Estimating Social Preferences and Kantian Morality in Strategic Interactions

Ingela Alger, Boris van Leeuwen

▶ **To cite this version:**

Ingela Alger, Boris van Leeuwen. Estimating Social Preferences and Kantian Morality in Strategic Interactions. 2021. hal-03142431

HAL Id: hal-03142431
https://hal.science/hal-03142431

Preprint submitted on 16 Feb 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

ESTIMATING SOCIAL PREFERENCES AND KANTIAN MORALITY IN STRATEGIC INTERACTIONS*

Boris van Leeuwen† Ingela Alger‡

14th January 2021

Abstract: Theoretical work 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. A finite mixture analysis shows that the subject pool is well described as consisting of two types. One combines inequity aversion and Kantian morality, while the other combines spite and Kantian morality. The value of adding Kantian morality to well-established preference classes is also evaluated.

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

Keywords: Social preferences, Kantian morality, other-regarding preferences, 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, University of Copenhagen, Stockholm School of Economics and the conference on Markets, Morality, and Social Responsibility (Toulouse) 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 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. 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 consequentialistic concerns such as altruism and inequity aversion. We do this by way of conducting an experimental study.

The laboratory experiment consists of 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 consequentialistic 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 he played against himself. By contrast, a subject with purely consequentialistic 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]

[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

Positing a utility function with three parameters capturing attitudes towards unfavorable inequity, favorable inequity, and the Kantian moral concern, 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]

Not surprisingly, the estimations at the level of the individual subjects reveal a lot of heterogeneity in preferences. While many subjects appear to be averse to unfavorable inequity, some appear to be either indifferent or either like or dislike favorable inequity. The behavior of most subjects is compatible with some concern for Kantian morality, and allowing for this motivational factor significantly improves the fit of the model to the data. Kantian morality further appears in all the aggregate estimations. The representative agent in the subject pool combines "behindness aversion" with Kantian morality.[4] 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 "spite" or "competitiveness" – an aversion to being behind and taste for being ahead – with Kantian morality.

Importantly, allowing for Kantian morality substantially improves the fit of the model. Model selection criteria and out-of-sample predictions indeed favor models with Kantian morality over those without. Comparing our main estimates to those based on a utility function with negative reciprocity as in Charness and Rabin (2002) instead of the Kantian moral concern further shows that the value added of Kantian morality is in the same ballpark as such well-established motives as inequity aversion, altruism, and reciprocity. Moreover, the out-of-sample predictions are more accurate with preferences that combine

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.

[4]Interestingly, this is in line with the theoretical prediction of Alger et al. (2020), who show in a general model that preferences that combine material self-interest, a Kantian moral concern and other-regard at the material payoff level is what should be expected in most human populations.

Kantian morality with attitudes towards the realized payoff distribution, than any of the preferences without the Kantian moral concern.

Our paper fits in the large literature that estimates or tests models of social preferences.[5] In relation to this literature, our main contribution is that we allow for the possibility 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.[6] 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 rich 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 allows 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 and Kantian morality. As indicated by our results, most subjects indeed appear to have such complex preferences. Furthermore, our data enables use of 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 are presented in Section 4, wherein we also report robustness checks and several measures of the value added of Kantian morality in our experiment. Section 5 concludes.

[5] 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 a recent survey, see Cooper and Kagel (2015). Closest to our work is the recent study by Bruhin et al. (2019), who use the same finite mixture approach as we do, but who do not consider Kantian morality.

[6] 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

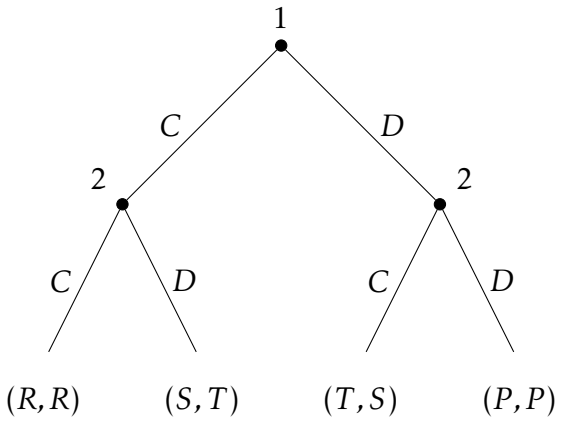
In the experiment, subjects play three types of well-known game protocols, illustrated in Figure 1: the Sequential Prisoner's Dilemma protocol (SPD), shown in Figure 1a, the mini Trust Game protocol (TG), shown in Figure 1b, and the mini Ultimatum Game protocol (UG), shown in Figure 1c.[7] We use the 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 j). Each strategy pair (x, y) determines the realization probability $\eta_{(x,y)}(\gamma)$ of each play γ 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 $\eta_{(x,y)}((2, D, C)) = \frac{(1-y_1)\cdot x_3}{2}$.

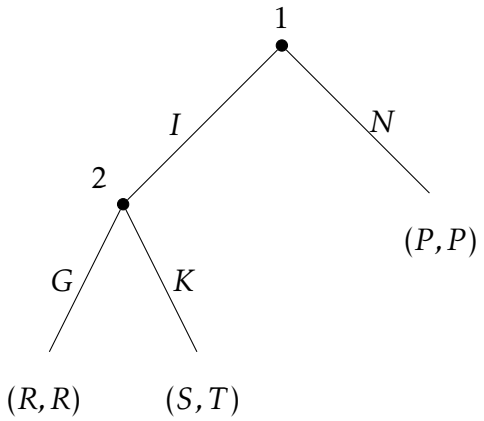
Turning to the two other game protocols, when the trust game protocol is symmetrically randomized, a behavior strategy is a vector, $x = (x_1, x_2) \in [0, 1]^2$, where x_1 is the

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

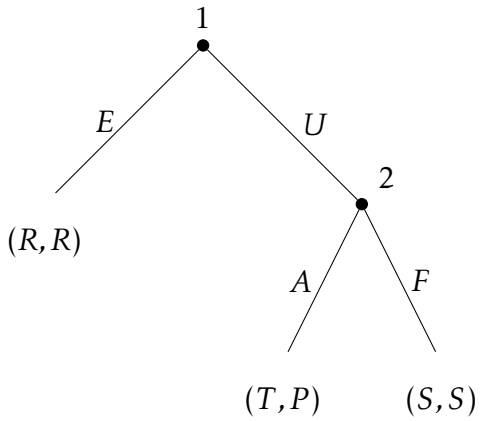
Figure 1: Game protocols



(a) Sequential Prisoner's Dilemma game protocol

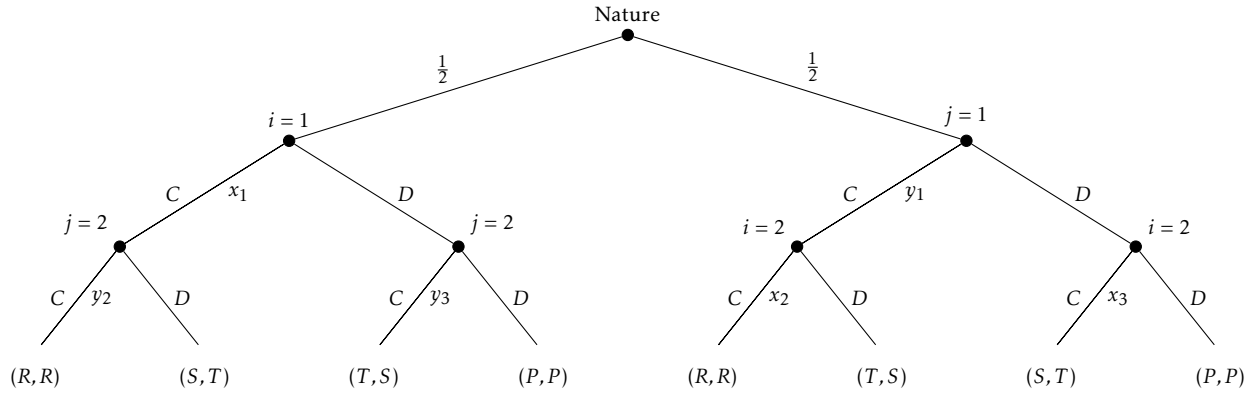


(b) Trust Game protocol



(c) Ultimatum Game protocol

Figure 2: Meta-game protocol for the SPD



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 $\eta_{(x,y)}(\gamma)$ to denote the probability of each play γ 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 Social preferences and Kantian morality

Let the expected utility of a subject i playing against a subject j be

$$\begin{aligned} u_i(x,y) \;=\; & (1-\kappa_i)\cdot\sum_{\gamma}\eta_{(x,y)}(\gamma)\cdot\pi_i(\gamma) \\ & -\alpha_i\cdot\sum_{\gamma}\eta_{(x,y)}(\gamma)\cdot\max\left\{0,\pi_{ij}(\gamma)-\pi_i(\gamma)\right\} \\ & -\beta_i\cdot\sum_{\gamma}\eta_{(x,y)}(\gamma)\cdot\max\left\{0,\pi_i(\gamma)-\pi_{ij}(\gamma)\right\} \\ & +\kappa_i\cdot\sum_{\gamma}\eta_{(x,x)}(\gamma)\cdot\pi_i(\gamma), \end{aligned} \tag{1}$$

where x and y are i's and j's behavior strategy, respectively, π_i is i's material utility following play γ and π_{ij} is j's material utility following play γ. This utility function has three

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 $\alpha_i < 0$) from disadvantageous inequity, i.e., from falling short in terms of material utility 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 utility. The third parameter, κ_i, captures a Kantian moral concern (à la *Homo moralis*, Alger & Weibull, 2013). It places weight on the expected material utility 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 γ would occur is $\eta_{(x,x)}(\gamma)$. 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 utility 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 utility if used by both subjects (see Alger & Weibull, 2013, for a discussion).

The utility function in (1) nests many familiar utility functions in the literature. Clearly, setting all three parameters to zero, $\alpha_i = \beta_i = \kappa_i = 0$, 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 $\kappa_i = 0$. One obtains Becker's (1974) model of pure altruism by setting $\kappa_i = 0$ and $\alpha_i = -\beta_i$, for some $\beta_i \in (0,1)$.[8] Here β_i is the individual's "degree of altruism", the weight placed on the other subject's material utility, while the weight $1 - \beta_i$ is placed on own material utility. Pure *Homo moralis* preferences are obtained by setting $\alpha_i = \beta_i = 0$ and $\kappa_i \in (0,1)$. Here κ_i is the individual's "degree of Kantian morality", the weight placed on the material utility 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 utility, given the strategy profile (x, y) effectively played. The utility function in (1) also nests the Charness and Rabin (2002) model without reciprocity. In Section 4.4 we extend the utility function to also accommodate reciprocity as formalized in Charness and Rabin (2002).

