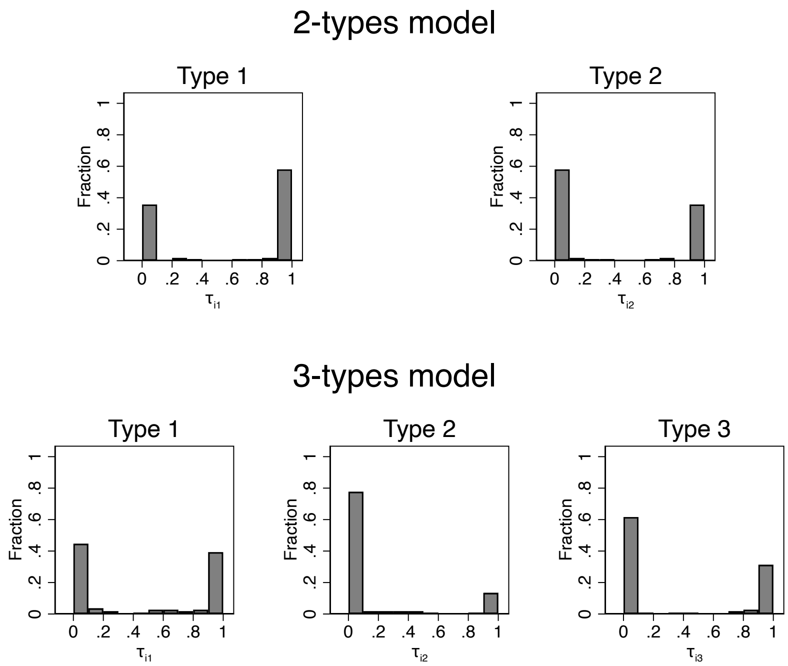
$$ICL = -2\ln L + d\ln N + EN(\boldsymbol{\tau}), \tag{8}$$

where the log-likelihood function $\ln L$ is defined as in (6), d is the number of estimated parameters, and N is the number of individuals in our sample. The last term in (8) is the entropy

$$EN(\boldsymbol{\tau}) = -\sum_{k=1}^{K}\sum_{i=1}^{N}\tau_{i,k}\ln\tau_{i,k}, \tag{9}$$

where $\tau_{i,k}$ is the estimated posterior probability of individual i belonging to type k, as defined in (7). This implies that the stronger individuals are assigned to types (i.e. all $\tau_{i,k}$'s

Figure 6: Posterior probabilities of type classifications



Notes: Distributions of the estimated posterior probability $\tau_{i,k}$ of individual i belonging to type k for the two-types and three-types finite mixture models reported in Table 4.

close to zero or one), the lower the entropy will be. In other words, the ICL extends the BIC by adding an additional penalty if individuals are assigned imprecisely to types.

Figure 6 shows the distributions of the estimated posterior probability $\tau_{i,k}$ (of individual i belonging to type k) for the two-type and three-type models. In all cases, most estimated $\tau_{i,k}$ are very close to zero or 1, which implies that most individuals are quite precisely assigned to a type. For the two-types model, virtually all estimated $\tau_{i,k}$ are close to zero or one. For the three-types model, a few individuals are imprecisely classified.

Bruhin et al. (2019) use the 'normalized entropy criterion' (NEC, Celeux & Soromenho, 1996), which is defined as:

$$NEC = \frac{EN(\boldsymbol{\tau})}{\ln L(K) - \ln L(1)}, \tag{10}$$

where $\ln L(1)$ is the log-likelihood of the representative agent model and $\ln L(K)$ the log-likelihood of the model with K types. Hence, the NEC weighs the precision of the type classifications $\tau_{i,k}$ by the increase in the log-likelihood compared to the representative agent model.

Table 4 shows statistics for both the ICL and the NEC. For both metrics, a lower score indicates a more preferred model. The NEC selects the two-types model and the ICL selects the three-types model. Table A.5 in Appendix A1 shows estimates and goodness-of-fit metrics for a four-types model. The four-types model performs worse on both criteria than the three-types models in Table 4. Note that marginal improvement in the ICL score is largest when going from the representative agent to the two-types model. In sum, assuming two types instead of a representative agent brings us a long way in capturing the heterogeneity in the population.

4.3 Robustness

Here we examine the robustness of the results reported above by allowing for risk aversion, rational expectations, and for game-specific noise parameters λ. We only discuss the main findings; more details are provided in Appendix A4.

4.3.1 Risk attitudes

In the main analysis, we imposed risk neutrality. However, since each subject in our experiment faces risky decisions (the monetary payoff depends on the decision of the opponent, which the subject does not know when making the decisions), we here report estimations allowing for risk aversion.[22] Thus, we will here take the term $\pi_i(\zeta)$ in the utility function in (1) to be the Bernoulli function value that the individual attaches to the monetary payoffs under play ζ. In a recent paper, Apesteguia and Ballester (2018) show that estimating risk aversion parameters using a random utility model may be problematic. To avoid this, we estimate the social preference and Kantian morality parameters imposing

[22]Following Rabin (2000), expected utility theory may not be best-suited to capture small-stakes risk aversion, and behavior in line with risk aversion may also be explained by other sources as loss aversion or mental accounting (Rabin & Thaler, 2001). Our experiment is not designed to disentangle different sources, however.

risk attitudes. Here we present the results for the aggregate estimates, for which we estimate mixture models under the assumption that all subjects have logarithmic utility over monetary outcomes, i.e., $\pi_i(\zeta) = \ln m_i(\zeta)$ and $\pi_j^i(\zeta) = \ln m_j(\zeta)$, where $\left(m_i(\zeta), m_j(\zeta)\right)$ is the monetary payoff allocation after a play ζ. In Appendix A4.1 we also present the individual estimates, allowing for the risk attitude to vary between individuals.[23]

Table 5 shows the estimates of finite mixture models under logarithmic utility. Comparing these results with those in Table 4, one sees that, qualitatively, estimates of the parameters α_k and κ_k are not much affected, although the Kantian morality parameter values are higher under risk aversion than under risk neutrality. The finite mixture estimates of the parameters β_k tend to be higher under risk neutrality than under risk aversion. Moreover, under risk neutrality, all estimates of β_k are non-negative, in contrast to the risk aversion estimates, where we observe $\beta_k < 0$ for some types k.[24] To see why risk aversion leads to lower degrees of aheadness aversion—sometimes even aheadness loving—than under risk neutrality, consider the Ultimatum Game protocol. In the Ultimatum Game, both risk aversion and aheadness aversion ($\beta > 0$) would induce one to choose the equal split E over the unequal split U. Hence, for a risk-averse individual who plays E, we may obtain a larger estimated β under risk neutrality than under risk aversion. While the prevalence of spite may appear surprising, it is in line with the theoretical prediction of Alger et al. (2020). They show in a general model that preferences that combine material self-interest, a Kantian moral concern and a social concern at the material payoff level is what should be expected in most human populations, and they identify evolutionary scenarios in which spite rather than altruism towards the other when ahead is favored.

The ICL criterion allows comparison of the fit of the risk-aversion and risk-neutral mod-

[23] At the individual level, the respective parameter estimates under risk neutrality and risk aversion preferences are strongly correlated (see Appendix A4.1 for a detailed analysis). Under risk aversion however, we observe a substantial directional shift in the estimates of β_i compared to the risk-neutral case. While most estimates of β_i are positive under risk neutrality, most estimates of β_i are negative under risk aversion. There is also a shift in the estimates of κ_i towards higher values.

[24] Table A.6 shows that the assignment of subjects to types for the risk-neutral two-types (panel B) model, is very similar to when we impose logarithmic $r_k = 1$. For the three-types models (panel C), some who are classified as "Type 2" with $r_k = 1$ are classified as "Type 1" under risk-neutrality and vice versa.

Table 5: Estimates at the aggregate level (logarithmic utility)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.13	0.05	0.24	0.12	−0.03	0.24
	(0.02)	(0.03)	(0.04)	(0.06)	(0.05)	(0.04)
β_k	−0.01	0.09	−0.29	0.22	−0.06	−0.29
	(0.03)	(0.03)	(0.07)	(0.04)	(0.07)	(0.08)
κ_k	0.20	0.23	0.17	0.24	0.21	0.17
	(0.01)	(0.02)	(0.02)	(0.06)	(0.05)	(0.02)
λ_k	0.24	0.27	0.16	0.20	0.31	0.15
	(0.02)	(0.02)	(0.02)	(0.03)	(0.05)	(0.01)
ϕ_k	1.00	0.59	0.41	0.29	0.30	0.41
	(-)	(0.06)	(0.06)	(0.07)	(0.07)	(0.06)
$\ln L$	-2356.8	-2165.6		-2140.0		
$EN(\tau)$	0.00	5.43		17.02		
ICL	4732.5	4379.1		4363.1		
NEC	-	0.028		0.078		

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 112 subjects. In these estimations, we impose logarithmic Bernouilli utility for all types.

els, respectively (see Tables 4 and 5). For any given number of types, the risk-aversion model has a considerably lower ICL score than the risk-neutral model. For the three-types model, for example, the ICL score under the risk aversion assumption is quite a bit lower than under risk neutrality (4363.1 versus 4531.9), showing that the risk-aversion model considerably improves the fit over the risk-neutrality model.

4.3.2 Rational expectations

So far, we assumed that people maximize expected utility given their (reported) subjective expectations. In Appendix A4.2, we estimate the preference parameters taking rational expectations instead. At the individual level, the estimated individual preference parameters under subjective and rational expectations are significantly correlated. At the aggregate level, the finite mixture models under rational expectations (see Table A.18 in Appendix A4.2) are qualitatively similar to those under subjective expectations for most types, although we also observe some differences for a part of the population. In particular, we observe that Type 2 in the two-types model, and Type 3 in the three-types model, now display spite ($\alpha_k > 0, \beta_k < 0$) with strong morality ($\kappa_k > 0$). This contrasts with the estimates under subjective expectations where these types combined behindness aversion with milder morality.[25] Given a number of types, the ICL scores under rational expectations are higher than under subjective expectations, indicating a worse fit under rational expectations.

4.3.3 Game-specific noise parameters

In the main analyses, we assume that the noise parameter λ is the same across game protocols. However, it could be that the error variance, and hence the noise parameter, is greater in certain type of game protocols. In Table A.19 in Appendix A4.3 we show finite mixture models where we allow for different noise parameters λ for each game protocol type (SPD, TG, UG). The estimates of the preference parameters of the 1-type and 2-types model are nearly identical to those in Table 4. For the 3-types model we observe some minor differences, but still the types are qualitatively similar. In particular, all the estimates of κ are significant, and in the same ballpark as in the main analysis.

5 The value added of distributional preferences, Kantian morality, and reciprocity

In this section, we extend the pre-registered analysis to also include the reciprocity parameters δ_i and γ_i of the utility function in (1), and we benchmark the added value of the Kantian morality parameter κ_i against the four other parameters, α_i, β_i, δ_i, and γ_i. We here restrict attention to the estimates based on risk neutrality; we also present some of these analyses allowing for risk aversion in Appendix A4.1.

[25]Table A.6 (panels D and E) shows that the assignment of subjects to types is similar under subjective and rational expectations.

5.1 Aggregate estimations

Table 6 shows the estimates of the finite mixture estimates for the model allowing for distributional preferences (α_k, β_k), Kantian morality (κ_k), and reciprocity (δ_k, γ_k). Including the reciprocity parameters δ_k and γ_k has limited effects on the parameter estimates compared to the pre-registered models in Table 4. In particular, the estimated κ_k is nearly identical for all the types. Also, for most types the parameter estimates of α_k and β_k are nearly identical, and the estimates of δ_k and γ_k are not significantly different from zero. The only major change appears for Type 2 in the two-types model and Type 3 in the three-types model, for which the estimates of both α_k and β_k are larger than those in Table 4, and both the δ_k and γ_k estimates are significantly different from zero.[26] Interestingly, these δ_k and γ_k estimates are all negative.[27] Thus, subjects classified as belonging to these types appear to exhibit *less* behindness aversion when the opponent has 'behaved unkindly' as a first mover, and *less* aheadness aversion when the opponent has 'behaved kindly' as a first mover. Although counter-intuitive, these findings are in line with those of Charness and Rabin (2002) (see their Tables VI and VII and the associated discussion). By contrast, Bruhin et al. (2019) find the conjectured signs for the reciprocity parameter estimates that are significantly different from zero (see their Tables 1 and 2).

To study the value-added of Kantian morality and reciprocity, we compare one-, two-, and three-types models allowing for combinations of distributional preferences (α, β), Kantian morality (κ) and reciprocity (δ,γ). Tables 4 and 6 showed the results for models allowing for (α, β, κ) and $(\alpha, \beta, \kappa, \delta, \gamma)$ respectively. In Tables A.10 and A.11 in Appendix A1 we further report the results for models allowing for only distributional preferences (α, β) and distributional preferences in combination with reciprocity $(\alpha, \beta, \delta, \gamma)$ respectively. Figure 7 shows the ICL scores for these models. Three things become clear from this figure.

[26] Table A.8 in Appendix A1 lists the chosen strategies for each type. Table A.6 shows that the assignment of subjects to types is largely similar with and without reciprocity (panels F and G), or with and without Kantian morality (panels H and I).

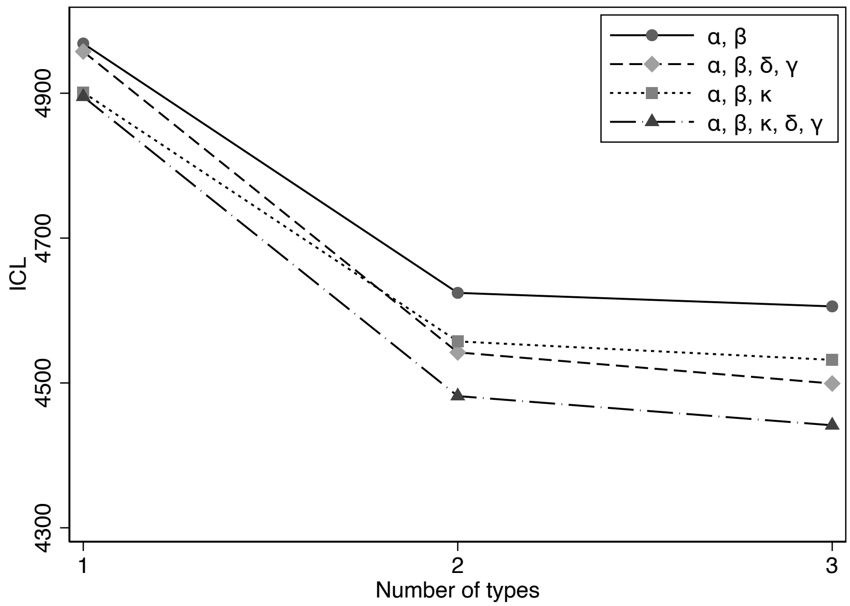
[27] The corresponding estimates that take subjects to be risk averse exhibit this counter-intuitive feature less; see Table A.15 in Appendix A4.1.

Table 6: Estimates at the aggregate level (distributional, Kantian morality, and reciprocity)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.17	0.08	0.27	0.13	0.02	0.24
	(0.02)	(0.03)	(0.06)	(0.06)	(0.04)	(0.06)
β_k	0.28	0.39	0.17	0.34	0.44	0.14
	(0.03)	(0.04)	(0.04)	(0.11)	(0.07)	(0.05)
κ_k	0.10	0.10	0.09	0.09	0.13	0.09
	(0.01)	(0.01)	(0.01)	(0.03)	(0.04)	(0.01)
δ_k	−0.04	0.05	−0.16	0.02	0.05	−0.13
	(0.02)	(0.03)	(0.06)	(0.06)	(0.05)	(0.06)
γ_k	−0.06	−0.07	−0.27	−0.20	0.09	−0.32
	(0.03)	(0.05)	(0.07)	(0.15)	(0.09)	(0.10)
λ_k	7.00	8.14	4.18	7.23	8.74	3.75
	(0.42)	(0.73)	(0.40)	(1.20)	(1.04)	(0.40)
ϕ_k	1.00	0.58	0.42	0.41	0.24	0.35
	(-)	(0.05)	(0.05)	(0.09)	(0.06)	(0.07)
$\ln L$	-2433.5	-2207.7		-2167.4		
$EN(\boldsymbol{\tau})$	0.00	5.10		12.32		
ICL	4895.3	4481.9		4441.6		
NEC	-	0.023		0.046		

Notes: Standard errors in parentheses. Based on our 'core sample' of 112 subjects, Table A.9 in Appendix A1 shows estimates based on all 136 subjects.

Figure 7: ICL scores



Notes: ICL scores of different finite mixture models. Lower ICL scores indicate a more preferred model. Figure based on our 'core sample' of 112 subjects.

