# Estimating Social Preferences and Kantian Morality in Strategic Interactions[*]

Boris van Leeuwen[†]

Ingela Alger[‡]

Jörgen W. Weibull[§]

30th November 2019

**Abstract:** Recent 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 laboratory experiments on strategic interaction in social dilemmas. Using a structural model, we estimate social preferences and morality concerns both at the individual level and the aggregate level. We observe considerable heterogeneity in social preferences and Kantian morality. A finite mixture analysis shows that the subject pool is well described as consisting of two types. One exhibits a combination of inequity aversion and Kantian morality, while the other combines spite and Kantian morality.

**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 Gijs van de Kuilen, Wieland Müller, Arthur Schram, and Sigrid Suetens for helpful suggestions and comments. Support by Knut and Alice Wallenberg Research Foundation is gratefully acknowledged. 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). I.A. and J.W. acknowledge IAST funding from the French National Research Agency (ANR) under grant ANR-17-EURE-0010 (Investissements d'Avenir program), and Chaire d'Excellence ANR-12-CHEX-0012-01 for I.A., and Chaire IDEX ANR-11-IDEX-0002-02 for J.W.

[†] Department of Economics and CentER, 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

[§] Stockholm School of Economics and Institute for Advanced Study in Toulouse. jorgen.weibull@hhs.se

# 1 Introduction

Social dilemmas, in which individuals must forego some personal benefit for a collective benefit to be generated, are ubiquitous. An investor's fear that an entrepreneur will "take the money and run" may prevent enterprises from being created. In bargaining situations, agreements may fail to materialize because one party was offended by the offer by the other party. Absent enforceable contracts between the parties, or the threat of future punishments such as in repeated games, the interacting parties' ability to realize potential mutual benefits depends to a large extent on the parties' attitudes. In particular, mutual benefits typically fail to materialize if each party is selfish, but may materialize if the parties have other-regarding preferences, such as altruism or inequity aversion. Understanding the nature of preferences is thus crucial for understanding behavior in social dilemmas.

While preferences such as 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), and image concerns (Bénabou & Tirole, 2006; Ellingsen & Johannesson, 2008) have received much attention in the behavioral and experimental economics literature, moral concerns have been less studied, although they may clearly matter in social dilemmas (Binmore, 1998). We here contribute to filling this gap by reporting structurally estimated preference parameters of a general utility function that nests several much studied preference classes, such as pure self-interest, altruism, spite (negative altruism), and inequity aversion, as well as a form of Kantian morality, so called *Homo moralis* preferences, which have been theoretically shown to have a strong evolutionary foundation (Alger & Weibull, 2013; Alger, Weibull, & Lehmann, 2019). The data was collected in a laboratory experiment using a variety of strategic interactions.

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. The subjects are randomly and anonymously matched. Each subject had to make multiple decisions in each of 18 different games (six games in each game class). In order to identify preferences in such strategic interactions, the analyst needs to either hypothesize subjects' beliefs about the behavior of their opponents (for example by some equilibrium hypothesis) or else try to elicit each subject's belief in each strategic interaction. We have chosen the second route. The ability to control for subjects' beliefs when trying to identify their preferences is indeed crucial for the estimation of other-regarding preferences (see Miettinen, Kosfeld, Fehr, and Weibull (2019) and references therein). Social image concerns (Bénabou & Tirole, 2006) are muted since subjects are anonymously and randomly matched.

On the basis of observed individual choices and reported beliefs, we estimate the preference parameter values for each individual subject. We do so using a structural model, the use of which has become more commonplace in experimental and behavioral economics, including the estimation of social preferences (DellaVigna, 2018). For this purpose, we add idiosyncratic random noise terms to the parametric utility function, thereby obtaining logistic choice probabilities, and apply the maximum likelihood method. We also perform aggregate estimations, using a finite mixture approach, the same as that used by Bruhin, Fehr, and Schunk (2018) in their statistical analysis of social preferences.

Not surprisingly, we find a lot of heterogeneity. The standard deviations of the estimated probability distributions for the parameters are in some cases as much as three times the mean. While many subjects show aversion to unfavorable inequity, some are either indifferent or positive or negative concerning aversion to favorable inequity. Most subjects show some concern for Kantian morality, and allowing for this motivational factor significantly improved the fit of the model to the data. The representative agent in the subject pool places about 70% weight on own payoff, about 13% weight on not falling behind the opponent, and 17% weight on Kantian morality. Our finite mixture estimations capture the heterogeneity in a tractable way. Models with two or three types provide a much better fit than the representative agent model. The two types model has one type that combines inequity aversion with Kantian morality, while the other type combines "spite" or "competitiveness" – a dislike of being behind and joy of being ahead – with Kantian morality. Importantly, allowing for Kantian morality substantially improves the fit of the model. The model selection criteria indeed favor models with Kantian morality over those without. Moreover, the value added of Kantian morality is in the same ballpark of such well-established motives as inequity aversion, altruism, and reciprocity.