Because 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 allow for risk aversion. Thus, the term $\pi_i(\gamma)$ in equation (1) is the Bernoulli function value that the individual attaches to his or her monetary payoff under

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

play γ. We will call $\pi_i(\gamma)$ the individual's *material utility* under play γ. If the monetary payoff allocation after a play γ is $\big(m_i(\gamma), m_j(\gamma)\big)$, we assume that the individual's own material utility is of the CRRA form

$$\pi_i(\gamma) = \frac{m_i(\gamma)^{1-r_i} - 1}{1 - r_i}, \tag{2}$$

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.[9] Hence, subject i evaluates the opponent j's monetary payoff as follows:

$$\pi_{ij}(\gamma) = \frac{m_j(\gamma)^{1-r_i} - 1}{1 - r_i}. \tag{3}$$

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 $\pi_i(\gamma) = \ln m_i(\gamma)$ and $\pi_{ij}(\gamma) = \ln m_j(\gamma)$.

2.3 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*. To see why, consider a dictator 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 $m > 1$. 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 risk-neutral pure altruist i, with $\beta_i = -\alpha_i \geq \kappa_i = 0$, and thus a utility function of the form (the factor $1/2$ represents nature's draw of roles):

$$u_i(x, y) \;=\; \frac{1}{2}[(1 - \beta_i)(w - x + my) + \beta_i(mx + w - y)]. \tag{4}$$

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

$$u_i(x, y) \;=\; \frac{1}{2}[(1 - \kappa_i)(w - x + my) + \kappa_i(mx + w - x)]. \tag{5}$$

[9] 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).

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 altruists and Kantian moralists is the same here:

$$\frac{du_i(x,y)}{dx} = \frac{1}{2}[\beta_i m - (1-\beta_i)], \tag{6}$$

and

$$\frac{du_i(x,y)}{dx} = \frac{1}{2}(\kappa_i m - 1). \tag{7}$$

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 (6) 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 $1/m$.[10] Therefore, we would be unable to separate altruism from a Kantian concern using dictator games.[11]

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. 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). 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 altruism, inequity aversion or spite, which induce consequentialistic reasoning.

[10] 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).

[11] 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 $(\alpha_i \cdot \kappa_i > 0)$.

Concretely, consider first a (symmetrically randomized) Trust Game protocol (see Figure 1b) with $2R > T + S$, and suppose that an individual i believes that the opponent will play K ("keep") as second-mover. If this individual i has no Kantian morality and is either selfish or driven by behindness aversion ($\alpha_i > 0$), he will choose N ("not invest") as first-mover. By contrast, if he has Kantian morality of a sufficiently large degree κ_i, then he will, as first-mover, choose I ("invest"), because he would himself play G ("give back") as second mover. Likewise, in the (symmetrically randomized) Sequential Prisoner's Dilemma protocol (Figure 1a), suppose that $2R > T + S > 2P$ and consider a subject who believes that the other will choose D both as first-mover and as second-mover. Despite this belief, a subject i with a large enough degree of Kantian morality would nevertheless evaluate the play C followed by C, because this is the play he would choose if he met himself.

Turning finally to the Ultimatum Game protocol, as in Figure 1c, we use it to conduct a formal analysis of the effect of Kantian morality (a formal analysis of the other two game protocols is provided in Appendix A1). A risk-neutral subject i obtains the following expected utility from using behavior strategy $x = (x_1, x_2)$ when he believes that the opponent will use behavior strategy $\hat{y} = (\hat{y}_1, \hat{y}_2)$ (the randomization factor 1/2 has been omitted):

$$\begin{aligned} u_i(x,\hat{y}) \;=\; & (1-\kappa_i)[x_1 R + (1-x_1)\hat{y}_2 T + (1-x_1)(1-\hat{y}_2) S \\ & + \hat{y}_1 R + (1-\hat{y}_1) x_2 P + (1-\hat{y}_1)(1-x_2) S] \\ & - [\alpha_i (1-\hat{y}_1) x_2 + \beta_i (1-x_1)\hat{y}_2](T-P) \\ & + \kappa_i [x_1 R + (1-x_1) x_2 T + (1-x_1)(1-x_2) S \\ & + x_1 R + (1-x_1) x_2 P + (1-x_1)(1-x_2) S]. \end{aligned} \tag{8}$$

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

$$\begin{aligned} \frac{\partial u_i(x,\hat{y})}{\partial x_1} \;=\; & (1-\kappa_i)[R - \hat{y}_2 T - (1-\hat{y}_2) S] + \beta_i \cdot \hat{y}_2 (T-P) \\ & + \kappa_i \cdot [2(R-S) - x_2 (T + P - 2S)] \end{aligned} \tag{9}$$

$$\frac{\partial u_i(x,\hat{y})}{\partial x_2} = (1-\kappa_i)(1-\hat{y}_1)(P-S) - \alpha_i \cdot (1-\hat{y}_1)(T-P) + \kappa_i \cdot (1-x_1)(T+P-2S). \tag{10}$$

To see the two key effects of Kantian morality mentioned above, we compare an individual who is inequity averse but does not have a Kantian concern ($\kappa_i = 0$) to one who has a Kantian concern but is not inequity averse ($\alpha_i = \beta_i = 0$). 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 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 shows up in (10) if $\kappa_i \neq 0$). Second, the expressions (9) and (10) 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 $\kappa = 1 > \alpha = \beta = 0$, 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.

Clearly, disentangling an individual's social preferences from his or her Kantian moral preferences requires controlling for his or her beliefs about the opponent's play. We therefore elicit subjects' such beliefs (by way of the quadratic scoring rule). We describe the experimental procedures, including the belief elicitation procedure, in the next subsection.

2.4 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), 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. The order of the game protocols was randomly determined at the beginning of each session. For each game protocol, subjects first indicated what they would do at each decision node and second what they believed others would do at each decision node. In all game protocols, we used neutral labels. Two of the 18 game protocols were randomly selected for payment. 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.

[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 original parameters.

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.

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 A4. 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.[13]

Prior to describing how we will analyze the data, we present some descriptive statistics.

2.5 Descriptive statistics

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 trustworthiness (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 A3). Table A.2 in Appendix A3 presents all decisions in the risk elicitation task. Based on their lottery choice, most subjects (83%) are classified as being risk-averse.

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 control for the 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.

[13] See https://aspredicted.org/blind.php?x=4u5nu8.

3.1 Individual preferences

For each subject i, we estimate the individual's social and moral preference parameters α_i, β_i, and κ_i as specified in (1), using a standard additive error specification. We refer to these preference parameters using the vector $\theta_i = (\alpha_i, \beta_i, \kappa_i)$. For each individual, we infer the risk parameter r_i from the lottery choices in the Eckel and Grossman (2002) task (see Table A.2 in Appendix A3). As robustness checks, we also estimate r_i alongside the other parameters and we carry out the analysis under the alternative assumption that all subjects are risk neutral (all $r_i = 0$), see Section 4.3. 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):

$$p_i(x_i, \hat{y}_i, \theta_i, \lambda_i) = \frac{\exp\left[(u_i(x_i, \hat{y}_i, \theta_i))/\lambda_i\right]}{\sum_{x' \in X_g} \exp\left[(u_i(x', \hat{y}_i, \theta_i))/\lambda_i\right]}, \tag{11}$$

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 $\theta_i = (\alpha_i, \beta_i, \kappa_i)$ and the "noise" parameter λ_i for each individual i.[14] Then, the probability

[14] In the maximum likelihood estimations, we use 7 different starting values for each parameter, so $7^4 = 2,401$ starting values per individual i.

density function can be written as:

$$f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_i, \lambda_i) = \prod_{g \in G} \prod_{x \in X_g} p_i(x, \hat{y}_i, \theta_i, \lambda_i)^{I(i,g,x)}, \tag{12}$$

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 $\theta_k = (\alpha_k, \beta_k, \kappa_k)$ and the noise parameter λ_k.

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 under the assumption that all subjects have logarithmic utility over monetary outcomes (i.e. we impose $r_k = 1$ for all types k). Given that most subjects in our experiment are risk averse according to the lottery task, assuming homogeneous risk aversion seems a better approximation of the data than assuming homogeneous risk neutrality. In subsection 4.3 we relax this assumption and also run estimations where we estimate the CRRA parameter r_k alongside the social preference and morality parameters. As an additional robustness check, we also run the estimations imposing risk-neutrality (i.e. $r_k = 0$ for all types k).

The log-likelihood is then given by:

$$\ln L = \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \phi_k \cdot f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_k, \lambda_k) \right), \tag{13}$$

where ϕ_k is the population share of type k in the population. To maximize the log-likelihood in (13), we use an Expectation-Maximization (EM) algorithm (see for instance

McLachlan, Lee, & Rathnayake, 2019).[15] As part of the EM algorithm, we estimate the posterior probabilities $\tau_{i,k}$ that individual i belongs to type k by:

$$\tau_{i,k} = \frac{\phi_k \cdot f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_k, \lambda_k)}{\sum_{m=1}^{K} \phi_m \cdot f(\mathbf{x}_i, \hat{\mathbf{y}}_i, \theta_m, \lambda_m)}. \tag{14}$$

4 Results

4.1 Individual preferences

Figure 3 shows the marginal distributions of the estimated individual preference parameters α_i, β_i, and κ_i.[16] For all three parameters, we observe considerable heterogeneity. Most estimates of α_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 both α_i and κ_i estimates). By contrast, most estimates of β_i are negative, and this is again confirmed by a signed-rank test ($p = 0.003$). Hence, we find that most subjects are motivated by a combination of Kantian morality ($\kappa_i > 0$) and spite ($\alpha_i > 0, \beta_i < 0$).