First, as lower ICL scores indicate a more preferred model, all multi-types models strongly outperform the one-type (representative agent) models. Second, adding either Kantian morality (κ) or reciprocity (δ, γ) to the pure distributional model (α, β) reduce the ICL scores substantially, showing that both Kantian morality (κ) and reciprocity (δ, γ) improve the fit of the model. Third, Kantian morality (κ) and reciprocity (δ, γ) aren't substitutes: adding both Kantian morality (κ) and reciprocity (δ, γ) to the pure distributional model (α, β) yields a decrease in the ICL score which is approximately twice as large as adding only Kantian morality or only reciprocity.[28]

5.2 Individual estimations

Figure 8 shows the individual parameter estimates when allowing for distributional preferences (α_i, β_i), Kantian morality (κ_i) and reciprocity (δ_i, γ_i). As in the pre-registered model

[28]In the corresponding figure obtained when we allow for risk aversion—see Figure A.6 in Appendix A4.1—both specifications that include Kantian morality (i.e., with and without reciprocity) yield a substantial and similar decrease in the ICL scores compared to the specifications without κ.

without reciprocity, most individual estimates of α_i, β_i, and κ_i are positive. For the reciprocity parameters δ_i and γ_i, we observe considerable heterogeneity, and both negative and positive estimates. Table A.12 in Appendix A1 shows summary statistics.

To study the value-added of the different preference parameters, we consider all models that are nested in (1) and apply standard information criteria.[29] We use both the Bayesian information criterion (BIC) and Akaike's Information Criterion (AIC), each of which is based on the log-likelihoods and adds a penalty for each parameter. The lower score, the better fit. More precisely, the criteria are:

$$BIC = -2\ln(L) + d\ln(18), \tag{11}$$

and

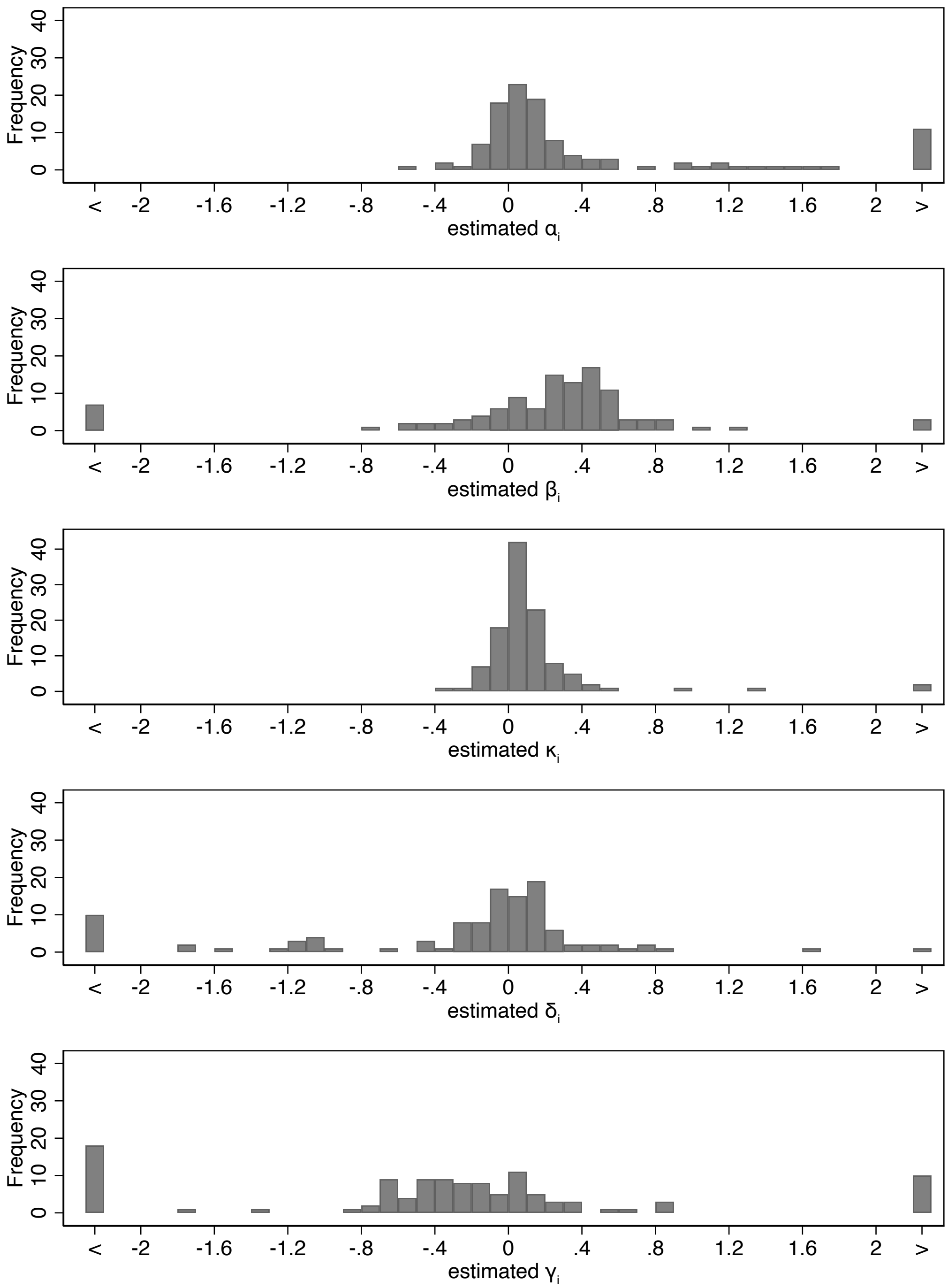
$$AIC = -2\ln(L) + 2d, \tag{12}$$

where $\ln(18)$ in (11) comes from the 18 observations per subject. Since $\ln 18 \approx 2.89 > 2$, BIC gives a heavier penalty per parameter than AIC.

Table 7 shows the results. The left side of the table shows which model provides the best fit according to BIC. For 21 subjects (18.8%) pure self-interest ($\alpha_i = \beta_i = \kappa_i = \delta_i = \gamma_i = 0$) has the lowest BIC score. This contrasts with the aggregate estimates (Table 6), where no purely self-interested type emerges. This difference may be the result of the relatively small number of observations for each individual estimation, giving less power to reject self-interest at the individual level. For the remaining 91 subjects, some combinations of social preferences and/or moral concerns improve the model's fit. In sum, for 21 subjects (18.8%), the model with the lowest BIC score includes κ_i. In comparison, α_i, β_i, δ_i and γ_i are included in the model with the lowest BIC score for 28 subjects (25.0%), 77 subjects (68.8%), 16 subjects (14.3%), and 45 subjects (40.2%) respectively. In particular β_i (aheadness aversion) and γ_i (positive reciprocity) play a big role and improve the fit for 69% and

[29]We only include the reciprocity parameters δ_i and γ_i in combination with α_i and β_i respectively.

Figure 8: Distributions of individual parameter estimates (distributional, morality, and reciprocity)



Note: All estimates of α_i, β_i, κ_i, γ_i, δ_i larger than 2 in absolute value are grouped in bins ("<" and ">") at the extremes of the horizontal axis. Figure based on our 'core sample' of 112 subjects. Figure A.3 in Appendix A1 shows a figure based on all 136 subjects.

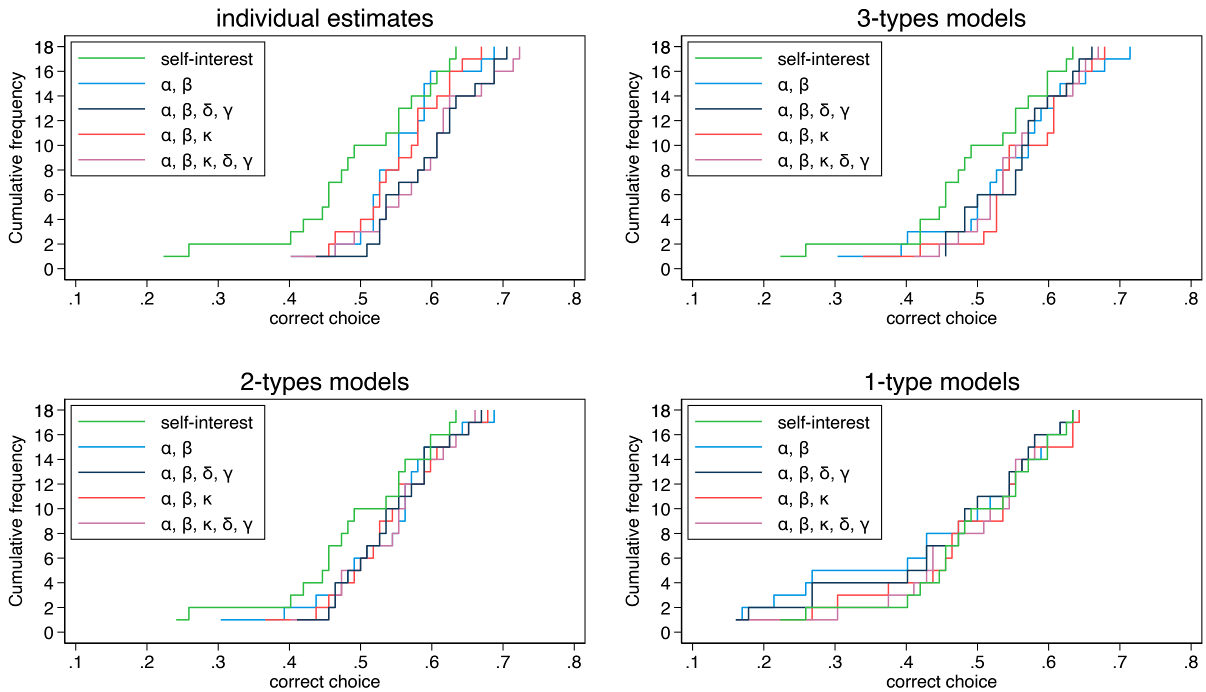
Table 7: Best individual fit

	BIC		AIC	
Parameters	Frequency	Percentage	Frequency	Percentage
$\alpha, \beta, \kappa, \delta, \gamma$			1	0.9
$\alpha, \beta, \kappa, \delta$				
$\alpha, \beta, \kappa, \gamma$			1	0.9
$\alpha, \beta, \delta, \gamma$	5	4.5	10	8.9
α, β, κ	1	0.9	1	0.9
α, β, δ	5	4.5	6	5.4
α, β, γ	6	5.4	9	8.0
α, κ, δ	3	2.7	4	3.6
β, κ, γ	5	4.5	5	4.5
α, β	2	1.8	1	0.9
α, κ	2	1.8	4	3.6
α, δ	3	2.7	3	2.7
β, κ	5	4.5	4	3.6
β, γ	29	25.9	27	24.1
α	1	0.9		
β	19	17.0	17	15.2
κ	5	4.5	5	4.5
-	21	18.8	14	12.5
Selected model includes:				
Parameter	Frequency	Percentage	Frequency	Percentage
α_i	28	25.0	40	35.7
β_i	77	68.8	82	73.2
κ_i	21	18.8	25	22.3
δ_i	16	14.3	24	21.4
γ_i	45	40.2	53	47.3

Notes: Entries in the top panel indicate the number of subjects for whom the specific model provides the lowest BIC or AIC score respectively. Entries in the bottom panel summarize how frequently a parameter was included in the the model the lowest BIC or AIC score respectively. Table based on our 'core sample' of 112 subjects.

40% of subjects respectively, while the value-added of each of the other three preference parameters is roughly in the same ballpark. The right side of Table 7 shows the results from the same exercise, but now applied to AIC. Then the best-fitting model at the individual level includes the parameter κ_i for 25 subjects (or 22.3%). Again, a smaller number of subjects than for β_i (82 subjects, or 73.2%) and γ_i (53 subjects, or 47.3%), but a number of subjects closer to α_i (40 subjects, or 35.7%) and also δ_i (24 subjects, or 21.4%).

Figure 9: Accuracy of out-of-sample predictions



Notes: Accuracy of out-of-sample predictions, based on individual estimates (top left panel) and finite mixture models with three-types, two-types, or a representative agent (1 type). Plots show cumulative frequency plots for the average fraction of correctly predicted choices per game protocol. Figure based on our 'core sample' of 112 subjects.

5.3 Out-of-sample predictions

So far, we evaluated the performance of different models based on information criteria. As an alternative, we consider the predictive accuracy of different models by conducting out-of-sample predictions. For each of the 18 game protocols, we estimate parameters based on the other 17 game protocols, and use the estimates to predict the choice for the one omitted game protocol. We conduct these analyses both at the individual level and the aggregate level.

Figure 9 illustrates the results, by comparing the predictive accuracy of models allowing for distributional preferences (α,β), distributional preferences in combination with either reciprocity ($\alpha,\beta,\gamma,\delta$) or Kantian morality ($\alpha,\beta,\kappa$) or both ($\alpha,\beta,\kappa,\gamma,\delta$).[30] The top

[30]In Table A.13 in Appendix A1 we list the average predictive accuracy for each of the models included in

left panel of Figure 9 compares the predictive accuracy based on individual estimates. All models clearly outperform random choice (which would lead to 20.8% accurate predictions in expectation). All models allowing for distributional preferences perform much better than when assuming self-interest, but the differences in predictive accuracy between these models are small. On average, the (α, β, κ)-model on average predicts 55.5% of choices correctly, somewhat more than the (α, β)-model which predicts 55.2% of choices correctly, and less than the (α, β, δ, γ)- and (α, β, κ, δ, γ)-models, which give 59.1% and 58.8% average accuracy, respectively. All models allowing for distributional preferences, reciprocity and/or Kantian morality perform much better than when assuming self-interest, which gives 48.8% average accuracy.

The other panels of Figure 9 show the predictive accuracy of finite mixture models. The bottom right panel of Figure 9 shows that assuming a representative agent (1 type) leads to much lower predictive accuracy. On average, the models assuming a representative agent achieve between 44.0% and 48.8% accuracy, much below the accuracy of the models allowing for individual estimates. The three-types and two-types models however, perform much better. As for the individual estimates, the two-types and three-types models with distributional preferences, reciprocity and/or Kantian morality outperform self-interest, but the differences between these models are modest. For the two-types models that go beyond self-interest, average accuracy is between 53.2% and 54.4%, while for the three-types models the range is 54.2% to 55.8%. This provides further evidence that the two-types model effectively captures the heterogeneity in preferences.

In sum, allowing for distributional preferences substantially improves the predictive accuracy over self-interest. However,the value-added of Kantian morality and reciprocity over distributional preferences is limited in the out-of-sample predictions. This contrasts with the improved within-sample fit when allowing for Kantian morality and reciprocity that we observed in Figure 7.[31]

Figure 9.

[31] In the out-of-sample predictions obtained when we allow for risk aversion—see Figure A.7 in Appendix

6 Concluding discussion

In this paper, we report results from a laboratory experiment designed to evaluate the explanatory power of Kantian morality in standard strategic interactions. To distinguish Kantian morality from other social concerns, we posit a general utility function that nests several much studied preference classes, such as pure self-interest, altruism, spite, inequity aversion, and reciprocity, and of course Kantian morality. We structurally estimate the preference parameters of this utility function controlling for the beliefs about opponent's play. We obtain both individual and aggregate estimates, where the latter consists of estimating the parameters for a representative agent, as well as identifying a small number of endogenously determined "preference types".

The individual estimates suggest substantial heterogeneity. This heterogeneity limits the usefulness of a representative agent approach. However, we find that the subjects' behaviors are well captured by models with two or three preference types. The two-types model suggests that roughly 60% of subjects display a combination of inequity aversion with Kantian morality, and the remaining share a combination of Kantian morality and behindness aversion. Quite remarkably, however, all the preference types—both the representative agent and the preference types within the two-types and the three-types model—have an estimated Kantian morality parameter κ of around 0.1. The finding that all types have a positive κ also holds when we allow for reciprocity, risk aversion, or rational expectations.

Compared with other experimental studies with structural preference estimations, our results agree with those of Bruhin et al. (2019) in that their behavioral data is largely consistent with there being a small number of "preference types". Our findings further agree with Bruhin et al. (2019) in that they do not either find evidence that the purely selfish *Homo oeconomicus* explains their behavioral data. A more detailed comparison is more in-

A4.1—the specifications that include Kantian morality (with and without reciprocity) perform substantially better than the other specifications in the one- and two-types models. Note that among all the specifications that differ from pure self-interest, these are the only two that make substantially better out-of-sample predictions than the others.

volved, since their experimental design differs from ours, and they do not include Kantian morality. Our results further agree broadly with those in the horse race study by Miettinen et al. (2020), although our richer data set allows us to capture the complex combination of subjects' motives that their study cannot address.

Our experimental design was motivated by findings in the theoretical literature that investigates the evolutionary foundations of preferences in strategic interactions (see Alger & Weibull, 2019; Alger, 2023, for recent surveys). Interestingly, our findings are in line with the theoretical prediction that evolution by natural selection favors preferences that combine not only self-interest and Kantian morality, but also either altruism or spite, when preferences are expressed at the level of material payoffs (Alger et al., 2020).[32] Indeed, our finite mixture estimates show that essentially all types combine self-interest, Kantian morality, and some concern for the other's payoff.

However, our analysis also reveals some intriguing findings. The estimated attitude towards being ahead materially is qualitatively different in the estimates that assume risk neutrality and those that assume risk aversion: while all types are either indifferent to other's payoff or altruistic towards the other when ahead under risk neutrality, a sizeable share of the subjects are classified as being spiteful when ahead under risk aversion. Furthermore, when reciprocity parameters are included in the posited utility function, all the estimates have counter-intuitive signs when we impose risk neutrality, but this is not the case when subjects are taken to be risk averse. These results clearly beg for further research.