Closest to our work are probably the papers by Bruhin et al. (2018) and Miettinen et al. (2019). The value added of the present study, in relation to Bruhin et al. (2018), is two-fold; we allow for the possibility of Kantian morality as part of the motivation behind choices, and we study strategic interactions (while they focus on allocation decisions for donors in dictator games and for second-movers in reciprocity games). In relation to Miettinen et al. (2019), the value added is four-fold: we study individual choices in 18 strategic interactions (while they study one), we elicit individual risk attitudes (while they assume risk neutrality), we make estimates of each subjects' vector of taste parameters, and we apply finite mixture methods.

More broadly speaking, our contribution fits in the large literature that estimates or tests models of social preferences.[1] In relation to this literature, our main contribution is that we estimate a deontological motive, namely a form of Kantian morality, in addition to social preferences.

The remainder of this paper is organized as follows. Section 2 introduces the class of preferences we estimate, Section 3 describes the experimental design, and Section 4 presents our econometric approach. The results are presented in Section 5, and Section 6 concludes.

[1] See, for example, Palfrey and Prisbrey (1997); Andreoni and Miller (2002); Charness and Rabin (2002); Engelmann and Strobel (2004); Fisman, Kariv, and Markovits (2007); Bellemare, Kröger, and Van Soest (2008); Blanco, Engelmann, and Normann (2011); DellaVigna, List, and Malmendier (2012); Ottoni-Wilhelm, Vesterlund, and Xie (2017) and, for a recent survey, see Cooper and Kagel (2015).

## 2 Social preferences and Kantian morality

We consider individual preferences such that each subject's expected utility can be written in the form

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

where $x$ is the behavior strategy used by the subject at hand, $i$, and $y$ the behavior strategy used by the subject $j$ with whom $i$ is matched. Each strategy pair $(x,y)$ determines the realization probability $\eta_{(x,y)}(\gamma)$ for each play $\gamma$ 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 1). 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 play $\gamma$. We will call $\pi_i(\gamma)$ the individual's *monetary utility* under play $\gamma$. If the monetary payoff allocation after a play $\gamma$ is $\left(m_i(\gamma),m_j(\gamma)\right)$, we assume that the individual's own monetary 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.[2] Hence, subject $i$ evaluates the opponent $j$'s

[2] There is experimental evidence that both students and financial professionals exhibit substantial 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).

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

For any given degree of relative risk aversion, the family of utility functions in (1) has three parameters. Two of them are the familiar measures of inequity aversion. The parameter $\alpha_i$ 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 monetary payoffs in the interaction. Likewise, the parameter $\beta_i$ 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 monetary utility.

The third parameter, $\kappa_i$, captures a Kantian moral concern (à la *Homo moralis*, Alger & Weibull, 2013). It places weight on the expected monetary 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 $\gamma$ would occur is $\eta_{(x,x)}(\gamma)$. In particular, a $\kappa_i$-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 monetary 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 other subjects. To choose a strategy $x$ in order to maximize the last term in (1) is to choose a strategy that maximizes monetary utility if used by both subjects in a pairwise interaction (see Alger & Weibull, 2013, for a discussion).

The utility function in equation (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 $\kappa_i = 0$. In that model, individuals care about own monetary utility and are also inequity averse. One obtains Becker's (1974) model of pure altruism by setting $\kappa_i = 0$ and $\alpha_i = -\beta_i$, for some $\beta_i \in (0,1)$.[3] Here $\beta_i$ is the individual's "degree of altruism", the weight placed on the other subject's monetary utility, while the weight $1 - \beta_i$ is placed on own monetary utility.

Pure *Homo moralis* preferences are obtained by setting $\alpha_i = \beta_i = 0$ and $\kappa_i \in (0,1)$. Here $\kappa_i$ is the individual's "degree of Kantian morality", the weight placed on the monetary utility that would be obtained if both subjects in the interaction at hand would use the same strategy, while the weight $1 - \kappa_i$ is placed on own monetary utility.

The utility function in (1) also nests the Charness and Rabin (2002) model without reciprocity. In Section 5.4 we extend the utility function to also accommodate reciprocity as formalized in Charness and Rabin (2002).