Table 2, which shows summary statistics for the parameter estimates, provides further support for the pattern observed in Figure 3. Median and mean estimates are positive for α_i and κ_i, but negative for β_i. Moreover, the relatively large standard deviations indicate that there is considerable heterogeneity in social preferences and Kantian morality.[17]

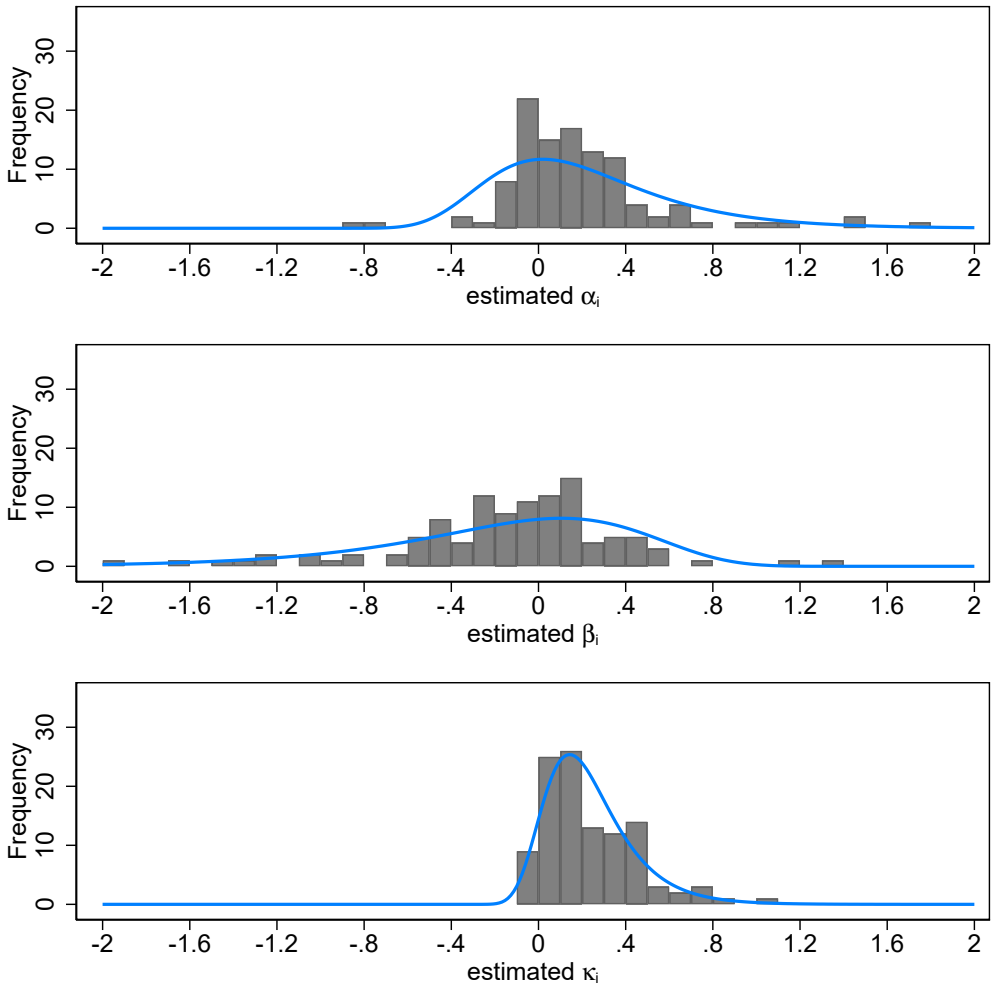
Figure 4 illustrates the pairwise correlations between the three preference parameter estimates. The left panel of Figure 4 shows that the estimates for α_i and β_i are negatively correlated (Spearman's $\rho = -0.295$, $p = 0.002$, $n = 109$), and again that there is substantial heterogeneity. For many individuals we observe a combination of $\alpha_i > 0$ and $\beta_i < 0$,

[15] We use 24 sets of starting values.

[16] 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 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 109 subjects for whom all three preference parameter estimates lie between -2 and 2. The fraction that we leave out in the main text (19,6%) 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 A3 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.

[17] For these estimates we used the risk elicitation task to determine r_i. In subsection 4.3 we provide robustness tests where we estimate r_i alongside the preference parameter, or impose risk neutrality.

Figure 3: Distributions of individual parameter estimates



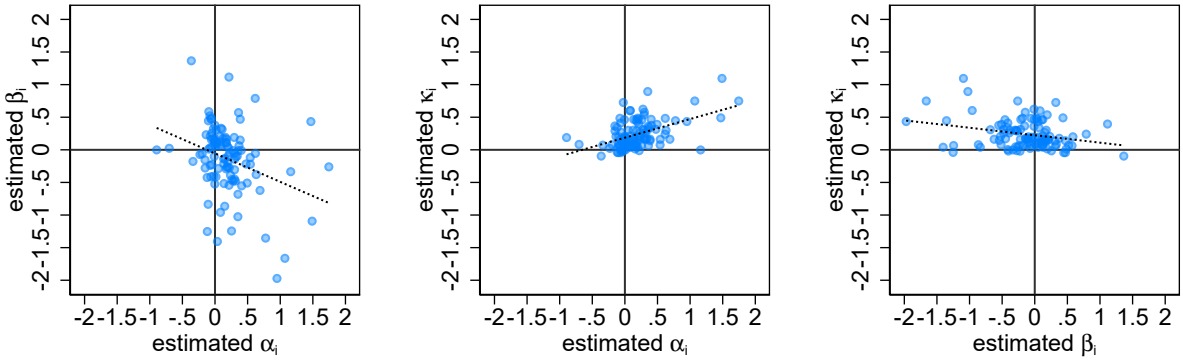
Note: Figure based on the 109 subjects for whom the α_i, β_i and κ_i estimates have absolute value below 2. The (blue) lines indicate fitted Gumbel distributions (see Appendix A2 for details). Figure A.2 shows a similar figure based on all 136 subjects.

Table 2: Individual parameter estimates

Parameter	Median	Mean	S.D.	Min	Max
α_i	0.14	0.19	0.38	−0.89	1.75
β_i	-0.06	−0.14	0.51	−1.97	1.37
κ_i	0.18	0.24	0.22	−0.10	1.10

Notes: Table based on the 109 subjects for whom the α_i, β_i and κ_i estimates have absolute value below 2. Table A.3 shows a similar table based on all 136 subjects.

Figure 4: Correlations between estimated preference parameters.



Notes: Each dot represents one subject. Dotted lines indicate linear predictions (intercept+slope). Specifically, we estimate $\beta_i = -0.05 - 0.44\alpha_i$, $\kappa_i = 0.19 + 0.28\alpha_i$ and $\kappa_i = 0.22 - 0.11\beta_i$. Figure based on the 109 subjects for whom the α_i, β_i and κ_i estimates have absolute value below 2.

in line with spiteful/competitive preferences, i.e., an individual dislikes being behind but likes being ahead of the other. The middle panel of Figure 4 reveals a strong and positive correlation between α_i and κ_i estimates (Spearman's $\rho = 0.423$, $p < 0.001$, $n = 109$). This means that many individuals combine a distaste for disadvantageous inequity, or, as Bruhin et al. (2019) call it, "behindness aversion," with Kantian morality. For the estimates of β_i and κ_i we find a negative correlation (Spearman's $\rho = -0.173$, $p = 0.071$, $n = 109$). We also use copula methods to describe the joint parameter 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 A2 provides more details.

Table 3: 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	0.14	0.06	0.27	−0.01	0.12	0.27
	(0.02)	(0.03)	(0.03)	(0.05)	(0.08)	(0.03)
β_k	0.00	0.09	−0.31	−0.07	0.24	−0.31
	(0.03)	(0.04)	(0.06)	(0.07)	(0.06)	(0.08)
κ_k	0.21	0.23	0.18	0.21	0.23	0.18
	(0.01)	(0.02)	(0.02)	(0.03)	(0.06)	(0.02)
λ_k	0.25	0.28	0.16	0.32	0.22	0.15
	(0.02)	(0.02)	(0.01)	(0.05)	(0.03)	(0.02)
ϕ_k	1.00	0.62	0.38	0.33	0.30	0.37
	(-)	(0.05)	(0.05)	(0.07)	(0.08)	(0.05)
$\ln L$	-2336.9	-2154.4		-2131.7		
$EN(\boldsymbol{\tau})$	0.00	4.25		17.80		
ICL	4692.6	4355.3		4346.9		
NEC	-	0.023		0.087		

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 109 subjects. In these estimations, we impose $r_k = 1$ (i.e. logarithmic utility) for all types. Table A.4 in Appendix A3 shows estimates based on the full sample. Table A.5 in Appendix A3 shows the estimates of a 4-type model.

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 3 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 $\alpha_0 = 0.14$, $\beta_0 = 0.00$, and $\kappa_0 = 0.21$, where the index 0 stands for the representative agent. In other words, the representative agent dislikes disadvantageous inequity, is indifferent with respect to advantageous inequality, and has a positive degree of Kantian morality.

The representative agent thus exhibits Kantian morality and behindness aversion.

4.2.2 The two- and three-type models

As can be seen in Table 3, 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 much stronger heterogeneity in terms of the inequity aversion parameters α_k and β_k: some types exhibit behindness aversion ($\alpha_k > 0$) while other types are (close to) indifferent to behindness ($\alpha_k \approx 0$); and some types disliking behind ahead ($\beta_k > 0$) while other types like it ($\beta_k < 0$).

More specifically, when assuming two types, the most common type (Type 1) exhibits (mild) inequity aversion, with parameter estimates $\alpha_1 = 0.06$ and $\beta_1 = 0.09$, combined with a degree of Kantian morality $\kappa_1 = 0.23$. This type represents about 62% of the subjects. The other type, Type 2, exhibits a combination of strong *spite* ("negative altruism") and Kantian morality, with $\alpha_2 = 0.27$, $\beta_2 = -0.31$, and $\kappa_2 = 0.18$. [18]

For each subject i, we estimate the posterior probability $\tau_{i,k}$ that i belongs to type k (as defined in (14)). 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.6 in Appendix A3 lists the chosen strategies per game protocol type based on this classification. "Type 2 subjects", who combine spite and Kantian morality, mostly choose to always defect (D, D, D) in the SPDs (in 87%) of the cases, while "Type 1 subjects", who combine inequity aversion and Kantian morality, choose (D, D, D) less frequently (38%) and often conditionally cooperate (C, C, D) instead (32%). Similarly, in the TGs, Type 2 subjects most frequently choose not to invest as first mover and to "keep" as second mover (N, K) (85%), 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 (74%) and accept unfair offers as a second mover (97%). Instead, Type 1 subjects most frequently propose an equal payoff (68%) and accept fewer unequal offers (88%).

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 spite with Kantian morality, and represents a similar fraction of the population (37%).[19] The

[18] The finding that a sizeable share of the subjects (here 38%) are both spiteful ($\alpha_k > 0$ and $\beta_k < 0$) and moral ($\kappa_k > 0$) agrees with a recent theoretical result that preference evolution in some settings leads to a combination of self-interest, spite and Kantian morality (see Alger et al., 2020).

[19] In panel A of Table A.7 (see Appendix A3), we show a transition matrix for the two-types and three-

new Type 2 combines (relatively strong) inequity aversion with Kantian morality. It represents around 17% of the population. Type 1 is very close to *Homo moralis*. The inequity aversion parameters α_1 and β_1 are not significantly different from zero (at the 5% level), while the Kantian morality κ_1 is positive and significant. This type represents 33% of the population. In sum: under the three-types approach, Type 1 displays Kantian morality, Type 2 is inequity averse and moral, and Type 3 is spiteful and moral.