Our posited utility function is richer than most examined before: in addition to Kantian morality, it allows for altruism, spite, inequity aversion, and reciprocity. As is the case for all other similar studies, it could be, however, that other motivations not included in the posited utility function drive (part) of the behavior. For future research, it would

[32] This result does not contradict that of Alger and Weibull (2013), according to which evolution by natural selection favors a convex combination between self-interest and Kantian morality. Indeed, Alger et al. (2020) confirm in their model that such preferences are indeed favored by evolution when it is own and other's reproductive success that appear as arguments in the utility function, rather than (trivial) material payoffs.

be interesting to study the value added of Kantian morality compared to other motivations like guilt aversion and image concerns. It would further be interesting to examine whether results similar to ours also obtain in a representative sample, along the lines of the studies by Bellemare et al. (2008) and Cettolin and Suetens (2018). While evolutionary theory suggests that the qualitative nature of preferences guiding behavior in strategic interactions should be similar across the world, certain differences between populations may be expected to influence the relative importance of self-interest, social concerns, and Kantian morality. In particular, since evolutionary theory suggests that migration patterns and the involvement in inter-group conflict are expected to impact preferences guiding behavior in strategic interactions (Alger et al., 2020; Choi & Bowles, 2007), this theory delivers testable predictions that may help explain cross-cultural differences (Falk et al., 2018) and also perhaps differences between men and women (Croson & Gneezy, 2009). Finally, it would be interesting to investigate more precisely whether Kantian morality can help explain the formation of social norms (Bicchieri, 2005; Krupka & Weber, 2013; Elster, 1989), as well as the documented enhancement of pro-social behaviors triggered by role uncertainty (Iriberri & Rey-Biel, 2011). Related to the last issue, our experimental design is adapted to detect *Homo moralis* preferences in *ex ante* symmetric situations, because the current theoretical models define these preferences in such settings; future work may reveal fruitful ways to formalize, and also test, a similar form of Kantian morality in asymmetric settings.

References

Alger, I. (2023). Evolutionarily stable preferences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *378*(1876), 20210505.

Alger, I., & Laslier, J.-F. (2022). Homo moralis goes to the voting booth: Coordination and information aggregation. *Journal of Theoretical Politics*, *0*(0), 09516298221081811.

Alger, I., & Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, *81*(6), 2269–2302.

Alger, I., & Weibull, J. W. (2017). Strategic behavior of moralists and altruists. *Games*, *8*(3).

Alger, I., & Weibull, J. W. (2019). Evolutionary models of preference formation. *Annual Review of Economics*, *11*, 329–354.

Alger, I., Weibull, J. W., & Lehmann, L. (2020). Evolution of preferences in structured populations: genes, guns, and culture. *Journal of Economic Theory*, *185*(104951).

Andersson, O., Holm, H. J., Tyran, J.-R., & Wengström, E. (2014). Deciding for others reduces loss aversion. *Management Science*, *62*(1), 29–36.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *Economic Journal*, *100*(401), 464–477.

Andreoni, J., & Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753.

Apesteguia, J., & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, *126*(1), 74–106.

Bardsley, N., & Moffatt, P. G. (2007). The Experimetrics of Public Goods: Inferring Motivations from Contributions. *Theory and Decision*, *62*(2), 161–193. doi: 10.1007/s11238-006-9013-3

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, *97*(2), 170-176. doi: 10.1257/aer.97.2.170

Becker, G. S. (1974). A theory of social interactions. *Journal of Political Economy*, *82*(6), 1063–1093.

Bellemare, C., Kröger, S., & Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica*, *76*(4), 815-839.

Bénabou, R., Falk, A., Henkel, L., & Tirole, J. (2020). *Eliciting moral preferences: theory and experiment* (mimeo). Toulouse School of Economics.

Bénabou, R., & Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, *96*(5), 1652-1678.

Bicchieri, C. (2005). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press. doi: 10.1017/CBO9780511616037

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, *22*(7), 719–725.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior*, *72*(2), 321–338.

Bolton, G. E., & Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American Economic Review*, *90*(1), 166–193.

Bomze, I., Schachinger, W., & Weibull, J. (2021). Does moral play equilibrate? *Economic Theory*, *71*(1), 305–315. Retrieved from https://doi.org/10.1007/s00199-020-01246-4 doi: 10.1007/s00199-020-01246-4

Breitmoser, Y. (2013). Estimation of social preferences in generalized dictator games. *Economics Letters*, *121*(2), 192–197.

Brock, J. M., Lange, A., & Ozbay, E. Y. (2013). Dictating the risk: Experimental evidence on giving in risky environments. *American Economic Review*, *103*(1), 415–37.

Bruhin, A., Fehr, E., & Schunk, D. (2019). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, *17*(4), 1025–1069.

Cappelen, A. W., Konow, J., Sørensen, E. Ø., & Tungodden, B. (2013). Just luck: An experimental study of risk-taking and fairness. *American Economic Review*, *103*(4), 1398–1413.

Capraro, V., & Rand, D. G. (2018). Do the Right Thing: Preferences for Moral Behavior, Rather Than Equity or Efficiency per se, Drive Human Prosociality. *Judgment and Decision Making*, *13*(1), 99–111.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*(2), 195–212.

Cettolin, E., & Suetens, S. (2018). Return on trust is lower for immigrants. *Economic Journal*, *129*(621), 1992-2009.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, *117*(3), 817–869.

Cherubini, U., Luciano, E., & Vecchiato, W. (2004). *Copula methods in finance*. John Wiley & Sons.

Choi, J.-K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science*, *318*(5850), 636–640.

Cooper, D. J., & Kagel, J. H. (2015). Other-regarding preferences: A selective survey of experimental results. In *The Handbook of Experimental Economics, Volume 2* (pp. 217–289). Princeton University Press.

Croson, R. (2000). Thinking like a game theorist: factors affecting the frequency of equilibrium play. *Journal of economic behavior & organization*, *41*(3), 299–314.

Croson, R., & Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, *47*(2), 448–474.

Daley, B., & Sadowski, P. (2017). Magical thinking: A representation result. *Theoretical Economics*, *12*(2), 909-956.

Daube, M., & Ulph, D. (2016). Moral Behaviour, Altruism and Environmental Policy. *Environmental and Resource Economics*, *63*(2), 505–522. Retrieved from https://doi.org/10.1007/s10640-014-9836-2 doi: 10.1007/s10640-014-9836-2

DellaVigna, S. (2018). Structural Behavioral Economics. In D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of Behavioral Economics* (p. 613-723). New York: Elsevier.

DellaVigna, S., List, J. A., & Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, *127*(1), 1–56.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic behavior*, *47*(2), 268–298.

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295.

Eichner, T., & Pethig, R. (2021). Climate policy and moral consumers*. *The Scandinavian Journal of Economics*, *123*(4), 1190-1226. doi: https://doi.org/10.1111/sjoe.12450

Eichner, T., & Pethig, R. (2022). Kantians defy the economists' mantra of uniform pigovian emissions taxes. *Ecological Economics*, *200*, 107514.

Ellingsen, T., & Johannesson, M. (2008). Pride and prejudice: The human side of incentive theory. *American Economic Review*, *98*(3), 990-1008.

Elster, J. (1989). Social norms and economic theory. *Journal of economic perspectives*, *3*(4), 99–117.

Engelmann, D. (2012). How not to extend models of inequality aversion. *Journal of Economic Behavior & Organization*, *81*(2), 599–605.

Engelmann, D., & Strobel, M. (2004). Inequality aversion, efficiency, and maximin preferences in simple distribution experiments. *American Economic Review*, *94*(4), 857-869.

Exley, C. L. (2016). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies*, *83*(2), 587–628.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., & Sunde, U. (2018). Global evidence on economic preferences. *Quarterly Journal of Economics*, *133*(4), 1645-1692.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*(2), 293–315.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, *114*(3), 817–868.

Fisman, B. R., Kariv, S., & Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, *97*(5), 1858–1876.

Gächter, S., & Renner, E. (2010). The effects of (incentivized) belief elicitation in public goods experiments. *Experimental Economics*, *13*(3), 364–377.

Gauriot, R., Heger, S. A., & Slonim, R. (2020). Altruism or diminishing marginal utility? *Journal of Economic Behavior & Organization*, *180*, 24–48.

Iriberri, N., & Rey-Biel, P. (2011). The role of role uncertainty in modified dictator games. *Experimental Economics*, *14*(2), 160–180.

Iriberri, N., & Rey-Biel, P. (2013). Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do? *Quantitative Economics*, *4*(3), 515–547.

Joe, H., & Xu, J. J. (1996). *The estimation method of inference functions for margins for multivariate models* (Technical Report No. 166). Department of Statistics, University of British Columbia.

Krawczyk, M., & Le Lec, F. (2010). 'Give me a chance!' An experiment in social decision under risk. *Experimental Economics*, *13*(4), 500–511.

Krawczyk, M., & Le Lec, F. (2016). Dictating the risk: experimental evidence on giving in risky environments: comment. *American Economic Review*, *106*(3), 836–39.

Krupka, E. L., & Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, *11*(3), 495–524.

Laffont, J.-J. (1975). Macroeconomic constraints, economic efficiency and ethics: An introduction to kantian economics. *Economica*, *42*(168), 430–437.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (p. 105–142). New York: Academic Press.

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, *6*(1), 355-378.

Miettinen, T., Kosfeld, M., Fehr, E., & Weibull, J. W. (2020). Revealed preferences in a sequential prisoners' dilemma: a horse-race between six utility functions. *Journal of*

Economic Behavior and Organization, *173*, 1-25.

Nunnari, S., & Pozzi, M. (2022). *Meta-analysis of inequality aversion estimates* (mimeo).

Ottoni-Wilhelm, M., Vesterlund, L., & Xie, H. (2017). Why do people give? testing pure and impure altruism. *American Economic Review*, *107*(11), 3617–33.

Palfrey, T. R., & Prisbrey, J. E. (1997). Anomalous behavior in public goods experiments: how much and why? *American Economic Review*, *87*, 829–846.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review*, 1281–1302.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica*, *68*(5), 1281–1292.

Rabin, M., & Thaler, R. H. (2001). Anomalies: Risk aversion. *Journal of Economic Perspectives*, *15*(1), 219–232.

Ross, L., Greene, D., & House, P. (1977). The “false consensus effect”: an egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279-301.

Roth, B., & Voskort, A. (2014). Stereotypes and false consensus: How financial professionals predict risk preferences. *Journal of Economic Behavior & Organization*, *107*, 553-565.

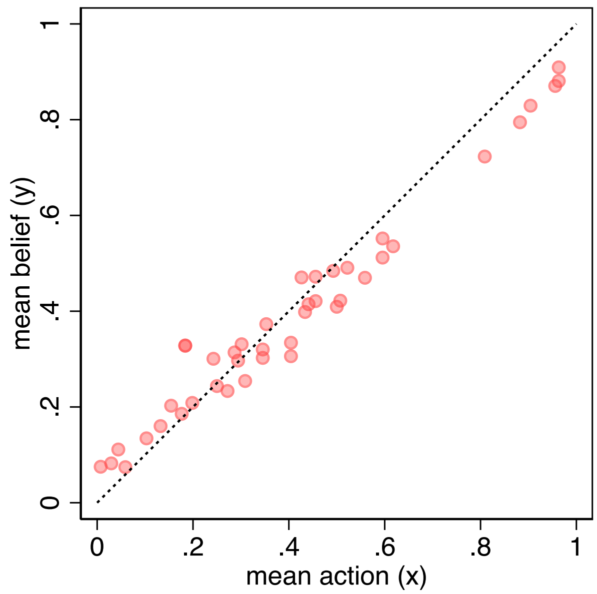
Sarkisian, R. (2017). Team incentives under moral and altruistic preferences: Which team to choose? *Games*, *8*(3). doi: 10.3390/g8030037

Wilcox, N. T., & Feltovich, N. (2000). *Thinking like a game theorist: Comment* (mimeo).

Appendices (For Online Publication)

Appendix A1 Additional tables and figures

Figure A.1: Correlations between mean actions and beliefs



Note: Each dot represents the mean action x and mean stated belief y for each of the actions (listed in Table 1). Means are taken across all 136 subjects.

Table A.1: Lottery choices

Lottery	Outcomes A	Outcomes B	Frequency	Percentage	r_i
Sessions 2-8					
1	18	18	50	43.9%	1.61
2	22	15	24	21.1%	1.00
3	26	12	18	15.8%	0.39
4	30	9	3	2.6%	0.25
5	34	6	8	7.0%	0.08
6	37	2	11	9.7%	-0.09
Session 1					
1	18	18	5	22.7%	4.71
2	22	16	3	13.6%	2.95
3	26	14	6	27.3%	1.19
4	30	12	4	18.2%	0.77
5	34	10	2	9.1%	0.32
6	40	4	2	9.1%	-0.13

Notes: Lottery choices in the Eckel and Grossman (2002) risk elicitation task. 'Outcomes' are the payoffs denoted in "points", see Appendix A7 for the instructions. The final column lists the implied r_i parameters for each lottery choice. Note that after the first session, we slightly adjusted the outcomes to better estimate r_i. Table based on all 136 subjects.

Table A.2: Behavioral predictions (more types)

No.	T	R	P	S	strong altruism $\alpha = -0.5$ $\beta = 0.5$ $\delta = 0$ $\gamma = 0$ $\kappa = 0$	strong morality $\alpha = 0$ $\beta = 0$ $\delta = 0$ $\gamma = 0$ $\kappa = 0.5$	intermediate altruism $\alpha = -0.2$ $\beta = 0.2$ $\delta = 0$ $\gamma = 0$ $\kappa = 0$	intermediate morality $\alpha = 0$ $\beta = 0$ $\delta = 0$ $\gamma = 0$ $\kappa = 0.2$	moderate altruism $\alpha = -0.1$ $\beta = 0.1$ $\delta = 0$ $\gamma = 0$ $\kappa = 0$	moderate morality $\alpha = 0$ $\beta = 0$ $\delta = 0$ $\gamma = 0$ $\kappa = 0.1$
Sequential Prisoner's Dilemmas										
1	90	45	15	10	(C,D,C)	(D,D,C)	(C,D,C)	(D,D,C)	(C,D,C)	(D,D,C)
2	90	55	20	10	(C,C,C)	(C,C,D)	(C,D,C)	(C,D,D)	(C,D,D)	(C,D,D)
3	80	65	25	20	(C,C,C)	(C,C,D)	(C,D,C)	(C,C,D)	(C,D,C)	(C,D,D)
4	90	65	25	10	(C,C,C)	(C,C,D)	(C,D,C)	(C,C,D)	(C,D,D)	(C,D,D)
5	80	75	30	20	(C,C,C)	(C,C,D)	(C,C,C)	(C,C,D)	(C,C,D)	(C,C,D)
6	90	75	30	10	(C,C,C)	(C,C,D)	(C,C,D)	(C,C,D)	(C,D,D)	(C,C,D)
Trust Games										
7	80	50	30	20	(I,G)	(I,K)	(I,K)	(I,K)	(I,K)	(I,K)
8	90	50	30	10	(I,G)	(I,K)	(I,K)	(N,K)	(N,K)	(N,K)
9	80	60	30	20	(I,G)	(I,G)	(I,K)	(I,K)	(I,K)	(I,K)
10	90	60	30	10	(I,G)	(I,G)	(I,K)	(I,K)	(N,K)	(N,K)
11	80	70	30	20	(I,G)	(I,G)	(I,G)	(I,G)	(I,K)	(I,K)
12	90	70	30	10	(I,G)	(I,G)	(I,K)	(I,G)	(I,K)	(I,K)
Ultimatum Games										
13	60	50	40	10	(E,A)	(U,A)	(U,A)	(U,A)	(U,A)	(U,A)
14	65	50	35	10	(E,A)	(U,A)	(U,A)	(U,A)	(U,A)	(U,A)
15	70	50	30	10	(E,A)	(U,A)	(U,A)	(U,A)	(U,A)	(U,A)
16	75	50	25	10	(E,A)	(U,A)	(U,A)	(U,A)	(U,A)	(U,A)
17	80	50	20	10	(E,A)	(U,A)	(U,A)	(U,A)	(U,A)	(U,A)
18	85	50	15	10	(E,A)	(U,A)	(U,A)	(U,A)	(U,A)	(U,A)

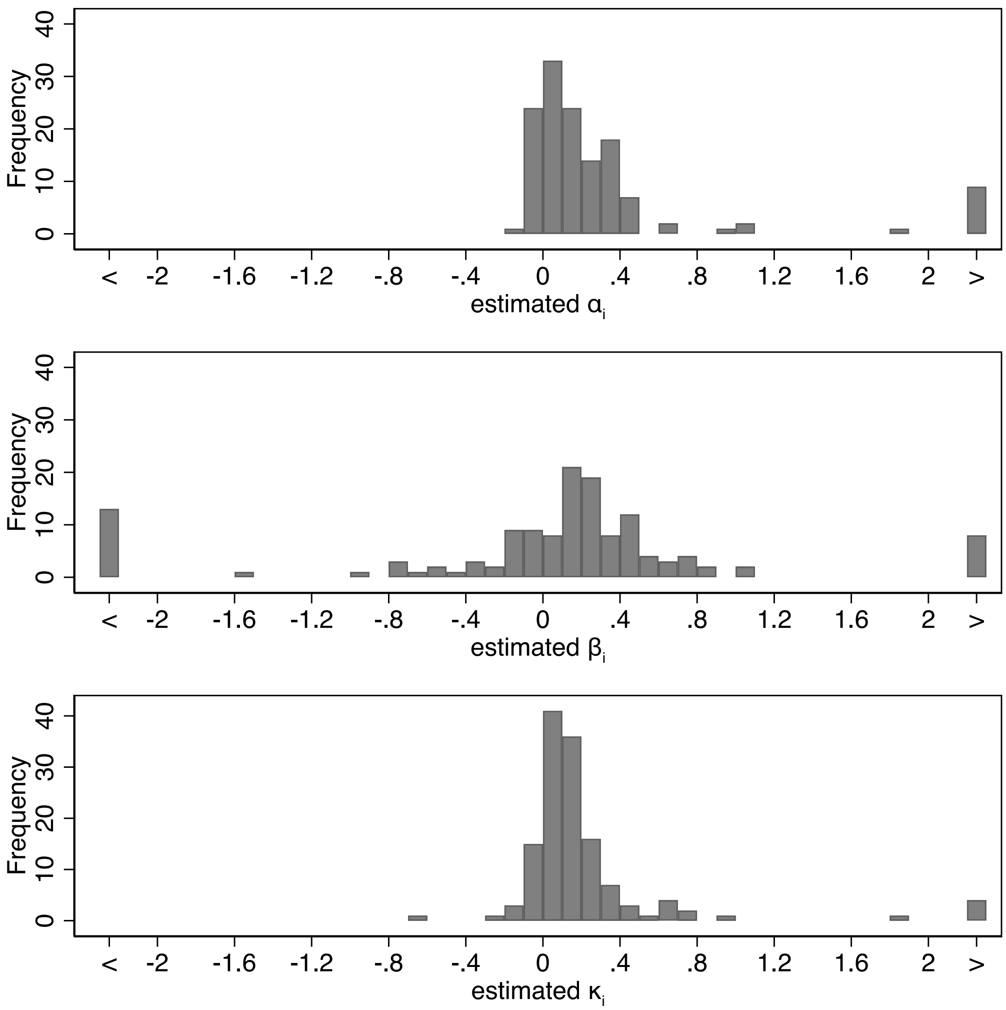
Notes: Predicted behavioral strategies, assuming rational expectations (see Table 1 for average play in each game protocol).