In terms of chosen strategies, Type 3 behaves almost identical as Type 2 in the two-types model. The new Types 1 and Type 2 differ in some respects. In the SPDs, the new 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, Type 2 refuses unequal offers more frequently than Type 1.

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 0.2, 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 ($\alpha_k = \beta_k = \kappa_k = 0$).[20] 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

$$ICL = -2\ln L + d\ln N + EN(\boldsymbol{\tau}), \tag{15}$$

where the log-likelihood function $\ln L$ is defined as in (13), d is the number of estimated parameters, and N is the number of individuals in our sample. The last term in (15) is

types models. All but one subject who is 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.

[20] This is in line with the findings by Bruhin et al. (2019).

the entropy

$$EN(\boldsymbol{\tau}) = -\sum_{k=1}^{K}\sum_{i=1}^{N}\tau_{i,k}\ln\tau_{i,k}, \tag{16}$$

where $\tau_{i,k}$ is the estimated posterior probability of individual i belonging to type k, as defined in (14). This implies that the stronger individuals are assigned to types (i.e. all $\tau_{i,k}$'s 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 5 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, some individuals are imprecisely classified to either Type 1 or Type 2.

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

$$NEC = \frac{EN(\boldsymbol{\tau})}{\ln L(K) - \ln L(1)}, \tag{17}$$

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.

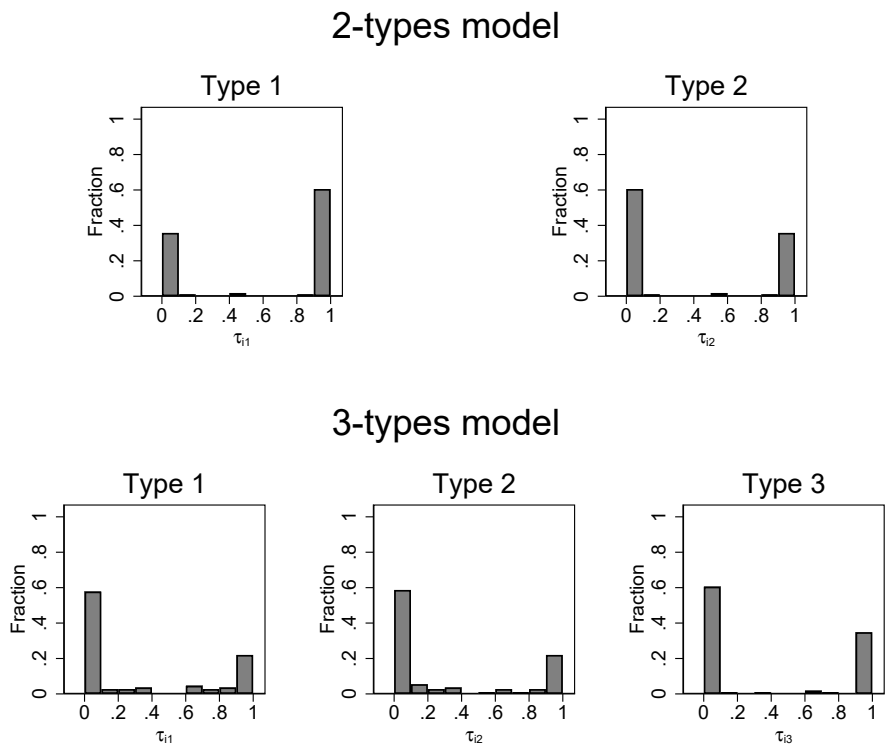
Table 3 shows statistics for both the ICL and the NEC. For both metrics, a lower score indicates a more preferred model. The NEC selects the two-types model and the ICL selects the three-types model. Table A.5 in Appendix A3 shows estimates and goodness-of-fit metrics for a four-types model. The four-types model performs worse on both criteria than the two-types and three-types models in Table 3. 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

4.3.1 Estimating risk attitudes

In the main analysis, we imposed values for the CRRA parameter r. In the individual estimations, we based r_i on the decision in the lottery task while in the aggregate esti-

Figure 5: 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 3.

mations we assumed that everyone has logarithmic utility (i.e. $r_k = 1$). We also estimate the CRRA parameter r alongside the social preference and Kantian morality parameters. Doing so does not affect our estimates by much.

First, at the individual level, estimating r_i alongside the preference parameters (α_i, β_i, κ_i) does not affect the estimates by much. The estimated preference parameters are strongly correlated (Spearman rank correlations: $\rho = 0.639, p < 0.001$, $n = 109$ for α_i, $\rho = 0.566, p < 0.001$, $n = 109$ for β_i, and $\rho = 0.606, p < 0.001$ for κ_i) although the correlation between the imposed and estimated r_i values is weak ($\rho = 0.069, p = 0.478$). The estimates of α_i, β_i and κ_i are not systematically smaller or larger using either method (signed-rank tests, $p = 0.198$, $n = 109$ for α_i, $p = 0.228$, $n = 109$ for β_i, and $p = 0.388$, $n = 109$ for κ_i).

Second, estimations of the finite mixture models that include the estimation of a CRRA parameter r_k for each type k lead to a value for r_k close to 1 in most cases (see Table A.8 in Appendix A3). As a result, the estimated social preference and Kantian morality parameters change very little.[21]

4.3.2 Risk neutrality

In yet another robustness check we estimate the social preference and Kantian morality parameters under the alternative assumption that all subjects are risk neutral (i.e., $r_i = 0$ for all subjects i). Figure 6 shows scatter plots of individual parameter estimates under both assumptions, with estimates under risk neutrality on the horizontal axis and estimates under constant (individual specific) relative risk aversion (CRRA) 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 $\rho = 0.802$. For the inequity parameter β_i (when ahead) it is $\rho = 0.774$, and for the Kantian morality parameter κ_i it is $\rho = 0.627$ (all three rank correlations hold for $p < 0.001$, $n = 109$).

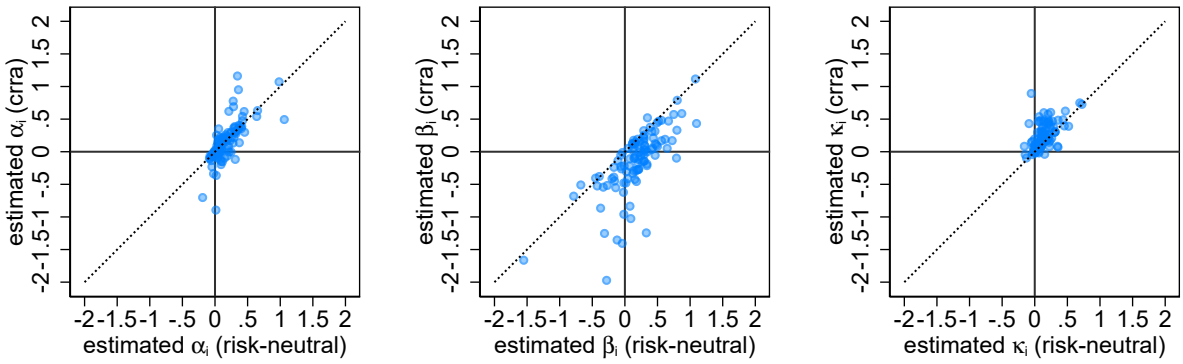
The middle panel in Figure 6 also shows that the β_i estimates are much higher under risk neutrality than under CRRA.[22] Indeed, for 94 out of 109 subjects, the risk-neutral estimate is higher than the CRRA estimate (signed-rank test, $p < 0.001$).[23] By contrast,

[21] Table A.7 shows that for both two-types (panel B) and three-types (panel C) models, subjects are separated in almost the same groups as when we impose $r_k = 1$.

[22] One can easily see how assuming risk neutrality would bias 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.

[23] Moreover, for most subjects (80 out of 109), β_i is positive under risk neutrality (signed-rank test, $p < 0.001$).

Figure 6: 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 109 subjects.

the risk-neutral estimates of κ_i (80 out of 109, signed-rank test: $p < 0.001$) and α_i (64 out of 109, signed-rank test: $p = 0.068$) are lower for most subjects than under CRRA.[24] For the majority of subjects (72 out of 109), assuming CRRA preferences instead of risk neutrality leads to a higher log-likelihood, indeed indicating a better fit under CRRA preferences.

Table 4 shows the estimates of finite mixture models under risk neutrality. Comparing these results with those in Table 3, one sees that, qualitatively, estimates of the parameters α_k and κ_k are not much affected. For all types in Tables 3 and 4, α_k and κ_k are positive, under both risk hypotheses, although the Kantian morality parameter values somewhat lower under risk neutrality than under CRRA. In line with the individual parameter estimates (see the middle panel in Figure 6), the finite mixture estimates of the parameters β tend to be much higher under risk neutrality than under CRRA. Moreover, under risk-neutrality, all estimates of β_k are non-negative, in contrast to the CRRA estimates, where we observed $\beta_k < 0$ for some types k.[25]

The ICL criterion allows comparison of the fit of the CRRA and risk-neutral models, respectively (see Tables 3 and 4). For any given number of types, the CRRA model has

[24] Most risk-neutral estimates of κ_i (96 out of 109) and α_i (92 out of 109) are positive (signed-rank tests, $p < 0.001$)

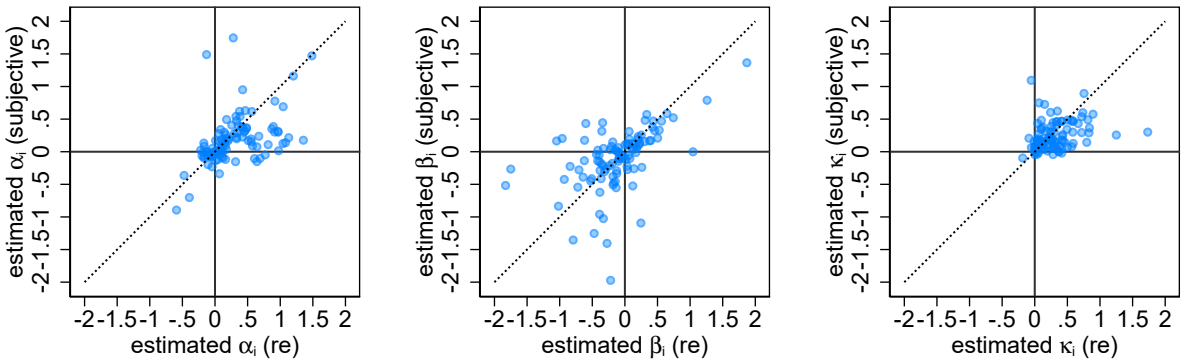
[25] Table A.7 shows that the assignment of subjects to types for the risk-neutral two-types (panel D) model, is very similar to when we impose $r_k = 1$. For the three-types models (panel E), the type classification is again similar under both assumptions, but some who are classified as "Type 2" with $r_k = 1$ are classified as "Type 1" under risk-neutrality.