Table A.3: Individual parameter estimates (all subjects)

Parameter	Median	Mean	S.D.	Min	Max
α_i	0.14	1964.71	11769.12	−0.19	89924.06
β_i	0.17	−664.37	5600.44	−54498.23	10936.72
κ_i	0.11	963.47	5610.93	−0.68	39590.11

Notes: Table based on estimates from all 136 subjects.

Figure A.2: Distributions of individual parameter estimates (all subjects)



Note: All estimates of α_i, β_i and κ_i larger than 2 in absolute value are grouped in bins ("<" and ">") at the extremes of the horizontal axis. Figure based on all 136 subjects.

Table A.4: Estimates at the aggregate level (all subjects)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.18	0.15	0.17	0.21	0.03	0.17
	(0.01)	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)
β_k	0.25	0.37	−0.02	0.26	0.52	−0.03
	(0.02)	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)
κ_k	0.11	0.12	0.09	0.12	0.16	0.09
	(0.01)	(0.01)	(0.01)	(0.01)	(0.04)	(0.02)
λ_k	7.89	9.68	3.74	10.40	6.52	3.59
	(0.46)	(0.75)	(0.51)	(0.96)	(0.68)	(0.29)
ϕ_k	1.00	0.63	0.37	0.49	0.17	0.34
	(-)	(0.06)	(0.06)	(0.06)	(0.04)	(0.04)
$\ln L$	-3026.9	-2762.6		-2706.6		
$EN(\boldsymbol{\tau})$	0.00	6.06		13.64		
ICL	6073.4	5575.5		5495.6		
NEC	-	0.023		0.043		

Notes: Standard errors in parentheses. Table based on all 136 subjects.

Table A.5: The 4-types model

	Type 1	Type 2	Type 3	Type 4
α_k	0.17	0.16	0.02	0.16
	(0.03)	(0.03)	(0.03)	(0.06)
β_k	−0.01	0.17	0.46	0.51
	(0.04)	(0.05)	(0.06)	(0.08)
κ_k	0.09	0.12	0.15	0.04
	(0.02)	(0.01)	(0.05)	(0.03)
λ_k	3.63	8.66	8.56	4.67
	(0.29)	(1.54)	(1.12)	(0.61)
ϕ_k	0.34	0.35	0.19	0.12
	(0.05)	(0.07)	(0.06)	(0.05)
$\ln L$	-2212.6			
$EN(\boldsymbol{\tau})$	23.45			
ICL	4538.3			
NEC	0.103			

Notes: Standard errors in parentheses. Estimation results from models with 1, 2 and 3 types can be found in Table 4. Based on our 'core sample' of 112 subjects.

Table A.6: Transitions between types

Panel A: 2 types and 3 types

3 types	2 types Type 1	Type 2
Type 1	52	3
Type 2	17	0
Type 3	0	40

Panel B: 2 types, ln and risk neutral

2 types ($r = 0$)	2 types ($r = 1$) Type 1	Type 2
Type 1	63	6
Type 2	2	41

Panel C: 3 types, ln and risk neutral

3 types ($r = 0$)	3 types ($r = 1$) Type 1	Type 2	Type 3
Type 1	19	28	8
Type 2	13	4	0
Type 3	0	2	38

Panel D: 2 types, subjective and rational expectations

2 types (subj. exp.)	2 types (rational exp.) Type 1	Type 2
Type 1	59	10
Type 2	4	39

Panel E: 3 types, subjective and rational expectations

3 types (subj. exp.)	3 types (rational exp.) Type 1	Type 2	Type 3
Type 1	35	10	10
Type 2	3	14	0
Type 3	3	0	37

Panel F: 2 types, with and without reciprocity (δ, γ)

2 types (α, β, κ)	2 types ($\alpha, \beta, \kappa, \delta, \gamma$) Type 1	Type 2
Type 1	62	7
Type 2	2	41

Panel G: 3 types, with and without reciprocity (δ, γ)

3 types (α, β, κ)	3 types ($\alpha, \beta, \kappa, \delta, \gamma$) Type 1	Type 2	Type 3
Type 1	42	10	3
Type 2	0	17	0
Type 3	3	0	37

Panel H: 2 types, with and without Kantian morality (κ)

2 types ($\alpha, \beta, \delta, \gamma$)	2 types ($\alpha, \beta, \kappa, \delta, \gamma$) Type 1	Type 2
Type 1	64	8
Type 2	0	48

Panel I: 3 types, with and without Kantian morality (κ)

3 types ($\alpha, \beta, \delta, \gamma$)	3 types ($\alpha, \beta, \kappa, \delta, \gamma$) Type 1	Type 2	Type 3
Type 1	44	0	1
Type 2	1	27	0
Type 3	0	0	39

Notes: Each panel shows transition matrices between types in different finite mixture models. Subjects are assigned a type based on the posterior probability $\tau_{i,k}$ (that subject i belongs to type k, see eq. (7)).

Table A.7: Strategies by type (distributional pref., and Kantian morality)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
Sequential Prisoner's Dilemmas						
C,C,C	2%	3%	0%	2%	5%	0%
C,C,D	21%	30%	5%	20%	61%	5%
C,D,C	0%	1%	0%	1%	0%	0%
C,D,D	9%	10%	7%	12%	6%	7%
D,C,C	1%	2%	0%	2%	5%	0%
D,C,D	7%	10%	2%	11%	8%	2%
D,D,C	3%	3%	2%	2%	6%	2%
D,D,D	57%	40%	84%	51%	10%	84%
Trust Games						
I,G	28%	43%	5%	31%	78%	4%
I,K	16%	21%	9%	23%	11%	10%
N,G	4%	6%	3%	8%	1%	2%
N,K	51%	30%	83%	39%	10%	85%
Ultimatum Games						
E,A	42%	56%	21%	54%	59%	20%
E,F	8%	10%	3%	10%	11%	3%
U,A	50%	34%	75%	36%	30%	77%
U,F	0%	0%	0%	0%	0%	0%

Notes: Relative frequencies (in %) of chosen strategies based on the 1, 2, and three-types models reported in Table 4. Subjects are assigned a type based on the type posterior probability $\tau_{i,k}$ (that subject i belongs to type k, see eq. (7)).

Table A.8: Strategies by type (distributional pref., Kantian morality, and reciprocity)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
Sequential Prisoner's Dilemmas						
C, C, C	2%	3%	0%	1%	6%	0%
C, C, D	21%	33%	5%	18%	51%	3%
C, D, C	0%	1%	0%	1%	1%	0%
C, D, D	9%	12%	5%	15%	6%	5%
D, C, C	1%	3%	0%	1%	4%	0%
D, C, D	7%	10%	3%	9%	12%	2%
D, D, C	3%	3%	2%	2%	5%	2%
D, D, D	57%	35%	85%	53%	15%	88%
Trust Games						
I, G	28%	46%	5%	24%	73%	3%
I, K	16%	22%	8%	28%	10%	8%
N, G	4%	6%	3%	6%	5%	2%
N, K	51%	26%	84%	42%	11%	88%
Ultimatum Games						
E, A	42%	55%	26%	51%	57%	22%
E, F	8%	11%	3%	9%	13%	3%
U, A	50%	34%	71%	40%	30%	74%
U, F	0%	0%	0%	0%	0%	0%

Notes: Relative frequencies (in %) of chosen strategies based on the 1, 2, and three-types models reported in Table 6. Subjects are assigned a type based on the type posterior probability $\tau_{i,k}$ (that subject i belongs to type k, see eq. (7)).

Table A.9: Estimates at the aggregate level (distributional, morality, and reciprocity; all subjects)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.18	0.11	0.24	0.13	0.03	0.24
	(0.02)	(0.02)	(0.05)	(0.05)	(0.04)	(0.04)
β_k	0.29	0.42	0.11	0.37	0.51	0.10
	(0.02)	(0.04)	(0.05)	(0.05)	(0.08)	(0.06)
κ_k	0.10	0.10	0.09	0.09	0.13	0.09
	(0.01)	(0.02)	(0.01)	(0.02)	(0.04)	(0.02)
δ_k	−0.04	0.04	−0.13	0.03	0.08	−0.13
	(0.02)	(0.03)	(0.05)	(0.04)	(0.05)	(0.03)
γ_k	−0.06	−0.10	−0.35	−0.24	0.16	−0.35
	(0.03)	(0.04)	(0.10)	(0.10)	(0.30)	(0.35)
λ_k	7.69	8.86	3.89	8.09	9.90	3.79
	(0.46)	(0.59)	(0.40)	(0.81)	(1.30)	(0.37)
ϕ_k	1.00	0.64	0.36	0.42	0.24	0.34
	(-)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)
$\ln L$	-3018.3	-2713.0		-2640.4		
$EN(\boldsymbol{\tau})$	0.00	4.56		12.72		
ICL	6066.1	5494.5		5391.8		
NEC	-	0.015		0.034		

Notes: Standard errors in parentheses. Based on all 136 subjects.

Table A.10: Estimates at the aggregate level (distributional preferences)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.06	0.03	0.06	0.03	0.03	0.06
	(0.01)	(0.02)	(0.01)	(0.05)	(0.04)	(0.01)
β_k	0.32	0.44	0.08	0.28	0.57	0.08
	(0.02)	(0.03)	(0.04)	(0.04)	(0.04)	(0.04)
λ_k	6.98	8.09	3.75	8.23	6.84	3.32
	(0.46)	(0.68)	(0.53)	(2.16)	(0.95)	(0.32)
ϕ_k	1.00	0.61	0.39	0.37	0.29	0.34
	(-)	(0.07)	(0.07)	(0.09)	(0.06)	(0.06)
$\ln L$	-2477.2	-2292.5		-2267.1		
$EN(\boldsymbol{\tau})$	0.00	6.38		19.66		
ICL	4968.6	4624.4		4605.7		
NEC	-	0.035		0.094		

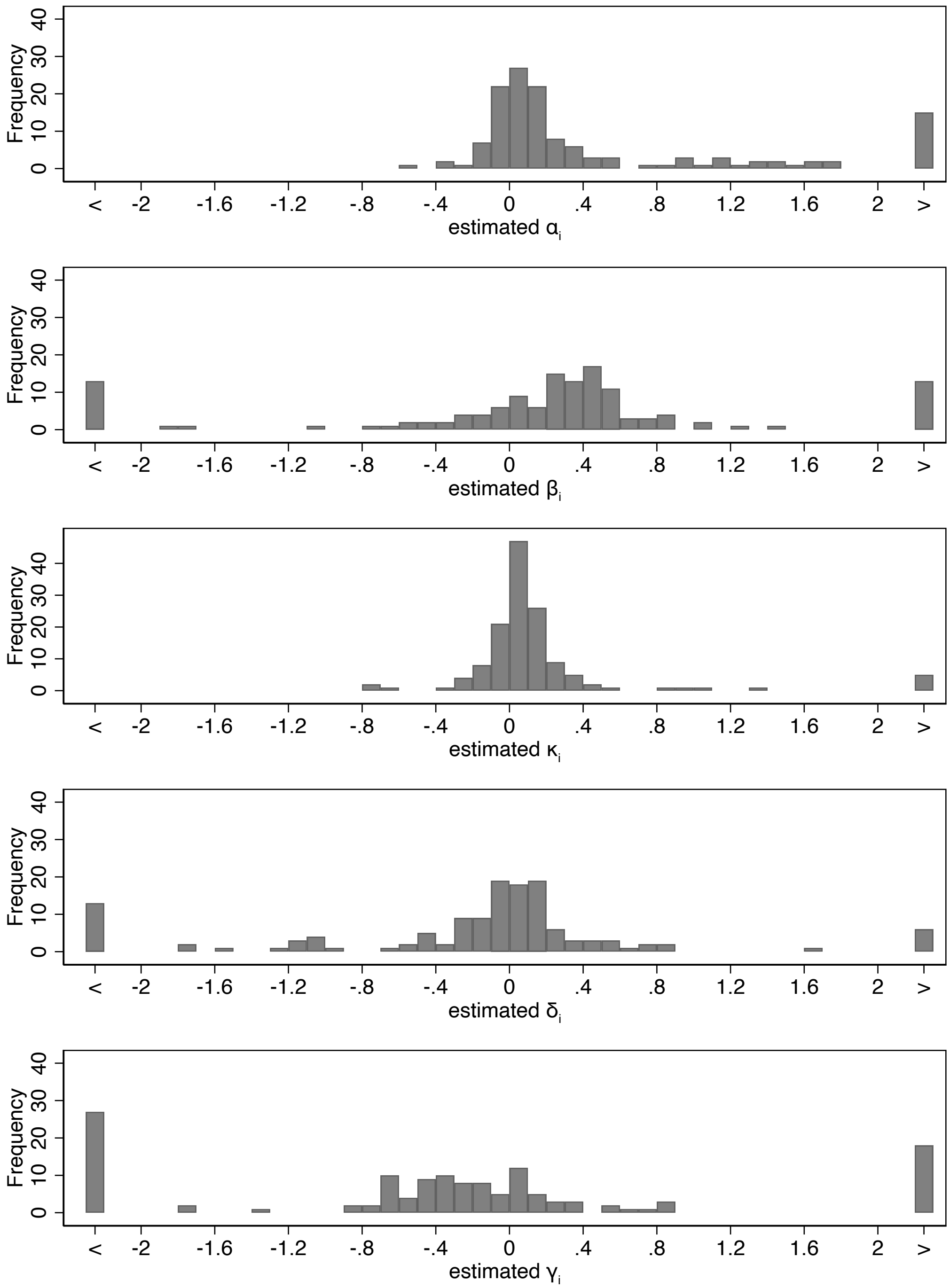
Notes: Standard errors in parentheses. Based on our 'core sample' of 112 subjects.

Table A.11: Estimates at the aggregate level (distributional, reciprocity)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.07	0.00	0.15	0.05	−0.08	0.14
	(0.02)	(0.03)	(0.05)	(0.04)	(0.05)	(0.05)
β_k	0.37	0.47	0.20	0.42	0.53	0.19
	(0.02)	(0.04)	(0.04)	(0.07)	(0.05)	(0.05)
δ_k	−0.02	0.06	−0.13	0.03	0.10	−0.13
	(0.02)	(0.03)	(0.05)	(0.05)	(0.05)	(0.05)
γ_k	−0.08	−0.09	−0.31	−0.22	0.04	−0.30
	(0.03)	(0.04)	(0.08)	(0.13)	(0.07)	(0.12)
λ_k	6.77	7.62	3.71	7.09	7.69	3.57
	(0.40)	(0.69)	(0.41)	(1.11)	(0.99)	(0.37)
ϕ_k	1.00	0.64	0.36	0.40	0.26	0.34
	(-)	(0.06)	(0.06)	(0.08)	(0.06)	(0.06)
$\ln L$	-2467.0	-2243.3		-2202.3		
$EN(\boldsymbol{\tau})$	0.00	3.73		14.36		
ICL	4957.6	4542.2		4499.2		
NEC	-	0.017		0.054		

Notes: Standard errors in parentheses. Based on our 'core sample' of 112 subjects.

Figure A.3: Distributions of individual parameter estimates (distributional, morality, and reciprocity; all subjects)



Note: All estimates of α_i, β_i, κ_i, δ_i, and γ_i larger than 2 in absolute value are grouped in bins ("<" and ">") at the extremes of the horizontal axis. Figure based on all 136 subjects.

Table A.12: Individual parameter estimates (incl. reciprocity)

Parameter	Median	Mean	S.D.	Min	Max
α_i	0.13	0.71	2.25	−0.52	21.07
β_i	0.29	0.19	7.53	−38.86	58.43
κ_i	0.06	0.23	1.12	−0.30	11.31
δ_i	−0.01	−0.40	1.38	−9.19	3.48
γ_i	−0.27	−4.62	18.95	−103.51	22.56

Notes: Table based on our 'core sample' of 112 subjects.

Table A.13: Average predictive accuracy of out-of-sample predictions

	Individual	3-types	2-types	1-type
self-interest	48.8	48.9	48.8	48.8
α, β	55.2	54.2	53.2	44.0
α, β, κ	55.5	55.8	53.9	47.9
$\alpha, \beta, \delta, \gamma$	59.1	55.3	54.0	45.1
$\alpha, \beta, \kappa, \delta, \gamma$	58.8	55.2	54.4	48.1

Notes: Table shows the average predictive accuracy (in percentages) for the out of sample predictions reported in Section 5.3 and Figure 9. Table based on our 'core sample' of 112 subjects.

Appendix A2 Distinguishing Kantian morality from social preferences

A2.1 Dictator game protocols

To see why Dictator game protocols would not allow us to distinguish between social preferences and Kantian morality à la *Homo moralis*, consider such a game in which the donor may transfer any part of his endowment w to the recipient, and the amount transferred will be multiplied by a factor $m > 1$.[33] Suppose that both players face an equal probability of being the donor, and denote by $x \in [0, w]$ and $y \in [0, w]$ their respective strategies (how much to give in the donor role). Consider first a pure altruist i, with $\beta_i = -\alpha_i \geq \kappa_i = \delta_i = \gamma_i = 0$, and thus a utility function of the form (the factor 1/2 represents nature's draw of roles):

$$u_i(x, y) \quad = \quad \frac{1}{2}\left[(1-\beta_i)(w-x+my)+\beta_i(mx+w-y)\right]. \tag{13}$$

If instead i is a pure *Homo moralis*, with $\kappa_i \geq \alpha_i = \beta_i = \delta_i = \gamma_i = 0$, then his or her expected utility is:

$$u_i(x, y) \quad = \quad \frac{1}{2}\left[(1-\kappa_i)(w-x+my)+\kappa_i(mx+w-x)\right]. \tag{14}$$

Comparison of the second terms in these utility functions reveals that while an altruist cares about the other individual's monetary payoff $(mx + w - y)/2$ (which depends on the other's strategy y), an individual driven by Kantian morality instead cares about the monetary payoff $(mx + w - x)/2$, which would result if both players were to use i's strategy x. Nonetheless, as shown by the derivatives with respect to own strategy x, the trade-off for

[33] The same argument applies if $m = 1$ as long as the subject's marginal utility from money is decreasing.

altruists and Kantian moralists is the same here:

$$\frac{du_i(x,y)}{dx} = \frac{1}{2}[\beta_i m - (1-\beta_i)], \tag{15}$$

and

$$\frac{du_i(x,y)}{dx} = \frac{1}{2}(\kappa_i m - 1). \tag{16}$$

Whether an altruist or a Kantian moralist, the individual either gives the whole endowment or nothing at all: indeed, dividing the right-hand side of (15) by $1-\beta_i$, and letting $\sigma_i \equiv \frac{\beta_i}{1-\beta_i}$, we see that the altruist gives everything if σ_i exceeds $1/m$ while the Kantian moralist gives everything if κ_i exceeds $1/m$.[34] Therefore, we would be unable to separate altruism from a Kantian concern using dictator games.[35]

A2.2 The Ultimatum, Trust, and Sequential Prisoner's Dilemma game protocols

Here we write the full expected utility expressions of a subject i with a utility function as in (1) in each of the three game protocols. The objective is to show the qualitative difference between Kantian morality on the one hand (as captured by κ_i), and social preferences on the other hand (as captured by α_i, β_i, δ_i, and γ_i).

Beginning with the Ultimatum Game protocol, as in Figure 1c, i obtains the following expected utility from using behavior strategy $x = (x_1, x_2)$ when he believes that the oppo-

[34]This observation is in line with a more general comparison of behavioral predictions for altruists and Kantian moralists in Alger and Weibull (2013), see also Alger and Weibull (2017).

[35]We would face the same identification problem with allocation tasks. Consider a subject i who faces the choice between the allocations (S,T) and (P,P), where the first entry is monetary payoff to self and the second entry is monetary payoff to the other subject, with $T > P > S$. A risk-neutral subject i with a utility function of the form in (1) strictly prefers (S,T) to (P,P) if and only if $\kappa_i(T-P) - \alpha_i(T-S) > P-S$. Hence, a subject who selects (S,T) can be driven either by pure altruism ($-\alpha_i > 0 = \kappa_i$), by pure Kantian morality ($\kappa_i > 0 = \alpha_i$), by a combination of these, or by a combination of behindness aversion and Kantian morality ($\kappa_i > \alpha_i > 0$).

nent will use behavior strategy $\hat{y} = (\hat{y}_1, \hat{y}_2)$ (the randomization factor 1/2 has been omitted):

$$\begin{aligned} u_i(x,\hat{y}) \;=\;& (1-\kappa_i)[x_1 R + (1-x_1)\hat{y}_2 T + (1-x_1)(1-\hat{y}_2)S \\ &+ \hat{y}_1 R + (1-\hat{y}_1)x_2 P + (1-\hat{y}_1)(1-x_2)S] \\ &- [(\alpha_i+\delta_i)(1-\hat{y}_1)x_2 + \beta_i(1-x_1)\hat{y}_2](T-P) \\ &+ \kappa_i[x_1 R + (1-x_1)x_2 T + (1-x_1)(1-x_2)S \\ &+ x_1 R + (1-x_1)x_2 P + (1-x_1)(1-x_2)S]. \end{aligned} \tag{17}$$

The partial derivatives with respect to x_1 and x_2 are thus:

$$\begin{aligned} \frac{\partial u_i(x,\hat{y})}{\partial x_1} \;=\;& (1-\kappa_i)[R - \hat{y}_2 T - (1-\hat{y}_2)S] + \beta_i \hat{y}_2 (T-P) \\ &+ \kappa_i[2(R-S) - x_2(T+P-2S)] \end{aligned} \tag{18}$$

$$\frac{\partial u_i(x,\hat{y})}{\partial x_2} = (1-\kappa_i)(1-\hat{y}_1)(P-S) - (\alpha_i+\delta_i)(1-\hat{y}_1)(T-P) + \kappa_i(1-x_1)(T+P-2S). \tag{19}$$

Note that in this game protocol behindness aversion matters only following the unfair offer, in which case it is augmented by the negative reciprocity parameter δ_i. Hence, in the following discussion we will refer to the term $\alpha_i + \delta_i$ simply as behindness aversion. To see the two key effects of Kantian morality mentioned in the main text, we compare an individual who is inequity averse but does not have a Kantian concern ($(\alpha_i+\delta_i)\beta_i > 0 = \kappa_i$) to one who has a Kantian concern but is not inequity averse ($\kappa_i > 0 = \alpha_i + \delta_i = \beta_i$). First, when considering the effect of his choice as a first-mover, x_1, the inequity-averse individual pays no attention to his choice as a second-mover, while the Kantian moralist does (i.e., x_2 shows up in the derivative in (18) if and only if $\kappa_i \neq 0$). Likewise, when considering the effect of his choice as a second-mover, x_2, the inequity-averse individual pays no attention to his choice as a first-mover, while the Kantian moralist does (i.e., x_1 appears in (19) if and only if $\kappa_i \neq 0$). Second, the expressions (18) and (19) show that beliefs about the opponent's play (information that we elicit from the subjects) matter less for a pure Kantian moralist than for a purely inequity averse individual. In the extreme case where $1 = \kappa_i > \alpha_i + \delta_i = \beta_i = 0$, the Kantian moralist chooses the strategy that would maximize the expected material payoff should both players choose it, irrespective of what (s)he believes the opponent will play.

In the Trust Game protocol (Figure 1b), a behavior strategy is a vector $x = (x_1, x_2) \in X = [0,1]^2$, where x_1 is the probability with which the player trusts the receiver, and x_2 the probability with which he honors trust (if the sender trusts him).[36] Then the expected utility (as defined in (1)) from playing $x = (x_1, x_2)$ against $y = (y_1, y_2)$ is (omitting the factor 1/2):

$$\begin{aligned} u_i(x,y) = {} & (1-\kappa_i)[x_1[y_2R + (1-y_2)S] + (1-x_1)P] \\ & + (1-\kappa_i)[y_1[x_2R + (1-x_2)T] + (1-y_1)P] \\ & + \kappa_i\{x_1[x_2R + (1-x_2)S] + (1-x_1)P\} \\ & + \kappa_i\{x_1[x_2R + (1-x_2)T] + (1-x_1)P\} \\ & - [\alpha_i x_1(1-y_2) + (\beta_i + \gamma_i)y_1(1-x_2)](T-S). \end{aligned} \tag{20}$$

Note that in this game protocol it is the positive reciprocity parameter γ_i which appears: it augments aheadness aversion following a 'nice' first move by the opponent. Hence, for a subject who believes that the opponent plays $\hat{y}$:

$$\frac{\partial u_i(x,\hat{y})}{\partial x_1} = (1-\kappa_i)[S-P+\hat{y}_2(R-S)] + \kappa_i[x_2(2R-S-T)+S+T-2P] - \alpha_i(1-\hat{y}_2)(T-S), \tag{21}$$

and

$$\frac{\partial u_i(x,\hat{y})}{\partial x_2} = (1-\kappa_i)\hat{y}_1(R-T) + \kappa_i x_1(2R-S-T) + (\beta_i+\gamma_i)\hat{y}_1(T-S). \tag{22}$$

Again, the individual's own play as second mover, x_2, appears in the derivative for play as first mover, x_1, if and only if $\kappa_i \neq 0$ (see in (21)). Likewise, the individual's own play as first

[36]Since each player has only one decision node, the distinction between mixed and behavioral strategies is immaterial.

mover, x_1, appears in the derivative for play as second mover, x_2, if and only if $\kappa_i \neq 0$ (see in (22)).

We turn finally to the Sequential Prisoners' Dilemma game protocol (as in Figure 1a). As noted in the main text, Defect by the first mover is classified as misbehavior if and only if $2R > S + T$, while Cooperate by the first mover is classified as nice behavior if and only if $2R > S + T$. To account for this in the expression below, let q and p be dummy variables that take the value 1 if $2R > S + T$ and 0 otherwise. The expected utility (as defined in (1)) from playing $x = (x_1, x_2, x_3)$ against $y = (y_1, y_2, y_3)$ is then (again omitting the factor 1/2):

$$\begin{aligned} u_i(x,y) &= (1-\kappa_i)[x_1 y_2 R + x_1(1-y_2)S + (1-x_1)y_3 T + (1-x_1)(1-y_3)P] \\ &+ (1-\kappa_i)[y_1 x_2 R + y_1(1-x_2)T + (1-y_1)x_3 S + (1-y_1)(1-x_3)P] \\ &+ \kappa_i[x_1 x_2 R + x_1(1-x_2)S + (1-x_1)x_3 T + (1-x_1)(1-x_3)P] \\ &+ \kappa_i[x_1 x_2 R + x_1(1-x_2)T + (1-x_1)x_3 S + (1-x_1)(1-x_3)P] \\ &- \alpha_i x_1(1-y_2)(T-S) - (\alpha_i + q\delta_i)(1-y_1)x_3(T-S) \\ &- \beta_i(1-x_1)y_3(T-S) - (\beta_i + p\gamma_i)y_1(1-x_2)(T-S). \end{aligned} \quad (23)$$

Hence, for a subject who believes that the opponent would play $\hat{y}$ one obtains:

$$\begin{aligned} \frac{\partial u_i(x,\hat{y})}{\partial x_1} &= (1-\kappa_i)[S - P + \hat{y}_2(R-S) - \hat{y}_3(T-P)] \\ &+ \kappa_i[x_2(2R-S-T) + (1-x_3)(S+T-2P)] \\ &+ \beta_i \hat{y}_3(T-S) - \alpha_i(1-\hat{y}_2)(T-S), \end{aligned} \quad (24)$$

$$\frac{\partial u_i(x,\hat{y})}{\partial x_2} = (1-\kappa_i)\hat{y}_1(R-T) + \kappa_i x_1(2R-S-T) + (\beta_i + p\gamma_i)\hat{y}_1(T-S), \quad (25)$$

and

$$\frac{\partial u_i(x,\hat{y})}{\partial x_3} = (1-\kappa_i)(1-\hat{y}_1)(S-P) + \kappa_i(1-x_1)(T+S-2P) - (\alpha_i + q\delta_i)(1-\hat{y}_1)(T-S). \quad (26)$$

Again, these equations show that an individual with a Kantian moral concern ($\kappa_i > 0$) is not only influenced by his belief about the opponent's strategy, but also by what he would himself do at every decision node of the game tree.

Appendix A3 Copula estimation

We use copula methods to describe the joint parameter distributions for the individual estimates of α_i, β_i and κ_i. For this, let X_α, X_β and X_κ be random variables, possibly statistically dependent, with marginal CDFs F_α, F_β and F_κ. By Sklar's Theorem, their joint CDF can be written in the form

$$F\left(x_\alpha, x_\beta, x_\kappa\right) = C\left(F_\alpha\left(x_\alpha\right), F_\beta\left(x_\beta\right), F_\kappa\left(x_\kappa\right)\right).$$

We follow a two-step approach (Joe & Xu, 1996; Cherubini, Luciano, & Vecchiato, 2004). First, we fit the marginal distributions. For this, we assume that each preference parameter follows a Gumbel distribution, with CDF

$$F(x) = \exp\left[-e^{-(x-a)/b}\right],$$

where $a \in \mathbb{R}$ is usually called the *location*, and $b > 0$ the *scale*. The associated PDF is

$$f(x) = \frac{1}{b}\exp\left[-(x-a)/b - e^{-(x-a)/b}\right].$$

The empirical distributions of α_i and κ_i have a relatively long right tail (see Figure 4), which fits well with the Gumbel distribution. The empirical distribution of β_i has a relatively long left tail, therefore, we fit the reverse distribution, i.e. we fit the distribution of $-\beta_i$.

In the second step, we estimate the copula. We assume a Gumbel copula, which has the form:

$$\begin{aligned} C\left(F_\alpha\left(x_\alpha\right), F_{-\beta}\left(x_{-\beta}\right), F_\kappa\left(x_\kappa\right)\right) &= \\ &\exp\left(-\left[\left(-\ln F_\alpha\left(x_\alpha\right)\right)^\omega + \left(-\ln F_{-\beta}\left(x_{-\beta}\right)\right)^\omega + \left(-\ln F_\kappa\left(x_\kappa\right)\right)^\omega\right]^{1/\omega}\right) \end{aligned}$$

for some $\omega \geq 1$, where $\omega = 1$ represents statistical independence.

In both steps we use maximum likelihood to estimate parameters. Table A.14 shows the estimated parameters, and Figure 4 plots the estimated marginal distributions together with the empirical distributions. For the joint distribution, we estimate $\omega = 1.39$. To put this into perspective, this estimate implies a Kendall's tau of $\tau = 1 - \frac{1}{1.39} = 0.28$. This compares well to the bivariate correlations (see Section 4.1). Expressed in Kendall's tau, the correlation between α_i and $-\beta_i$ is $\tau = 0.16$, for α_i and κ_i we obtain $\tau = 0.32$ and for $-\beta_i$ and κ_i we obtain $\tau = 0.15$.

Table A.14: Individual parameter estimates (all subjects)

Panel A: Marginal distributions	α_i	$-\beta_i$	κ_i
a	0.08	−0.32	0.06
b	0.14	0.33	0.11
Panel B: Joint distribution			
ω	1.39		

Notes: Table based on estimates from our core sample of 112 subjects.