Table 4: Estimates at the aggregate level (assuming risk neutrality)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.17	0.13	0.19	0.18	0.01	0.19
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)
β_k	0.25	0.36	0.00	0.26	0.50	0.00
	(0.03)	(0.05)	(0.04)	(0.06)	(0.07)	(0.03)
κ_k	0.10	0.11	0.11	0.11	0.14	0.10
	(0.01)	(0.02)	(0.01)	(0.02)	(0.05)	(0.01)
λ_k	7.62	8.98	4.01	9.29	6.92	3.79
	(0.60)	(0.95)	(0.51)	(1.17)	(0.79)	(0.36)
ϕ_k	1.00	0.64	0.36	0.48	0.18	0.34
	(-)	(0.07)	(0.07)	(0.06)	(0.06)	(0.05)
$\ln L$	-2426.8	-2247.6		-2217.9		
$EN(\boldsymbol{\tau})$	0.00	5.31		14.20		
ICL	4872.5	4542.7		4515.6		
NEC	-	0.030		0.068		

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 109 subjects.

Figure 7: 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 109 subjects.

a lower ICL score than the risk-neutral model. For the three-types model, for example, the ICL score under the CRRA assumption is quite a bit lower than under risk neutrality (4343.6 versus 4515.6), showing that the CRRA model considerably improves the fit over the risk-neutrality model.

4.3.3 Rational expectations

So far, we assumed that people maximize expected utility given their (reported) subjective expectations. In this subsection we investigate what happens to the estimated preference parameters if we take rational expectations instead.

Figure 7 shows correlations between the individual estimates using subjective and rational expectations. For all three preference parameters, the estimates under the two assumptions are strongly correlated. For the inequity parameter α_i (when behind) the Spearman rank correlation is $\rho = 0.553$. For the inequity parameter β_i (when ahead) it is $\rho = 0.635$, and for the Kantian morality parameter κ_i it is $\rho = 0.344$ (for all three rank correlations: $p < 0.001$, $n = 109$). For most subjects, the log-likelihood is larger when we assume rational expectations instead of subjective expectations (67 out of 109, signed-rank test: $p = 0.043$), indicating that assuming rational expectations actually improves the fit for most subjects.

Table 5 shows the finite mixture estimates when we assume rational expectations. The representative agent with rational expectations is characterized by a combination of spite ($\alpha_k > 0, \beta_k < 0$) and morality ($\kappa_k > 0$). Compared to the model with subjective expectations

Table 5: 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	0.30	0.08	0.67	0.04	0.13	0.68
	(0.05)	(0.03)	(0.07)	(0.05)	(0.10)	(0.09)
β_k	−0.28	0.04	−0.52	−0.03	0.19	−0.51
	(0.05)	(0.06)	(0.09)	(0.07)	(0.14)	(0.11)
κ_k	0.36	0.31	0.31	0.18	0.49	0.32
	(0.03)	(0.04)	(0.03)	(0.02)	(0.10)	(0.05)
λ_k	0.31	0.28	0.23	0.28	0.24	0.21
	(0.02)	(0.02)	(0.02)	(0.02)	(0.04)	(0.03)
ϕ_k	1.00	0.54	0.46	0.42	0.19	0.39
	(-)	(0.05)	(0.05)	(0.05)	(0.04)	(0.05)
$\ln L$	-2462.0	-2157.1		-2097.6		
$EN(\boldsymbol{\tau})$	0.00	2.58		7.52		
ICL	4942.7	4539.0		4270.1		
NEC	-	0.008		0.025		

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 109 subjects. For all types, we assume logarithmic utility ($r_k = 1$) and rational expectations.

(see Table 3), the estimates for α_k and κ_k are larger when we assume rational expectations. The estimate for β_k is negative when we assume rational expectations, where it was zero 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 the two-types model, Type 1 again displays a combination of (mild) inequity aversion and morality, Type 2 combines spite with morality.[26] The ICL scores of both multi-type models are somewhat lower under rational expectations, indicating a slightly better fit under rational expectations. Most importantly however, the estimated preference parameters for the multi-type models are very similar under both assumptions.

[26]Table A.7 (panels F and G) shows that the assignment of subjects to types is similar under subjective and rational expectations.

4.4 The value added of Kantian morality

In the preceding sections, we showed that estimated Kantian morality parameters tend to be positive, both at the individual and aggregate level. In this subsection, we benchmark the added value of the Kantian morality parameter against other parameters, and also against reciprocity.

4.4.1 Individual estimations

We conduct likelihood-ratio tests to see if adding the Kantian morality parameter κ_i to a model with only the two social preference parameters α_i and β_i improves the fit. The likelihood-ratio tests reveal that adding κ_i improves the fit for 21 individuals at the 5% level (and for 32 individuals at the 10% level). For comparison, likelihood ratio tests when adding either α_i to (β_i, κ_i), or β_i to (α_i, κ_i), improves the fit at the 5% level for 20 and 26 individuals, respectively (at the 10% level, for 25 (α_i) and 37 (β_i) individuals). Hence, in terms of value added at the individual level, all three preference parameters are in roughly the same ballpark.

A more general approach is to consider all models that are nested in (1) and apply standard information criteria. 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:

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

and

$$AIC = -2\ln(L) + 2d, \tag{19}$$

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

Table 6 shows the results. The left panel shows which model provides the best fit according to BIC. For 37 subjects (33.9%) pure self-interest ($\alpha_i = \beta_i = \kappa_i = 0$) has the lowest BIC score. For the remaining 72 subjects, some combinations of social preferences and/or moral concerns improve the model's fit. For 23 subjects, (21.1%) pure *Homo moralis* preferences ($\alpha_i = \beta_i = 0$, $\kappa_i \neq 0$) provides the best individual fit. For another 11 subjects, models with κ_i in combination with α_i and/or β_i have the lowest BIC scores. In sum, for 34 subjects (31.2%), the model with the lowest BIC score includes κ_i. In comparison, α_i and β_i are included in the model with the lowest BIC score for 23 subjects (21.1%) and 35

Table 6: Best individual fit

	BIC		AIC	
Parameters	Frequency	Percentage	Frequency	Percentage
$\alpha_i, \beta_i, \kappa_i$	2	1.8	6	5.5
α_i, β_i	7	6.4	6	5.5
α_i, κ_i	5	4.6	8	7.3
β_i, κ_i	4	3.7	10	9.2
α_i	9	8.3	11	10.1
β_i	22	20.2	18	16.5
κ_i	23	21.1	24	22.0
-	37	33.9	26	23.9

Notes: Entries indicate the number of subjects for whom the specific model provides the lowest BIC or AIC score respectively. Table based on our 'core sample' of 109 subjects.

subjects (32.0%), respectively. The right panel 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 48 subjects (or 44.0%). Again, a larger number of subjects than for α_i (31 subjects, or 28.4%) and also slightly more subjects than β_i (40 subjects, or 36.7%).

4.4.2 Aggregate estimations

We also evaluate the value added of Kantian morality for the finite mixture estimations. Table A.9 in Appendix A3 shows estimates for finite mixture models with only α_k and β_k (i.e. where $\kappa_i = 0$). For any given number of types, these fixed mixture estimates give substantially higher ICL scores than the model including Kantian morality, indicating that fixed mixture estimates that include the parameter κ_i provide a better fit.

4.4.3 Reciprocity vs. Kantian morality

We can compare the value added of the Kantian morality parameter, to the value if one were to instead of Kantian morality add reciprocity. For this purpose, we modify the utility function in (1) to replace the Kantian morality term by a term that represents

negative reciprocity as in Charness and Rabin (2002), which leads to

$$
\begin{aligned}
u_i(x,y) \;=\; & \sum_{\gamma} \eta_{(x,y)}(\gamma)\cdot \pi_i(\gamma) \\
& - \alpha_i \cdot \sum_{\gamma} \eta_{(x,y)}(\gamma)\cdot \max\left\{0, \pi_{ij}(\gamma) - \pi_i(\gamma)\right\} \\
& - \beta_i \cdot \sum_{\gamma} \eta_{(x,y)}(\gamma)\cdot \max\left\{0, \pi_i(\gamma) - \pi_{ij}(\gamma)\right\} \\
& - \delta_i \cdot q \cdot \sum_{\gamma} \eta_{(x,y)}(\gamma)\cdot \max\left\{0, \pi_{ij}(\gamma) - \pi_i(\gamma)\right\},
\end{aligned}
\tag{20}
$$

where $q = 1$ if the other player 'misbehaved' and $q = 0$ otherwise. Following Charness and Rabin (2002), we label a first-mover action as misbehavior if it excludes an outcome that has maximal joint monetary payoffs. For our case this means that defecting as a first mover in a SPD protocol (if $2R > T+S$, which holds for 5 out of 6 SPDs), and not investing in a TG protocol constitutes misbehavior (note, however, that the δ_i term cancels in 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.

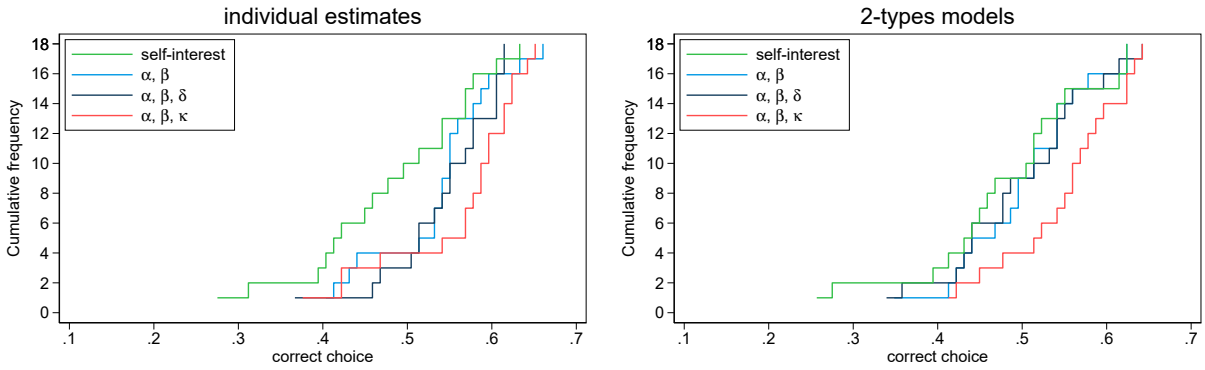
In Table A.10 in Appendix A3 we provide the results of finite mixture models based on (20). The three-types model has the lowest ICL score among the reciprocity models. Based on the ICL score, the three-types reciprocity model performs better than the mixture models with only α_k and β_k (see A.9). This shows that, adding reciprocity improves the fit of the model. Importantly however, the three-types model that allows for Kantian morality instead of reciprocity has an even lower ICL score, suggesting that Kantian morality adds more than reciprocity in our setting.