Appendix A4 Robustness

A4.1 Risk aversion

Here we will take the term $\pi_i(\zeta)$ in the utility function in (1) to be the Bernoulli function value that the individual attaches to his or her monetary payoff under play ζ. If the monetary payoff allocation after a play ζ is $\left(m_i(\zeta), m_j(\zeta)\right)$, we assume that the individual's own material utility is of the CRRA form

$$\pi_i(\zeta) = \frac{m_i(\zeta)^{1-r_i} - 1}{1 - r_i}, \tag{27}$$

where r_i is the (constant) *degree of relative risk aversion* of subject i. We further assume that each subject evaluates his or her opponent's monetary payoff in terms of own risk attitude.[37] Hence, subject i evaluates the opponent j's monetary payoff as follows:

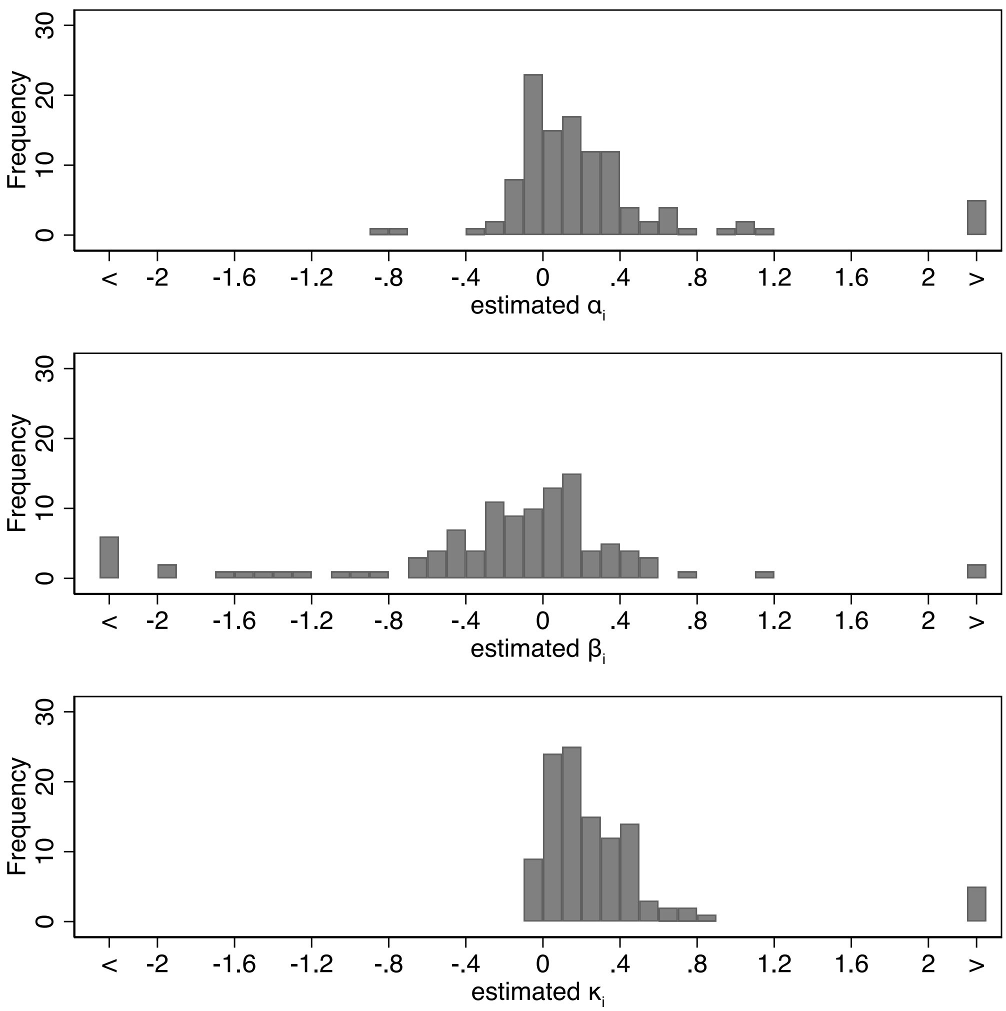
$$\pi_j^i(\zeta) = \frac{m_j(\zeta)^{1-r_i} - 1}{1 - r_i}. \tag{28}$$

Risk neutrality is the special case when $r_i = 0$, and we identify the special case $r_i = 1$ with logarithmic utility for money: then $\pi_i(\zeta) = \ln m_i(\zeta)$ and $\pi_j^i(\zeta) = \ln m_j(\zeta)$.

In a recent paper, Apesteguia and Ballester (2018) show that estimating CRRA parameters using a random utility model may be problematic. To avoid this, we estimate the social preference and Kantian morality parameters imposing risk parameters. At the individual level, we infer the risk parameter r_i from the lottery choices in the Eckel and Grossman (2002) task (see Table A.1 in Appendix A1). Figure A.4 shows the distributions of the parameter estimates when assuming these individual risk parameters. As under risk neu-

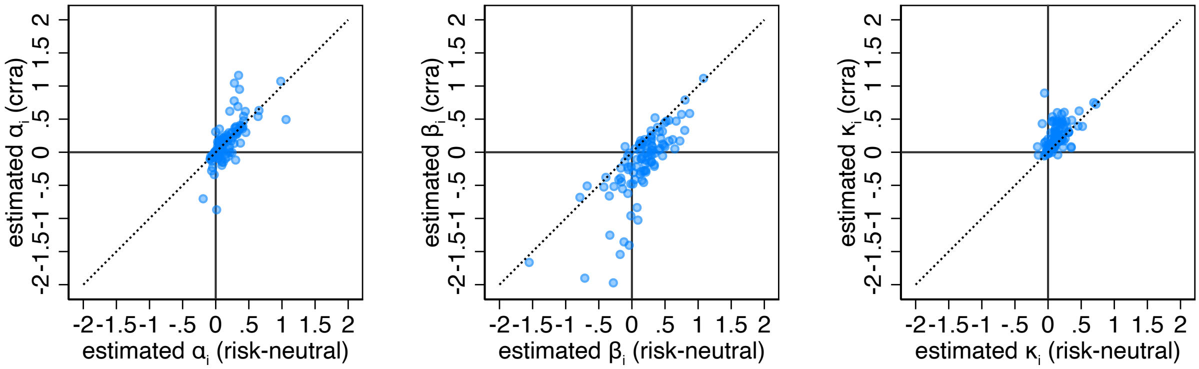
[37] There is experimental evidence that both students and financial professionals exhibit such false consensus (Roth & Voskort, 2014). Moreover, there is experimental evidence that people make the same decisions under risk (in the gain domain) for themselves and others (Andersson, Holm, Tyran, & Wengström, 2014; Exley, 2016), although Exley (2016) also shows that people sometimes act more averse to risk for others if it is in their material self-interest to do so. Gauriot, Heger, and Slonim (2020) show through simulations that estimates of social preferences parameters may depend on how subjects evaluate the monetary payoffs of others.

Figure A.4: Distributions of individual parameter estimates (allowing for risk aversion)



Note: All estimates of α_i, β_i and κ_i larger than 2 in absolute value are grouped in bins ("<" and ">") at the extremes of the horizontal axis. Figure based on all our 'core' sample of 112 subjects. For all subjects, we use the lottery choices to impose risk attitudes.

Figure A.5: Correlations between risk-neutral and CRRA estimates



Notes: Figures shows estimates smaller than 2 in absolute value. Dotted lines indicate 45 degree lines. Figure based on our 'core sample' of 112 subjects.

trality, most parameter estimates of α_i (76 out of 112) and κ_i (103 out of 112) are positive (signed-rank tests, $p < 0.001$). While we observed that most estimates of β_i are positive under risk neutrality, we now observe that most estimates of β_i (64 out of 112) are negative under CRRA preferences (signed-rank test, $p = 0.006$).

Figure A.5 shows scatter plots of individual parameter estimates under both assumptions, with estimates under risk neutrality on the horizontal axis and estimates under (individual specific) CRRA preferences on the vertical axis. Each dot represents an individual subject. The diagrams suggest that the risk-neutral and CRRA estimates are strongly correlated. Indeed, for the inequity parameter α_i (when behind) the Spearman rank correlation is $\rho = 0.706$. For the inequity parameter β_i (when ahead) it is $\rho = 0.747$, and for the Kantian morality parameter κ_i it is $\rho = 0.530$ (all three rank correlations hold for $p < 0.001$, $n = 112$).

The middle panel in Figure A.5 also shows that the β_i estimates are much higher under risk neutrality than under CRRA.[38] Indeed, for 97 out of 112 subjects, the risk-neutral

[38] As mentioned in the main text, one can easily see how assuming risk neutrality could lead to different estimates of β_k. Take for example the UG protocol. Both risk aversion and 'aheadness aversion' ($\beta_i > 0$) would induce one to choose E over U. To further see why risk aversion leads to lower degrees of aheadness aversion—sometimes even aheadness loving—and higher degrees of Kantian morality than under risk neutrality, let (for the sake of this argument) $u(m)$ denote the material utility associated with the monetary payoff m. If sufficiently strong, both aheadness aversion and Kantian morality can make an individual refrain from defecting in the Sequential Prisoner's Dilemma, keeping the money in the Trust game, and proposing the

estimate is higher than the CRRA estimate (signed-rank test, $p < 0.001$). By contrast, the risk-neutral estimates of κ_i (87 out of 112, signed-rank test: $p < 0.001$) are lower for most subjects than under CRRA. For α_i there is no clear directional shift as the risk-neutral estimates are higher than the CRRA estimates for 52 out of 112 subjects (signed-rank test: $p = 0.934$). For the majority of subjects (68 out of 112), assuming CRRA preferences instead of risk neutrality leads to a higher log-likelihood, indeed indicating a better fit under CRRA preferences.

Table 5 in the main text presents the estimates of the finite mixture estimation allowing for distributional preferences and Kantian morality under CRRA preferences (we impose logarithmic utility ($r_k = 1$) for all types). For a discussion of these estimates, see section 4.3.1 in the main text. Table A.15 shows the estimates of finite mixture models allowing for reciprocity on top of distributional preferences and Kantian morality, again assuming logarithmic utility. Compared to the models without reciprocity (see Table 5), the estimates of α_k, β_k, and κ_k for the 1-type and 2-types models are not much affected. For the 3-types model, Type 3 again combines spiteful or competitive preferences ($\alpha_k > 0$, $\beta_k < 0$) with Kantian morality ($\kappa_k = 0.17$). Types 1 and 2 differ somewhat from Types 1 and 2 without reciprocity. In contrast to the risk-neutral model (see Table 6), we now observe reciprocity parameters δ_k and γ_k with the expected positive sign. Type 1 in each of the models displays negative reciprocity, while Type 2 in the 3-types model displays positive reciprocity. This suggests that the negative estimates of the reciprocity parameters under risk neutrality might be partly explained by risk aversion. For Type 2 in the two-types model and Type 3 in the three-types model, we again estimate a negative value for γ_k, indicating very strong competitive or spiteful preferences.

Figure A.6 shows the ICL scores of different models allowing for distributional preferences, Kantian morality, reciprocity, and cominations thereof, assuming logarithmic utility

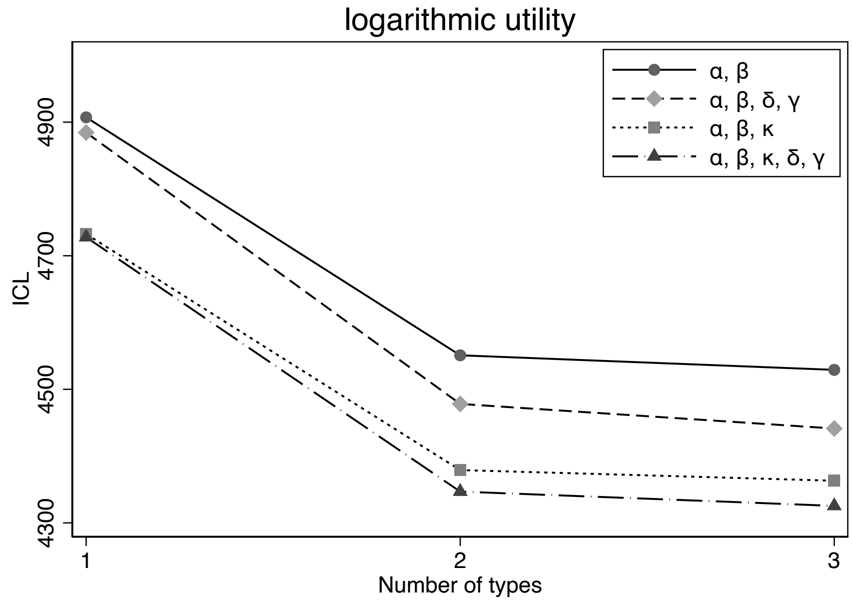
unequal split in the Ultimatum game. However, the effect appears for different reasons: while aheadness aversion entails disutility proportional to the difference $u(T) - u(S)$, Kantian morality generates utility proportional to $u(T) + u(S)$. Since $u(T) - u(S)$ relative to $u(T) + u(S)$ is smaller for a strictly concave function u than for a linear one, the above mentioned pro-social actions will appear to be driven more by Kantian morality than by aheadness aversion for a strictly concave function u than for a linear one.

Table A.15: Estimates at the aggregate level (distributional, Kantian morality, and reciprocity; under CRRA)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.10	−0.02	0.29	−0.03	−0.09	0.30
	(0.03)	(0.03)	(0.07)	(0.07)	(0.08)	(0.09)
β_k	−0.03	0.10	−0.23	0.07	0.16	−0.23
	(0.04)	(0.05)	(0.07)	(0.12)	(0.09)	(0.08)
κ_k	0.19	0.19	0.17	0.15	0.29	0.17
	(0.01)	(0.02)	(0.02)	(0.04)	(0.05)	(0.03)
δ_k	0.10	0.21	−0.08	0.27	0.06	−0.09
	(0.04)	(0.05)	(0.07)	(0.09)	(0.09)	(0.08)
γ_k	0.05	0.01	−0.31	−0.04	0.23	−0.32
	(0.04)	(0.05)	(0.14)	(0.21)	(0.12)	(0.13)
λ_k	0.24	0.26	0.17	0.25	0.24	0.17
	(0.01)	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)
ϕ_k	1.00	0.60	0.40	0.47	0.14	0.39
	(-)	(0.06)	(0.06)	(0.09)	(0.08)	(0.06)
$\ln L$	-2349.7	-2141.2		-2110.6		
$EN(\boldsymbol{\tau})$	0.00	3.19		10.02		
ICL	4727.8	4346.9		4325.5		
NEC	-	0.015		0.042		

Notes: Standard errors in parentheses. For all types, we impose logarithmic utility ($r_k = 1$). Table based on our 'core sample' of 112 subjects.

Figure A.6: ICL scores (under CRRA)



Notes: ICL scores of different finite mixture models. Lower ICL scores indicate a more preferred model. For all models, we impose logarithmic utility ($r_k = 1$). Figure based on our 'core sample' of 112 subjects.

($r_k = 1$). As under risk-neutrality (see Figure 7), the multi-types models again strongly outperform the representative agent models. Moreover, in line with the observations in section 4.3, allowing for CRRA preferences strongly improves the fit of the models, indicated by the lower ICL scores in Figure A.6 than in Figure 7. Also under logarithmic utility, we find that the models including the Kantian morality parameter κ, clearly outperform the models without κ.

We also evaluate the value-added of distributional preferences, Kantian morality, and reciprocity under risk aversion at the individual level. In Table A.16 we show which models are selected based on BIC and AIC criteria. As under risk neutrality (see Table 7), for 21 subjects (18.8%) self-interest is selected based on the BIC. Compared to the models under risk neutrality, we observe that when allowing for risk aversion models that include β or γ are less often selected, and that the selected models more often include the Kantian morality parameter κ. As under risk neutrality, β is most often included in the best-fitting models at the individual level (60 subjects, or 56.6% for BIC). Based on the lowest BIC score, including κ improves the fit for 31 subjects (27.7%), a number close to those for α (28 subjects, or 25.0%) and γ (32 subjects, or 28.6%). The negative reciprocity parameter δ is included for 16 subjects (14.3%).

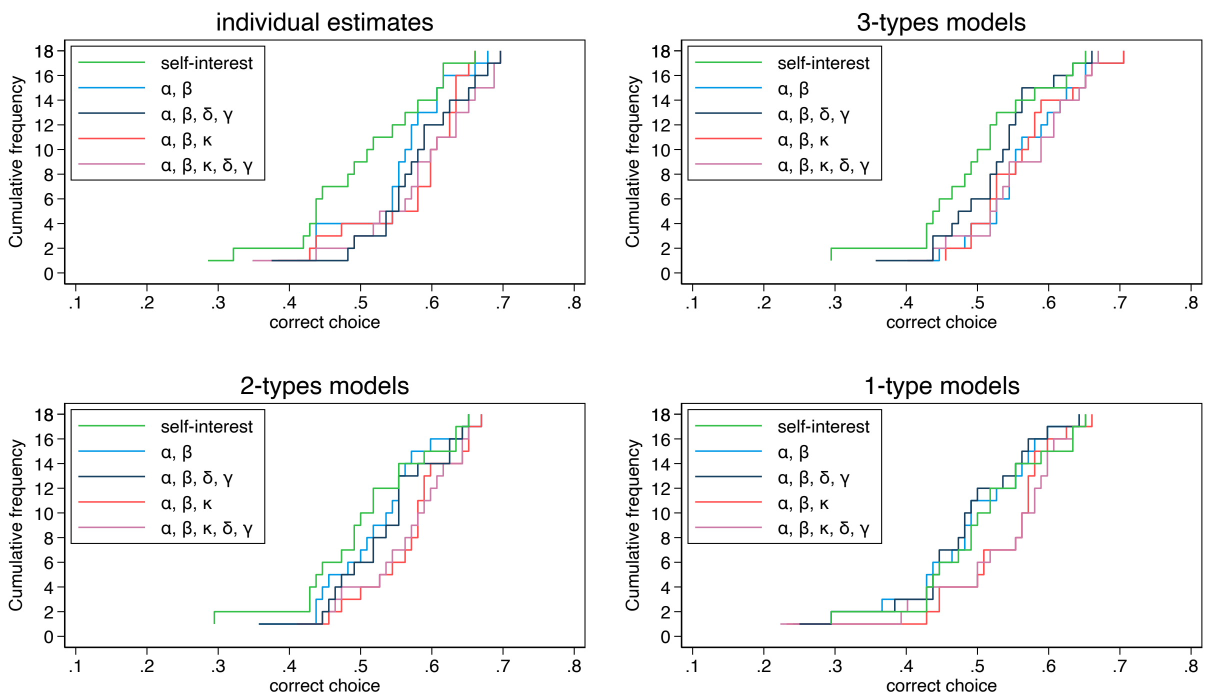
Table A.16: Best individual fit (under CRRA)

	BIC		AIC	
Parameters	Frequency	Percentage	Frequency	Percentage
$\alpha, \beta, \kappa, \delta, \gamma$			3	2.7
$\alpha, \beta, \kappa, \delta$				
$\alpha, \beta, \kappa, \gamma$			5	4.5
$\alpha, \beta, \delta, \gamma$	2	1.8	4	3.6
α, β, κ	1	0.9	1	0.9
α, β, δ	5	4.5	4	3.6
α, β, γ	4	3.6	4	3.6
α, κ, δ	2	1.8	2	1.8
β, κ, γ	7	6.2	9	8.0
α, β	4	3.6	3	2.7
α, κ			2	1.8
α, δ	7	6.2	7	6.2
β, κ	2	1.8	4	3.6
β, γ	19	17.0	18	16.1
α	3	2.7	4	3.6
β	16	14.3	12	10.7
κ	19	17.0	15	13.4
-	21	18.8	15	13.4
Selected model includes:				
Parameter	Frequency	Percentage	Frequency	Percentage
α_i	28	25.0	39	34.8
β_i	60	53.6	67	59.8
κ_i	31	27.7	41	36.6
δ_i	16	14.3	20	17.9
γ_i	32	28.6	43	38.4

Notes: Entries in the top panel indicate the number of subjects for whom the specific model provides the lowest BIC or AIC score respectively. Entries in the bottom panel summarize how frequently a parameter was included in the the model the lowest BIC or AIC score respectively. Table based on our 'core sample' of 112 subjects. For all subjects, we use the lottery choices to impose risk attitudes.