4.4.4 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 8 illustrates the results, by comparing the predictive accuracy of the model with α, β and κ, to self-interest ($\alpha = \beta = \kappa = 0$), a model without Kantian morality (α, β) and a model allowing for negative reciprocity (α, β, δ). The left panel of Figure 8

Figure 8: Accuracy of out-of-sample predictions



Notes: Accuracy of out-of-sample predictions, based on individual estimates (left panel) and finite mixture models with two-types (right panel). Plots show cumulative frequency plots for the average fraction of correctly predicted choices per game protocol. Figure based on our 'core sample' of 109 subjects.

compares the predictive accuracy based on individual estimates. All models clearly outperform random choice (which would lead to 20.8% accurate predictions in expectation). The model allowing for Kantian morality (α, β, κ) outperforms the other models in terms of predictive accuracy. The (α, β, κ)-model on average predicts 56.0% of choices correctly, somewhat more than the (α, β) and (α, β, δ) models, which give 53.4% and 54.3% average accuracy, respectively. All models allowing for social preferences and/or morality perform much better than when assuming self-interest, which gives 48.1% average accuracy.

The right panel of Figure 8 shows the predictive accuracy of finite mixture models assuming two types. Compared to the individual estimations, the gap between the model allowing for Kantian morality (α, β, κ) and the other models is larger for the two-types models. The two-types model with α, β and κ predicts 54.8% of choices correctly, which is better than the two-types (α, β)-model and reciprocity model (α, β, δ) which give 50.2% and 49.8% accuracy respectively. Note that the predictive accuracy of the two-types model allowing for Kantian morality (54.8%) is not far from the model allowing for Kantian morality with individual estimates (56.0%).[27] This provides further evidence that the model effectively captures the heterogeneity in preferences.

[27] When allowing for Kantian morality (α, β, κ), a model with a representative agent (1 type) performs worse (51.4% accuracy) than the two-types model, and a model with three-types performs only slightly better (55.4%) than the two-types model.

5 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, and inequity aversion, and of course Kantian morality. We structurally estimate the preference parameters of this utility function, allowing for risk aversion and 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 62% of the subjects display a combination of mild inequity aversion with Kantian morality, and the remaining 38% a combination of Kantian morality and strong spite. Within the three-types model, again one type is characterized by a combination of inequity aversion and Kantian morality (representing 30% of the population) and 37% of the population appear to combine Kantian morality with spite. However, now there is another type which displays only Kantian morality, representing 33% of the population. Quite remarkably, 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_k of around 0.2, which given the posited utility function means that the weight attached to the Kantian moral concern is about one quarter of the weight attached to the own material payoff.

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, for a recent survey). This literature shows that evolution by natural selection favors Kantian morality (see, in particular, Bergstrom (1995) and Alger and Weibull (2013)). As it turns out, our results are in fact in line with an even more recent contribution to this theoretical literature. In a model that enables analysis of the long-run impact of population structure on preferences, Alger et al. (2020) show that preferences that combine Kantian morality with either altruism or spite are favored by evolution by natural selection.[28]

[28]This result does not contradict that of Alger and Weibull (2013), which is shown by Alger et al. (2020)

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 involved, 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.

As for all laboratory experiments, establishing external validity would be highly desirable (Levitt & List, 2007). 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). Also, while our experiment was conducted on a WEIRD population (Henrich, Heine, & Norenzayan, 2010), evolutionary theory suggests that the qualitative nature of preferences guiding behavior in strategic interactions should be similar across the world, although 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).

References

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).

to also hold in their model when preferences are expressed with respect to effects of behavior on own and others' *fitness*. The result by Alger et al. (2020) that preferences favored by natural selection combine Kantian morality with either altruism or spite was obtained for preferences expressed with respect to effects of behavior on own and others' *material payoffs* (even marginal such effects).

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.

Bergstrom, T. C. (1995). On the evolution of altruistic ethical rules for siblings. *American Economic Review*, *85*(1), 58–81.

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 competi-

tion. *American Economic Review*, *90*(1), 166–193.

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

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.

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

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.

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

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.

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.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29.

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.

Levitt, S. D., & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, *21*(2), 153–174.

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.

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.

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

Appendices (For Online Publication)

Appendix A1 Distinguishing Kantian morality from social preferences

The Ultimatum Game protocol having been analyzed in detail in the main text (Subsection 2.3), we here analyze the other two game protocols. Throughout we assume risk neutrality; this is only for notational simplicity, the only difference being that the monetary payoffs would be replaced by the associated monetary utilities.

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).[29] 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 $1/2$):

$$\begin{aligned} u_i(x,y) \;=\; & (1-\kappa_i)[x_1[y_2R + (1-y_2)S] + (1-x_1)P] \\ & + (1-\kappa_i)[y_1[x_2R + (1-x_2)T] + (1-y_1)P] \\ & + \kappa_i\{x_1[x_2R + (1-x_2)S] + (1-x_1)P\} \\ & + \kappa_i\{x_1[x_2R + (1-x_2)T] + (1-x_1)P\} \\ & - [\alpha_i x_1(1-y_2) + \beta_i y_1(1-x_2)](T-S). \end{aligned} \tag{21}$$

Hence, for a subject who believes that the opponent plays $\hat{y}$:

$$\frac{\partial u_i(x,\hat{y})}{\partial x_1} = (1-\kappa_i)[S - P + \hat{y}_2(R-S)] + \kappa_i[x_2(2R-S-T) + S + T] - \alpha_i(1-\hat{y}_2)(T-S), \tag{22}$$

and

$$\frac{\partial u_i(x,\hat{y})}{\partial x_2} = (1-\kappa_i)\hat{y}_1(R-T) + \kappa_i x_1(2R-S-T) + \beta_i\hat{y}_1(T-S). \tag{23}$$

The social preference parameters α_i and β_i represent consequentialistic motives: they give weight to the monetary payoff consequences given what the subject believes about the opponent's actual play. By contrast, the Kantian morality parameter κ_i captures a deontological motive, such as "duty" or "to do the right thing", which ((following Alger & Weibull, 2013) we take to be to evaluate one's strategy in the light of what would happen if, hypothetically, the opponent would also use the same strategy.

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

Turning now to the Sequential Prisoners' Dilemma game protocol (as in Figure 1a), denote by x_1 the probability of playing C when moving first, x_2 the probability of playing C when moving second after play of C by the opponent, and and x_3 the probability of playing C when moving second after play of D by the opponent. Hence, the vector $x = (x_1, x_2, x_3) \in [0,1]^3$ is the player's behavior strategy in the symmetrically randomized sequential prisoners' dilemma. Then 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 (again omitting the factor $1/2$):

$$\begin{aligned} u_i(x,y) &= (1-\kappa_i)[x_1 y_2 R + x_1(1-y_2)S + (1-x_1)y_3 T + (1-x_1)(1-y_3)P] \\ &\quad + (1-\kappa_i)[y_1 x_2 R + y_1(1-x_2)T + (1-y_1)x_3 S + (1-y_1)(1-x_3)P] \\ &\quad + \kappa_i[x_1 x_2 R + x_1(1-x_2)S + (1-x_1)x_3 T + (1-x_1)(1-x_3)P] \\ &\quad + \kappa_i[x_1 x_2 R + x_1(1-x_2)T + (1-x_1)x_3 S + (1-x_1)(1-x_3)P] \\ &\quad - \alpha_i[x_1(1-y_2) + (1-y_1)x_3](T-S) \\ &\quad - \beta_i[(1-x_1)y_3 + y_1(1-x_2)](T-S). \end{aligned} \tag{24}$$

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

$$\begin{aligned} \frac{\partial u_i(x,\hat{y})}{\partial x_1} &= (1-\kappa_i)[S - P + \hat{y}_2(R-S) - \hat{y}_3(T-P)] \\ &\quad + \kappa_i[x_2(2R-S-T) + (1-x_3)(S+T-2P)] \\ &\quad + \beta_i \hat{y}_3(T-S) - \alpha_i(1-\hat{y}_2)(T-S), \end{aligned} \tag{25}$$

$$\frac{\partial u_i(x,\hat{y})}{\partial x_2} = (1-\kappa_i)\hat{y}_1(R-T) + \kappa_i x_1(2R-S-T) + \beta_i \hat{y}_1(T-S), \tag{26}$$

and

$$\frac{\partial u_i(x,\hat{y})}{\partial x_3} = (1-\kappa_i)(1-\hat{y}_1)(S-P) + \kappa_i(1-x_1)(T+S-2P) - \alpha_i(1-\hat{y}_1)(T-S). \tag{27}$$

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 A2 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

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

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

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

The empirical distributions of α_i and κ_i have a relatively long right tail (see 3, 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:

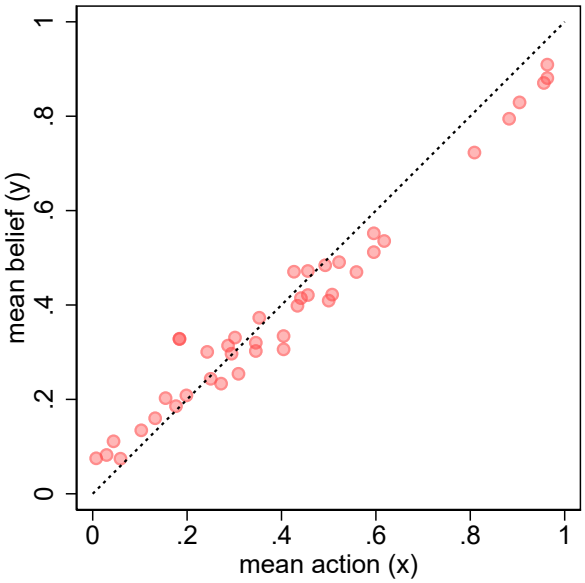
$$\begin{aligned} C\left(F_\alpha\left(x_\alpha\right), F_{-\beta}\left(x_{-\beta}\right), F_\kappa\left(x_\kappa\right)\right) \;=\; & \\ & \exp\left(-\left[\left(-\ln F_\alpha\left(x_\alpha\right)\right)^\omega + \left(-\ln F_{-\beta}\left(x_{-\beta}\right)\right)^\omega + \left(-\ln F_\kappa\left(x_\kappa\right)\right)^\omega\right]^{1/\omega}\right)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

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

In both steps we use maximum likelihood to estimate parameters. Table A.1 shows the estimated parameters, and Figure **??** plots the estimated marginal distributions together with the empirical distributions. For the joint distribution, we estimate $\omega = 1.30$. To put this into perspective, this estimate implies a Kendall's tau of $\tau = 1 - \frac{1}{1.30} = 0.23$. 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 $\tau = 0.21$, for α_i and κ_i we obtain $\tau = 0.29$ and for $-\beta_i$ and κ_i we obtain $\tau = 0.11$.