Figure A.7 shows the results of out-of-sample predictions under CRRA preferences. Table A.17 lists the average predictive accuracy for each model. The qualitative patterns in the data resemble those under risk neutrality (see Section 5.3 in the main text). At the individual level, all models allowing for distributional preferences perform much better

Figure A.7: Accuracy of out-of-sample predictions (under CRRA)



Notes: Accuracy of out-of-sample predictions, based on individual estimates (top left panel) and finite mixture models with three-types, two-types, or a representative agent (1 type). For the individual estimates, we impose risk parameters r_i based on the choices in the lottery task, in the finite mixture models we impose logarithmic utility ($r_k = 1$). Plots show cumulative frequency plots for the average fraction of correctly predicted choices per game protocol. Figure based on our 'core sample' of 112 subjects.

Table A.17: Average predictive accuracy of out-of-sample predictions (under CRRA)

	Individual	3-types	2-types	1-type
self-interest	49.8	49.3	49.7	49.7
α, β	55.1	56.1	52.1	48.0
α, β, κ	57.2	56.0	56.5	52.8
$\alpha, \beta, \delta, \gamma$	57.5	52.4	53.3	47.7
$\alpha, \beta, \kappa, \delta, \gamma$	58.2	56.2	56.2	52.7

Notes: Table shows the average predictive accuracy (in percentages) for the out of sample predictions reported in Figure A.7. Table based on our 'core sample' of 112 subjects.

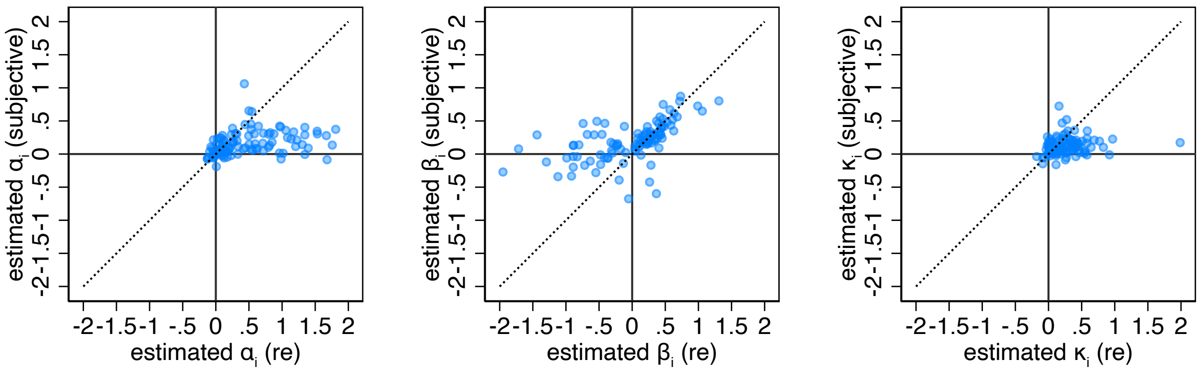
than self-interest. The average predictive accuracy for models allowing for distributional preferences ranges between 55.1% and 58.2%, compared to 49.8% under self-interest. Allowing for only one type (a representative agent) again results in relatively low predictive accuracy (between 48.0% and 52.8%), while models allowing for one or two types perform much better, especially when distributional preferences are combined with Kantian morality and/or reciprocity.

A4.2 Rational expectations

In the main analyses, we assume that subjects maximize their expected utility given their stated beliefs. As a robustness check, we also estimate preference parameters under the alternative assumption of rational expectations. Figure A.8 shows correlations between the individual estimates using subjective and rational expectations. For all three preference parameters, the estimates under the two assumptions are positively correlated. For the inequity parameter α_i (when behind) the Spearman rank correlation is $\rho = 0.402$ ($p < 0.001$, $n = 112$). For the inequity parameter β_i (when ahead) it is $\rho = 0.734$ ($p < 0.001$, $n = 112$), and for the Kantian morality parameter κ_i it is $\rho = 0.220$ ($p = 0.020$, $n = 112$).

Table A.18 shows the finite mixture estimates when we assume rational expectations. The representative agent with rational expectations is characterized by a combination of behindness aversion ($\alpha_k > 0, \beta_k = 0$) and morality ($\kappa_k > 0$). Compared to the model with subjective expectations (see Table 4 in the main text), the estimates for α_k and κ_k are larger

Figure A.8: Correlations between estimates using subjective and rational expectations



Notes: Figures shows estimates smaller than 2 in absolute value. Dotted lines indicate 45 degree lines. Figure based on our 'core sample' of 112 subjects.

Table A.18: Estimates at the aggregate level (assuming rational expectations)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.41	0.12	1.02	0.16	0.08	1.05
	(0.05)	(0.02)	(0.09)	(0.07)	(0.05)	(0.09)
β_k	0.00	0.35	−0.65	0.28	0.41	−0.69
	(0.07)	(0.03)	(0.12)	(0.05)	(0.10)	(0.16)
κ_k	0.27	0.14	0.49	0.09	0.24	0.51
	(0.03)	(0.03)	(0.05)	(0.03)	(0.06)	(0.06)
λ_k	10.67	7.84	10.35	8.14	6.07	10.11
	(0.62)	(0.57)	(1.00)	(0.77)	(0.90)	(0.88)
ϕ_k	1.00	0.56	0.44	0.37	0.22	0.42
	(-)	(0.05)	(0.05)	(0.06)	(0.06)	(0.06)
$\ln L$	-2689.8	-2316.0		-2261.9		
$EN(\boldsymbol{\tau})$	0.00	2.09		9.97		
ICL	5398.6	4676.5		4599.8		
NEC	-	0.006		0.023		

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 112 subjects. For all types, we assume risk neutrality ($r_k = 0$) and rational expectations.

when we assume rational expectations. The estimate for β_k is zero when we assume rational expectations, where it was positive under subjective expectations. For the representative agent model, the log-likelihood is lower when assuming rational expectations. For the two-types model and three-types model, assuming rational expectations leads to qualitatively similar results as under subjective expectations for some types, but quite different estimates for other types. For the two-types model, Type 1 again displays a combination of inequity aversion and morality, with estimates very close in magnitude to those under subjective expectations. Assuming rational expectations, Type 2 now combines strong spite ($\alpha_k > 0, \beta_k < 0$) with strong morality ($\kappa_k > 0$). This contrasts with the estimates under subjective expectations where Type 2 combined behindness aversion with milder morality.[39] For the three-types model, Types 1 and 2 are again very similar under both subjective and rational expectations, while Type 3's estimates are quite different. Given a number of types, the ICL scores under rational expectations are higher than under subjective expectations, indicating a worse fit under rational expectations. In sum, most estimated preference parameters for the multi-type models are very similar under both assumptions, while for some types the estimates differ. In combination with the higher ICL scores under rational expectations, this suggests that assuming subjective expectations is an important assumption for part of the population.

A4.3 Game protocol type specific noise parameters

In Table A.19 we present estimates of finite mixture models where we allow for different noise parameters λ for each game protocol type (SPD, TG, UG). Comparing the estimates in Table A.19 to those with a single λ for each type in Table 4, we find that the estimates of the preference parameters are nearly identical for the 1-type and 2-types models. For the 3-types models, the point estimates differ somewhat, but the three types are qualitatively similar. In both cases, Type 1 displays a combination of inequity aversion and Kantian

[39] Table A.6 (panels D and E) shows that the assignment of subjects to types is similar under subjective and rational expectations.

Table A.19: Estimates at the aggregate level (game protocol type specific noise parameters)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.16	0.11	0.15	0.19	0.10	0.08
	(0.01)	(0.02)	(0.03)	(0.03)	(0.02)	(0.04)
β_k	0.26	0.40	−0.02	0.08	0.42	−0.02
	(0.03)	(0.03)	(0.04)	(0.04)	(0.03)	(0.09)
κ_k	0.09	0.08	0.09	0.10	0.08	0.07
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.03)
$\lambda_{SPD,k}$	6.46	8.23	2.92	4.78	8.63	0.31
	(0.45)	(0.84)	(0.38)	(0.37)	(0.43)	(0.10)
$\lambda_{TG,k}$	10.25	12.94	3.56	3.80	12.05	4.35
	(1.39)	(2.41)	(0.70)	(0.74)	(0.60)	(0.96)
$\lambda_{UG,k}$	5.83	4.55	6.19	5.86	4.29	6.73
	(0.49)	(0.56)	(0.67)	(0.62)	(0.31)	(0.54)
ϕ_k	1.00	0.59	0.41	0.28	0.52	0.20
	(-)	(0.05)	(0.05)	(0.05)	(0.05)	(0.04)
$\ln L$	-2418.8	-2203.9		-2184.2		
$EN(\boldsymbol{\tau})$	0.00	4.13		14.56		
ICL	4866.0	4473.4		4477.3		
NEC	-	0.019		0.062		

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 112 subjects.

morality. Type 2 combines aheadness aversion with Kantian morality in both cases, although this type is also motivated by behindness aversion in Table A.19, where in Table 4 the α estimate was very close to zero. Type 3 combines aheadness aversion with a Kantian moral concern in both cases.

Appendix A5 Simulations

In the process of selecting game protocols for the experiment, we conducted some simulations to check whether we can retrieve the original parameters based on the set of experimental game protocols. In this appendix, we describe how such simulations were conducted.

Generating simulated data

First, for each (simulated) subject i, we randomly draw preference parameters $\alpha_i, \beta_i, \kappa_i$ independently from uniform distributions. For α_i and β_i, we draw the preference parameters from $U[-0.5, 0.5]$, while we draw κ_i from $U[0, 0.5]$. In the simulations, we set $\delta_i = 0$. Second, for each subject i and game protocol g we compute the expected utility for each possible pure strategy x_i based on utility function (1), with subjective beliefs $\hat{y}_i$ drawn independently from $U[0, 1]$. We then compute choice probabilities for each pure strategy x_i based on equation (4), where for all subjects we impose some fixed noise parameter λ_i. Based on these choice probabilities, we randomly select a behavioral strategy for each subject i and each game protocol g.

Estimation

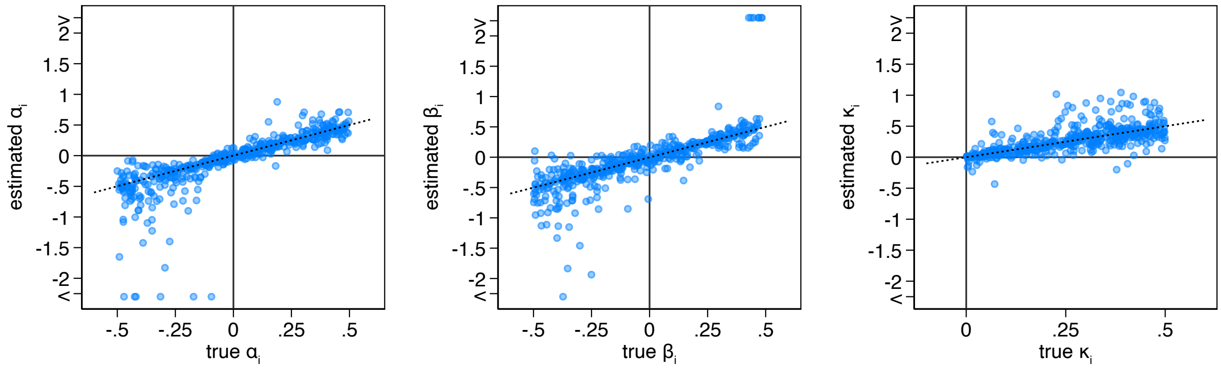
For the simulated data, we then estimate individual preference parameters as described in section 3.

Results

Figure A.9 shows the correlations between the simulated ('true') parameters and the estimated parameters, for 500 individuals with relatively low noise levels ($\lambda_i = 0.5$). Most estimated preference parameters lie very close to the 45-degree line, indicating that we can well retrieve the preference parameters.

Figures A.10 and A.11 show simulations with higher noise levels. When increasing the

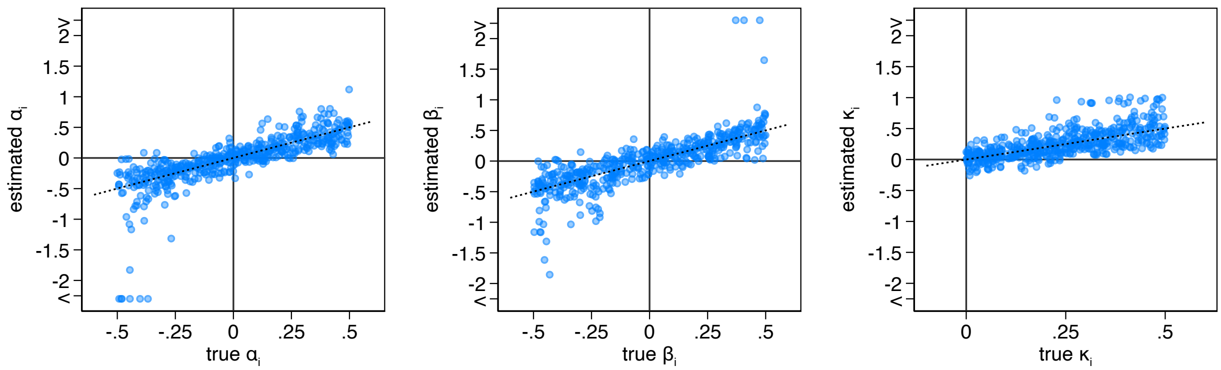
Figure A.9: Simulations ($\lambda_i = 0.5$)



Notes: Scatter plots shows the correlations between the simulated ('true') parameters and the estimated parameters. All estimated parameters larger then 2 in absolute value are grouped in the bins at the extremes of the vertical axes. Dotted lines indicate 45 degree lines. Figure based on 500 simulated subjects.

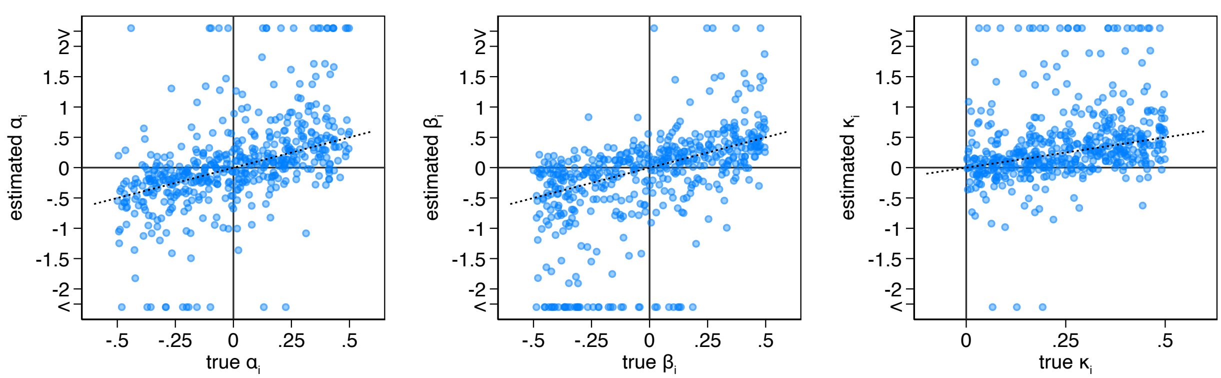
noise parameter to $\lambda_i = 5$, most estimated preference parameters are still close to the 45 degree line (see Figure A.10). Note that this noise level is roughly in the same ballpark as what we estimate in our pre-registered analyses (for the core sample, the median estimated $\lambda_i = 4.5$). When we further increase the noise parameter to $\lambda_i = 20$, estimated preference parameters lie much further from the 45-degree line, but we still observe strong correlations between the true and estimated parameters.

Figure A.10: Simulations ($\lambda_i = 5$)



Notes: Scatter plots shows the correlations between the simulated ('true') parameters and the estimated parameters. All estimated parameters larger then 2 in absolute value are grouped in the bins at the extremes of the vertical axes. Dotted lines indicate 45 degree lines. Figure based on 500 simulated subjects.