Appendix A3 Additional tables and figures

Figure A.1: Correlations between mean actions and beliefs



Note: Figure based on all 136 subjects.

Table A.1: Individual parameter estimates (all subjects)

Panel A: Marginal distributions	α_i	$-\beta_i$	κ_i
a	0.02	−0.11	0.14
b	0.34	0.49	0.16
Panel B: Joint distribution			
ω	1.30		

Notes: Table based on estimates from our core sample of 109 subjects.

Table A.2: Lottery choices

	Outcomes				
Lottery	A	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

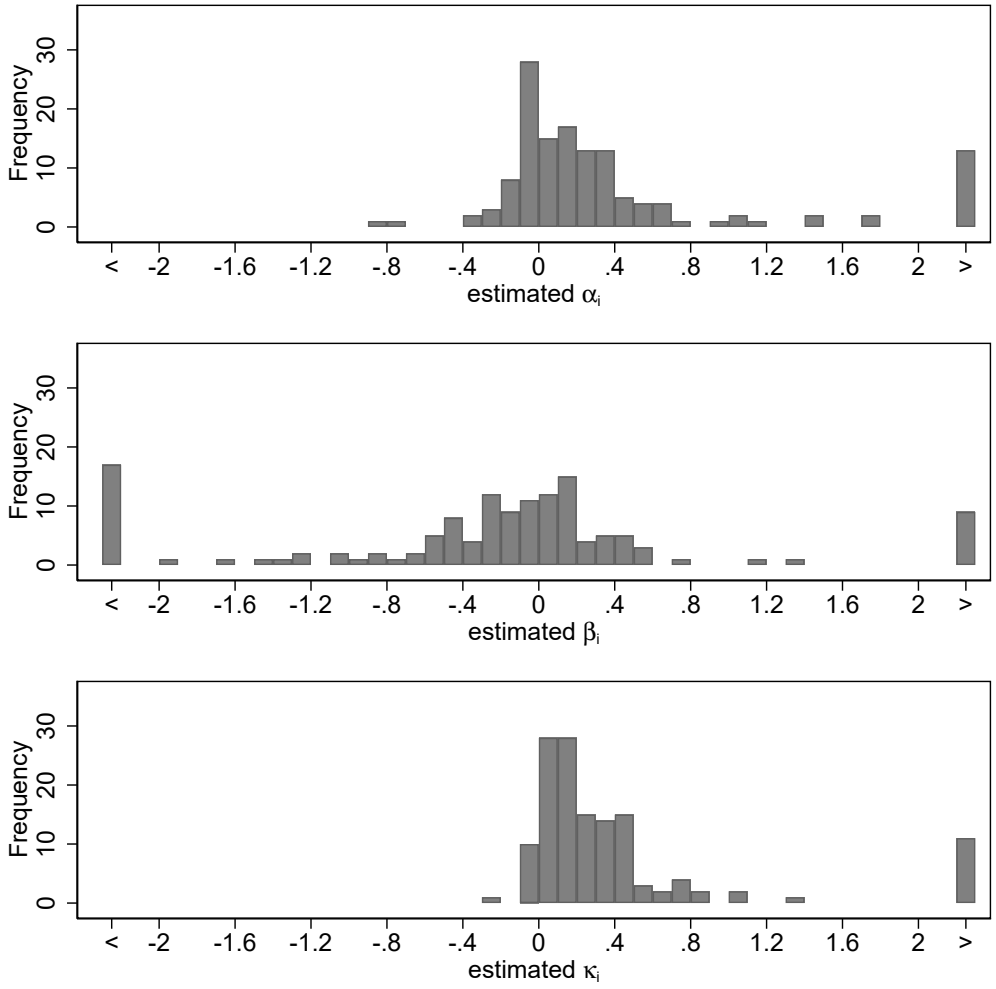
Notes: Lottery choices in the Eckel and Grossman (2002) risk elicitation task. 'Outcomes' are the payoffs denoted in "points", see Appendix A4 for the instructions. The final column lists the implied r_i parameters for each lottery choice. Note that after the first session, we slightly adjusted the outcomes to better estimate r_i. Table based on all 136 subjects.

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

Parameter	Median	Mean	S.D.	Min	Max
α_i	0.17	599.72	5938.54	−0.89	68186.74
β_i	−0.11	50.30	697.75	−496.78	8105.22
κ_i	0.20	189.62	2200.83	−0.29	25666.71

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	0.15	0.08	0.26	0.11	0.03	0.24
	(0.02)	(0.03)	(0.04)	(0.05)	(0.09)	(0.05)
β_k	0.00	0.11	−0.39	0.02	0.24	−0.45
	(0.02)	(0.03)	(0.11)	(0.05)	(0.07)	(0.11)
κ_k	0.21	0.24	0.17	0.19	0.35	0.16
	(0.01)	(0.02)	(0.02)	(0.02)	(0.10)	(0.03)
λ_k	0.26	0.29	0.17	0.31	0.22	0.16
	(0.01)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)
ϕ_k	1.00	0.60	0.40	0.48	0.17	0.36
	(-)	(0.05)	(0.05)	(0.06)	(0.05)	(0.04)
$\ln L$	-2898.0	-2638.3			-2587.6	
$EN(\boldsymbol{\tau})$	0.00	6.15			14.87	
ICL	5815.7	5326.9			5258.8	
NEC	-	0.024			0.048	

Notes: Standard errors in parentheses. Table based on all 136 subjects. For all types, we assume logarithmic utility ($r_k = 1$).

Table A.5: The 4-types model

	Type 1	Type 2	Type 3	Type 4
α_k	−0.02	0.47	0.27	0.13
	(0.07)	(0.12)	(0.04)	(0.11)
β_k	0.17	−1.10	−0.37	0.04
	(0.08)	(0.19)	(0.11)	(0.11)
κ_k	0.29	0.52	0.18	0.18
	(0.06)	(0.07)	(0.02)	(0.05)
λ_k	0.23	1.07	0.15	0.24
	(0.03)	(0.18)	(0.02)	(0.05)
ϕ_k	0.23	0.08	0.32	0.37
	(0.05)	(0.10)	(0.06)	(0.11)
$\ln L$		-2130.5		
$EN(\boldsymbol{\tau})$		23.95		
ICL		4374.1		
NEC		0.116		

Notes: Standard errors in parentheses. For all types, we impose log-utility (r_k = 1). Estimation results from models with 1, 2 and 3 types can be found in Table 3. Based on our 'core sample' of 109 subjects.

Table A.6: Strategies by type

	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%	4%	0%
C,C,D	21%	32%	5%	21%	43%	5%
C,D,C	0%	1%	0%	1%	0%	0%
C,D,D	9%	12%	4%	19%	4%	4%
D,C,C	2%	3%	0%	3%	2%	0%
D,C,D	7%	9%	3%	6%	13%	3%
D,D,C	2%	3%	2%	5%	1%	2%
D,D,D	57%	38%	87%	43%	24%	87%
Trust Games						
I,G	29%	43%	5%	31%	56%	5%
I,K	17%	23%	7%	38%	6%	7%
N,G	5%	5%	3%	2%	9%	3%
N,K	50%	28%	85%	30%	28%	84%
Ultimatum Games						
E,A	44%	57%	23%	52%	61%	22%
E,F	8%	12%	3%	9%	15%	3%
U,A	48%	32%	74%	39%	24%	74%
U,A	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 3. Subjects are assigned a type based on the type posterior probability $\tau_{i,k}$ (that subject i belongs to type k, see eq. (14)).

Table A.7: Transitions between types

Panel A: 2 types and 3 types (ln)

	2 types (ln)	
3 types (ln)	Type 1	Type 2
Type 1	36	0
Type 2	31	1
Type 3	0	41

Panel B: 2 types, ln and crra

	2 types (crra)	
2 types (ln)	Type 1	Type 2
Type 1	67	0
Type 2	1	41

Panel C: 3 types, ln and crra

	3 types (crra)		
3 types (ln)	Type 1	Type 2	Type 3
Type 1	31	4	1
Type 2	1	30	1
Type 3	0	0	41

Panel D: 2 types, ln and risk neutral

	2 types (risk neutral)	
2 types (ln)	Type 1	Type 2
Type 1	65	2
Type 2	5	37

Panel E: 3 types, ln and risk neutral

	3 types (risk neutral)		
3 types (ln)	Type 1	Type 2	Type 3
Type 1	31	3	2
Type 2	17	15	0
Type 3	5	0	36

Panel F: 2 types, subjective and rational expectations (ln)

	2 types (rational exp.)	
2 types (ln)	Type 1	Type 2
Type 1	57	10
Type 2	2	40

Panel G: 3 types, subjective and rational expectations (ln)

	3 types (rational exp.)		
3 types (ln)	Type 1	Type 2	Type 3
Type 1	32	3	1
Type 2	11	17	4
Type 3	4	0	37

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. (14)).

Table A.8: Estimates at the aggregate level, incl. CRRA parameter r_k

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.15	0.07	0.28	0.06	0.11	0.28
	(0.02)	(0.02)	(0.07)	(0.05)	(0.05)	(0.07)
β_k	0.03	0.13	−0.34	0.11	0.19	−0.37
	(0.03)	(0.03)	(0.17)	(0.12)	(0.07)	(0.14)
κ_k	0.19	0.21	0.19	0.18	0.26	0.20
	(0.01)	(0.02)	(0.03)	(0.03)	(0.03)	(0.03)
λ_k	0.37	0.44	0.12	2.74	0.10	0.11
	(0.11)	(0.20)	(0.06)	(0.44)	(0.07)	(0.07)
r_k	0.88	0.86	1.07	0.33	1.26	1.14
	(0.09)	(0.12)	(0.29)	(0.27)	(0.21)	(0.25)
ϕ_k	1.00	0.63	0.37	0.30	0.31	0.39
	(-)	(0.05)	(0.05)	(0.06)	(0.06)	(0.05)
$\ln L$	-2335.1	-2152.5			-2122.4	
$EN(\boldsymbol{\tau})$	0.00	4.17			14.80	
ICL	4693.6	4360.8			4339.3	
NEC	-	0.023			0.070	

Notes: Bootstrapped standard errors in parentheses. Table based on our 'core sample' of 109 subjects.