Figure A.11: Simulations ($\lambda_i = 20$)



Notes: Scatter plots shows the correlations between the simulated ('true') parameters and the estimated parameters. All estimated parameters larger then 2 in absolute value are grouped in the bins at the extremes of the vertical axes. Dotted lines indicate 45 degree lines. Figure based on 500 simulated subjects.

Appendix A6 Pre-registration

We pre-registered our main design elements (sample size, type of game protocols), and main analyses on aspredicted.org (see https://aspredicted.org/blind.php?x=4u5nu8). Below we reproduce the pre-registration.

1) Have any data been collected for this study already?

No, no data have been collected for this study yet

2) What's the main question being asked or hypothesis being tested in this study?

We test whether people's preferences in social dilemma situations (SPDs, TGs, UGs) can be well described by 'homo moralis' preferences as in Alger and Weibull (2013).

3) Describe the key dependent variable(s) specifying how they will be measured.

We measure actions and beliefs in SPDs TGs and UGs

4) How many and which conditions will participants be assigned to?

There is one condition (each participant plays the same games, in different, random order, for each session).

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

We will use maximum likelihood to estimate (individual) parameters of a utility function that includes three parameters, alpha, beta and gamma. Alpha en beta capture inequity aversion and gamma captures moral preferences. We use a logit specification. Using this model, we compare the predictive value (within the sample) of the general model to restricted versions of the model. The general model nests inequity aversion, altruism, homo moralis and selfish preferences.

6) Any secondary analyses?

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined. We will run 8 sessions, with 22 people invited for each session. If we have fewer than 120 subjects after the 8 sessions, we will run more sessions until we pass the 120 subjects minimum.

8) Anything else you would like to pre-register?
(e.g., data exclusions, variables collected for exploratory purposes, unusual analyses planned?)

Appendix A7 Experimental instructions

Welcome

Welcome to this experiment. All subjects receive the same instructions. Please read them carefully.

Do not communicate with any of the other subjects during the entire experiment. If you have any questions, raise your hand and wait until one of us comes to you to answer your question in private.

During the experiment you will receive points. These points are worth money. How many points (and hence how much money) you get depends on your own decisions, the decisions of others, and chance. At the end of the experiment the points that you got will be converted to euros and the amount will be paid to you privately, in cash.

Every point is equivalent to 0.17 euro.

Your decisions are anonymous. They will not be linked to your name in any way. Other subjects can never trace your decisions back to you.

Today's experiment consists of two parts. At the beginning of each part, you will receive new instructions. Your decisions made in one part will never affect outcomes in another part, so you can treat both parts as independent.

Decision situations I

In this part, you will participate in 18 different decision situations. For each decision situation, you will be randomly paired with someone else in the lab. Therefore, in each decision situation you will (most likely) be paired with a different subject than in the previous situation. You will never learn with whom you are paired.

The 18 decision situations will all be different, but they all involve two persons, and in all the decision situations one person is assigned to Role A (person A) while the other is assigned to Role B (person B). There are then two kinds of situations, as depicted in Figures 1 (below) and Figure 2 (on the next page).

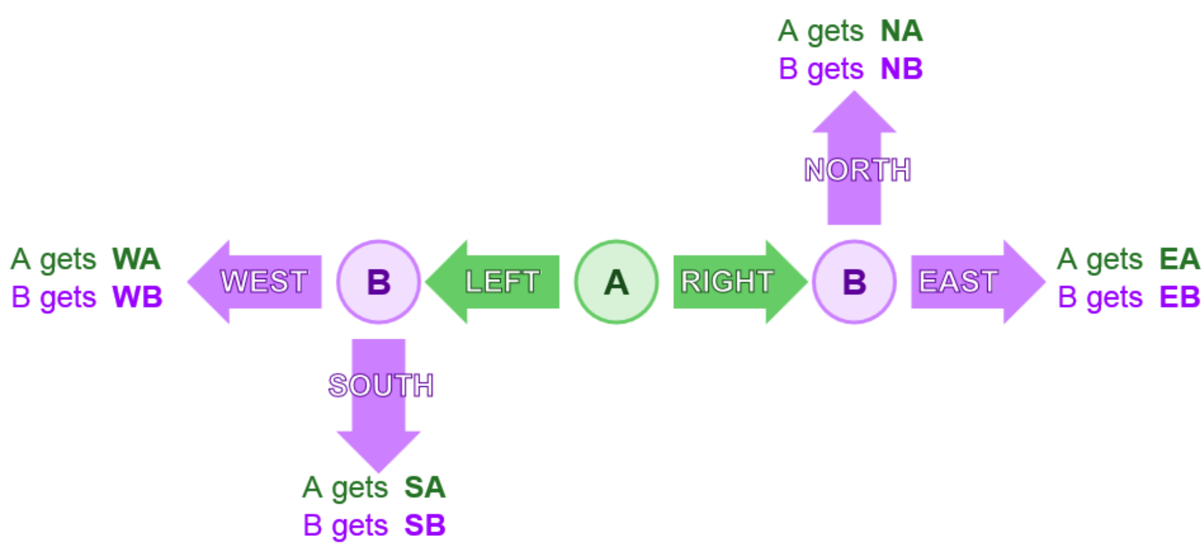
In the situation shown in Figure 1, person A first chooses LEFT or RIGHT. If A chooses LEFT, person B has to choose between WEST or SOUTH. If person A chooses RIGHT, person B has to choose between NORTH and EAST.

The choices of A and B jointly determine the number of points for A and B as follows:

- If A chooses LEFT and B chooses WEST, A gets WA points and B gets WB points
- If A chooses LEFT and B chooses SOUTH, A gets SA points and B gets SB points
- If A chooses RIGHT and B chooses NORTH, A gets NA points and B gets NB points
- If A chooses RIGHT and B chooses EAST, A gets EA points and B gets EB points

The values of WA, WB, SA, SB, NA, NB, EA and EB vary from one decision situation to another. At the beginning of each decision situation, you and all others in the lab will be informed of the values.

Figure 1



Decision situations II

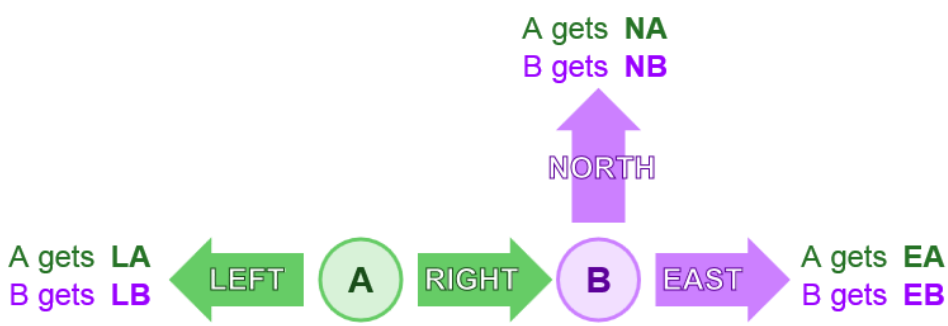
In the decision situation shown in Figure 2, person A first chooses LEFT or RIGHT. If A chooses LEFT, person B has no choice to make. If A chooses RIGHT, B has to choose between NORTH and EAST.

The choices of A and B jointly determine the number of points for A and B as follows:

- If A chooses LEFT, A gets LA points and B gets LB points
- If A chooses RIGHT and B chooses NORTH, A gets NA points and B gets NB points
- If A chooses RIGHT and B chooses EAST, A gets EA points and B gets EB points

The values of LA, LB, NA, NB, EA and EB vary from one decision situation to another. At the beginning of each decision situation, you and all others in the lab will be informed of the values.

Figure 2



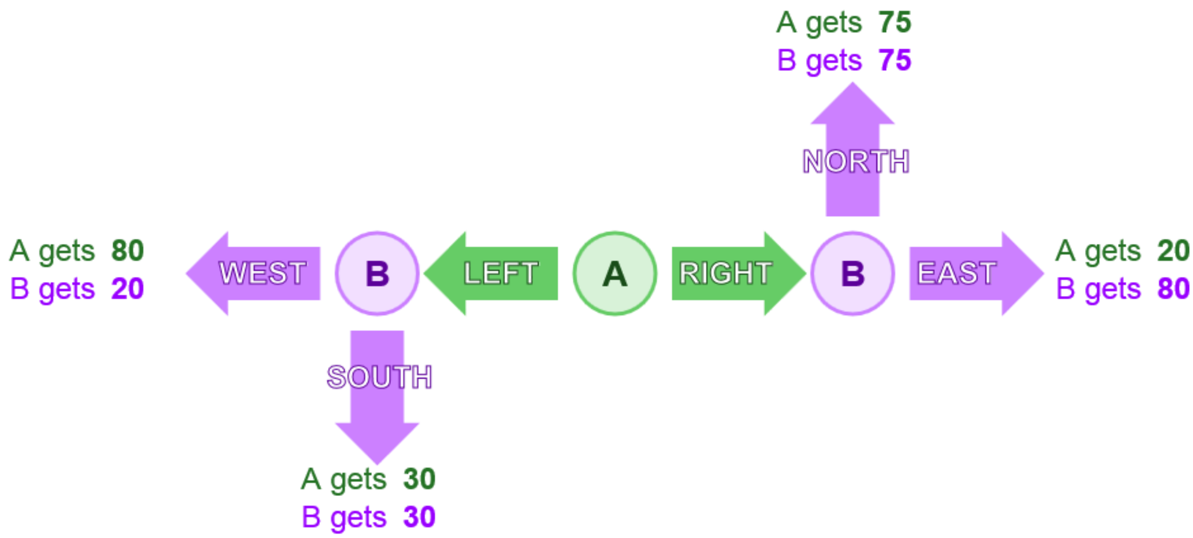
Example

The figure below gives an example of a decision situation. This decision situation is randomly selected. Remember that each of the 18 decision situations will be different.

In this example:

- If A chooses LEFT and B chooses WEST, A gets 80 points and B gets 20 points
- If A chooses LEFT and B chooses SOUTH, A gets 30 points and B gets 30 points
- If A chooses RIGHT and B chooses NORTH, A gets 75 points and B gets 75 points
- If A chooses RIGHT and B chooses EAST, A gets 20 points and B gets 80 points

If you want to see another example, click here



Decisions and payments

You will see 18 different decision situations. For each decision situation, you will be asked two things.

First, we will ask you what you want to do in Role A and what you want to do in Role B.

Second, we will ask you to guess what the others in the lab will do in Role A and what they will do in Role B. Specifically, we will ask you to guess:

- What percentage of the other people in the lab choose LEFT and what percentage choose RIGHT when in Role A
- What percentage of the other people in the lab choose WEST and what percentage choose SOUTH when facing that choice in Role B
- What percentage of the other people in the lab choose NORTH and what percentage choose EAST when facing that choice in Role B.

Both your decisions and your guesses will determine how many euros you get at the end of the experiment. Specifically, at the end of today's experiment, **two of the 18 decision situations will be randomly selected for payment: for one of these situations you**

get points from the decisions, while for the other situation you get points from your guesses. The same two decision situations will be selected for everyone in the lab.

Your decisions

For one decision situation you and the others in the lab get points from the decisions. For this situation, either you or the person you are paired with is assigned to Role A, while the other is assigned to Role B, with equal probability for each case. The number of points you and this other person get is then determined by your decision in the role to which you were assigned and the decision of the other person in the role to which (s)he was assigned.

Note that it is equally likely that your choices in role A or role B count. Think about flipping a coin: if heads comes up you will be in role A and if tails comes up you will be in role B. When you make your decisions, you do not know which role you have and you should therefore make decisions as if each role could determine the outcome, which is the case.

Your guesses

For another decision situation you and the others in the lab get points from the guesses. You get more points the closer your guesses are to what the others actually choose in both roles A and B. One of the guesses that you make in this situation will be randomly selected for payment. Specifically, you get between 0 and 50 points depending on the accuracy of your guess. If you want to earn as much as possible with your guesses, you should simply answer with what you really think is the most likely answer to each question. Your guesses do not have any impact on the number of points that the others in the lab get.

If you want to see how your earnings are calculated you can click here.

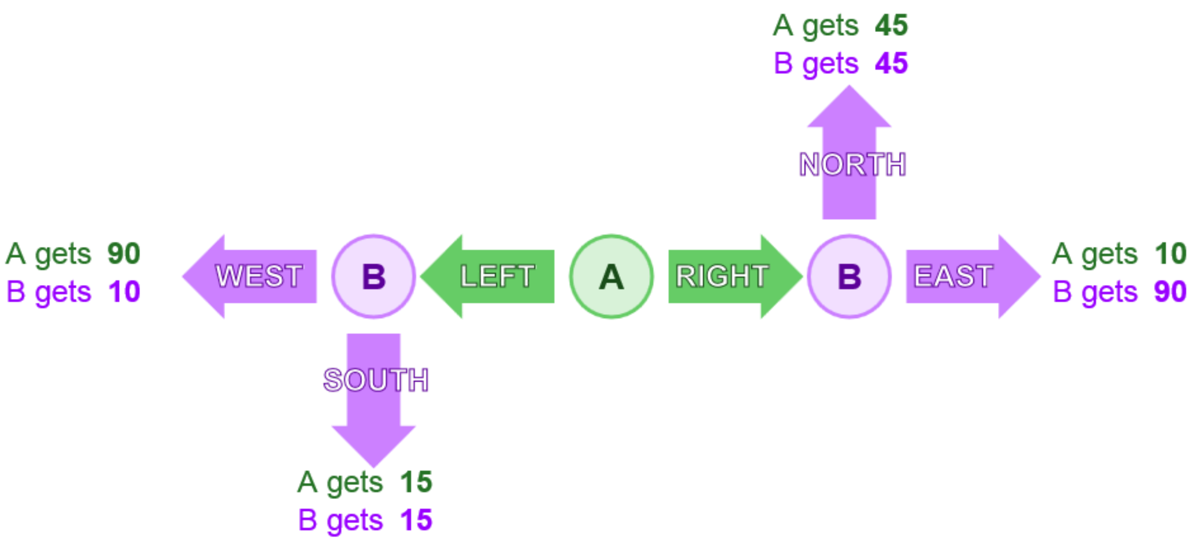
Decision screens

Below you can see and try the decision screens. First, you will see the screen where you will be asked for a decision in a decision situation. If you make a decision, you will be taken to the screen where you will be asked for a guess about what others will do.

In the examples below, all decision situations are chosen randomly. You can try the decision screens as often as you want.

Show example

Quiz questions I



Please answer the following quiz questions. If you have any questions please raise your hand.

The 18 decision situations:

O are always the same

O are sometimes the same

O are always different

The figure shows a possible decision situation. The figure merely serves as an example, the decision situation has been selected randomly.

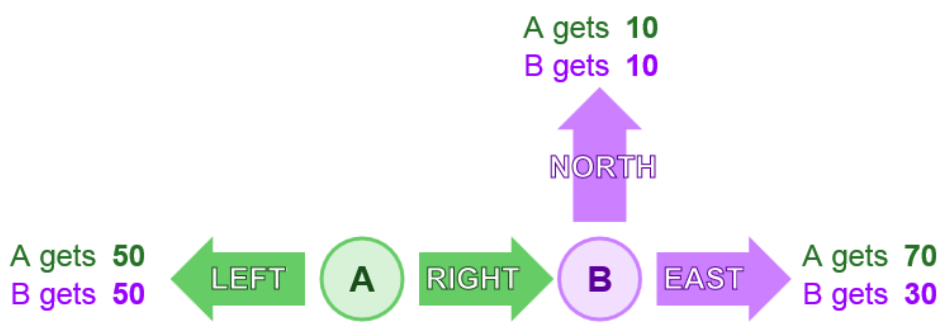
Suppose A chooses LEFT and B chooses SOUTH and EAST. How much would A and B earn?

A would earn: ___ points B would earn: ___ points

Suppose A chooses RIGHT and B chooses WEST and NORTH. How much would A and B earn?

A would earn: ___ points B would earn: ___ points

Quiz questions II



Please answer the following quiz questions. If you have any questions please raise your hand.

In each decision situation:

O you will have the same role (A or B)

O it is equally likely that you will be in role A or B

In each decision situation:

O you will be paired with the same subject

O you will be paired with a randomly determined subject

The figure shows a possible decision situation. The figure merely serves as an example, the decision situation has been selected randomly.

Suppose A chooses LEFT and B chooses NORTH. How much would A earn?

A would earn: ___ points B would earn: ___ points

Suppose A chooses RIGHT and B chooses EAST. How much would B earn?

A would earn: ___ points B would earn: ___ points

End of instructions

You have reached the end of the instructions. You can still go back by using the menu above. If you are ready, click on 'continue' below. If you need help, please raise your hand.

As soon as everyone has finished with instructions the experiment will start. During the experiment, you can take as much time as you need for each decision situation.

Part II

In this part you choose one of the six options listed below. You choose by clicking on the option you prefer. Each option has two possible outcomes (Outcome A or Outcome B) that are equally likely to occur. Think about the flip of a coin: heads (Outcome A) and tails (Outcome B) are equally likely.

At the end of the experiment, the computer will randomly select Outcome A or Outcome B. You will receive the number of points corresponding to the option you chose. For example: If you choose option 4 you will receive 30 points if Outcome A is selected by the computer and 9 points if Outcome B is selected by the computer.