Table A.9: Estimates at the aggregate level (without morality)

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	0.03	−0.12	0.09	−0.16	−0.08	0.09
	(0.02)	(0.03)	(0.02)	(0.05)	(0.07)	(0.02)
β_k	0.15	0.27	−0.22	0.10	0.45	−0.21
	(0.03)	(0.04)	(0.07)	(0.05)	(0.05)	(0.08)
λ_k	0.26	0.28	0.15	0.30	0.23	0.15
	(0.02)	(0.02)	(0.01)	(0.05)	(0.03)	(0.01)
ϕ_k	1.00	0.64	0.36	0.36	0.29	0.36
	(-)	(0.05)	(0.05)	(0.07)	(0.06)	(0.05)
$\ln L$	-2425.2	-2244.4		-2217.1		
$EN(\boldsymbol{\tau})$	0.00	4.96		17.55		
ICL	4864.4	4526.6		4503.3		
NEC	-	0.027		0.084		

Notes: Standard errors in parentheses. Based on our 'core sample' of 109 subjects. For all types, we assume logarithmic utility ($r_k = 1$).

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

	1 type	2 types		3 types		
	Rep. agent	Type 1	Type 2	Type 1	Type 2	Type 3
α_k	−0.08	−0.19	0.12	−0.13	0.12	−0.24
	(0.02)	(0.03)	(0.04)	(0.08)	(0.04)	(0.04)
β_k	0.15	0.26	−0.23	0.45	−0.21	0.10
	(0.03)	(0.03)	(0.06)	(0.05)	(0.07)	(0.05)
δ_k	0.16	0.28	−0.07	0.22	−0.07	0.32
	(0.03)	(0.04)	(0.05)	(0.14)	(0.05)	(0.08)
λ_k	0.25	0.26	0.15	0.22	0.15	0.28
	(0.02)	(0.02)	(0.01)	(0.03)	(0.02)	(0.05)
ϕ_k	1.00	0.64	0.36	0.29	0.36	0.35
	(-)	(0.05)	(0.05)	(0.07)	(0.05)	(0.07)
$\ln L$	-2409.6	-2214.0			-2186.2	
$EN(\boldsymbol{\tau})$	0.00	3.84			16.24	
ICL	4838.0	4474.1			4454.3	
NEC	-	0.020			0.073	

Notes: Standard errors in parentheses. Based on our 'core sample' of 109 subjects. For all types, we assume logarithmic utility ($r_k = 1$).

Appendix A4 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.

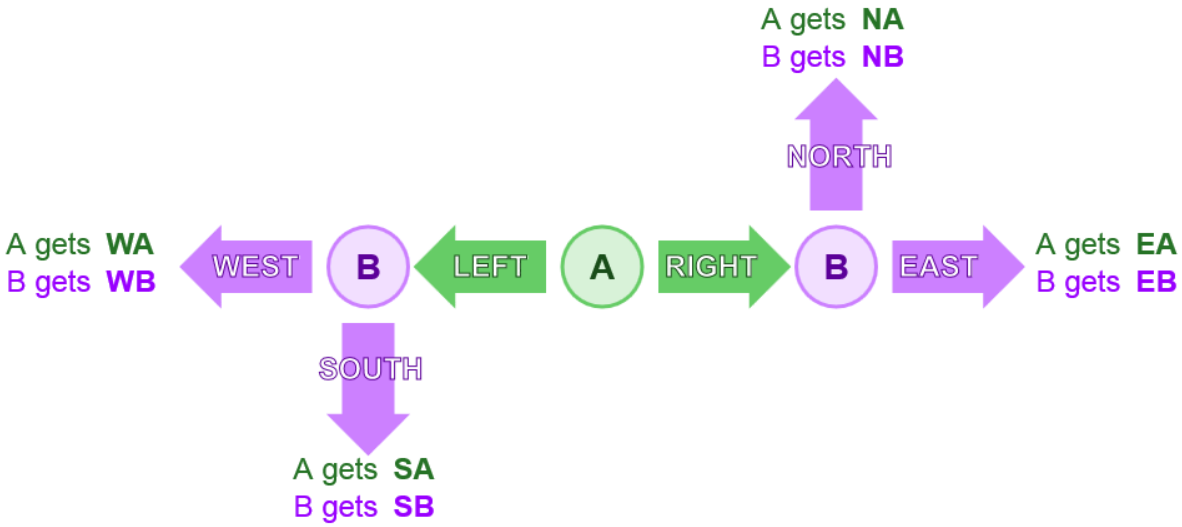
The choices of A and B jointly determine the number of points for A and B as follows:

- 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

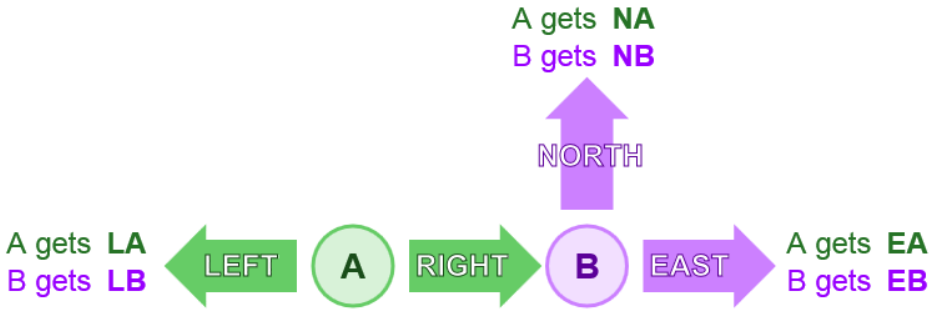
In the decision situation shown in Figure 2, person A first chooses LEFT or RIGHT. If A chooses LEFT, person B has no choice to make. If A chooses RIGHT, B has to choose between NORTH and EAST.

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



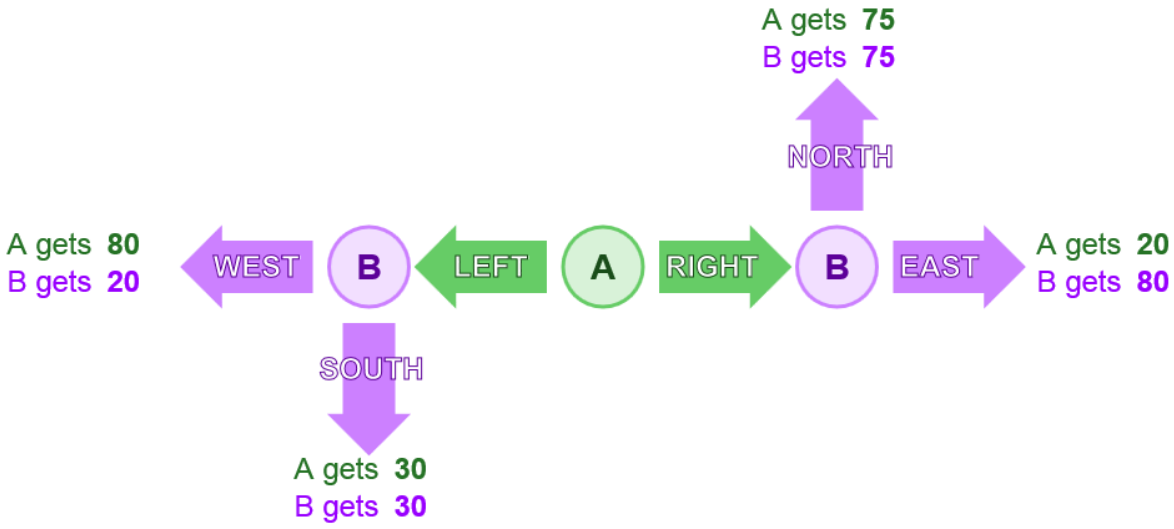
Example

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.

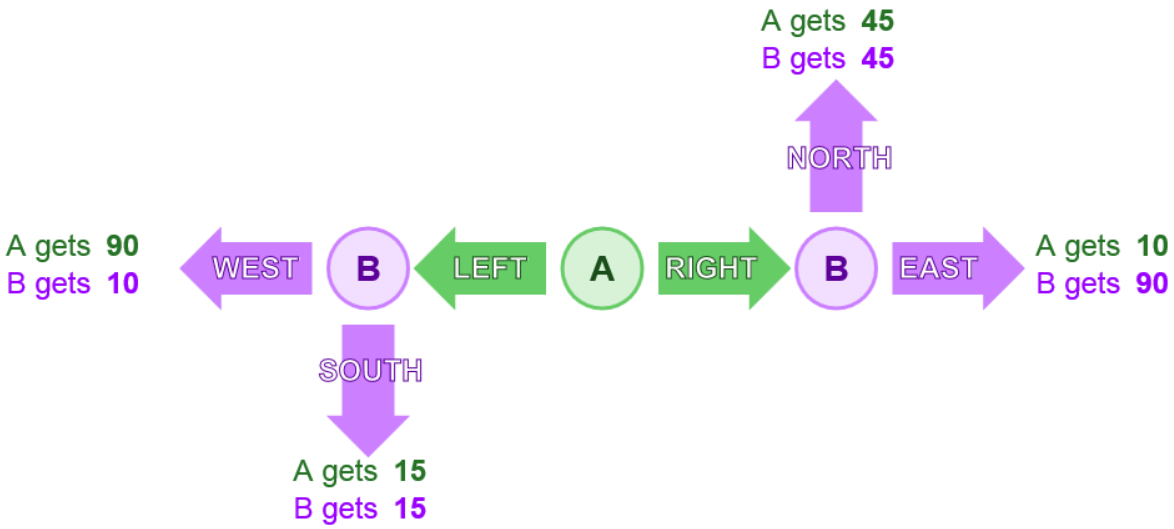
Decision screens

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.

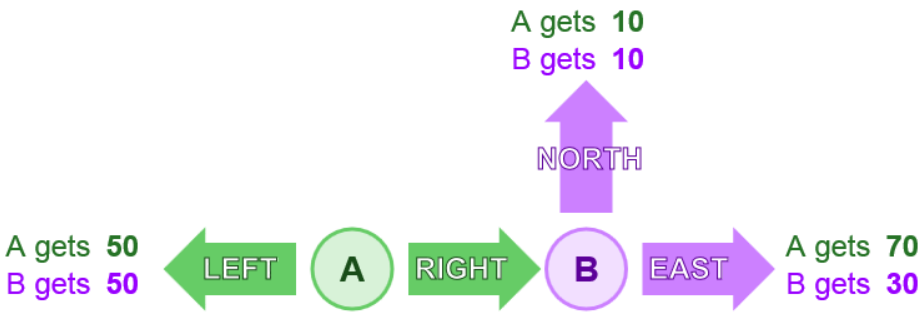
Suppose A chooses LEFT and B chooses SOUTH and EAST. How much would A and B earn?

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.