# 3 Experimental design and procedures

## 3.1 Game protocols

In the experiment, subjects play three types of well-known game protocols, illustrated in Figure 1: Sequential Prisoner's Dilemmas (SPD), shown in Figure 1a, mini Trust Game protocols (TG), shown in Figure 1b, and mini Ultimatum Game protocols (UG), shown in Figure 1c.[4] 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 $T > R > P > S$. Each subject plays 6 versions of each type of game protocol, for different monetary payoff assignments $T$, $R$, $P$ and $S$, see Table 1. [5]

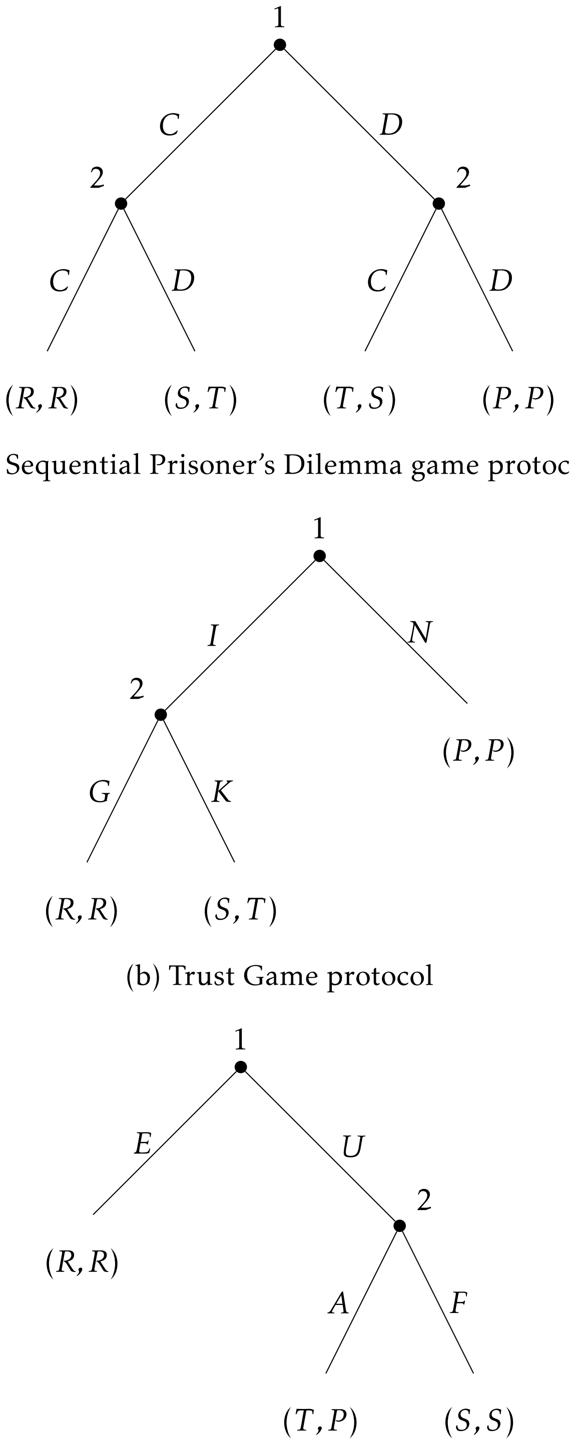
The term in the utility function in (1) that captures Kantian morality re-

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

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

[5] 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. More details are available upon request.

Figure 1: Game protocols



(a) Sequential Prisoner's Dilemma game protocol

(b) Trust Game protocol

(c) Ultimatum Game protocol

Table 1: Game protocols: monetary payoffs

| SPD protocols | | | | | TG protocols | | | | | UG protocols | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. | $T$ | $R$ | $P$ | $S$ | No. | $T$ | $R$ | $P$ | $S$ | No. | $T$ | $R$ | $P$ | $S$ |
| 1 | 90 | 45 | 15 | 10 | 7 | 80 | 50 | 30 | 20 | 13 | 60 | 50 | 40 | 10 |
| 2 | 90 | 55 | 20 | 10 | 8 | 90 | 50 | 30 | 10 | 14 | 65 | 50 | 35 | 10 |
| 3 | 80 | 65 | 25 | 20 | 9 | 80 | 60 | 30 | 20 | 15 | 70 | 50 | 30 | 10 |
| 4 | 90 | 65 | 25 | 10 | 10 | 90 | 60 | 30 | 10 | 16 | 75 | 50 | 25 | 10 |
| 5 | 80 | 75 | 30 | 20 | 11 | 80 | 70 | 30 | 20 | 17 | 80 | 50 | 20 | 10 |
| 6 | 90 | 75 | 30 | 10 | 12 | 90 | 70 | 30 | 10 | 18 | 85 | 50 | 15 | 10 |

*Notes:* All payoffs denoted in the experimental currency ("points"), see Section 3.3 for details on payments.

quires that a strategy can "meet itself", hence, that the interaction is symmetric. To symmetrize the game protocols in Figure 1, which are clearly 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. A behavior strategy $x$ therefore consists of specifying (potentially randomized) choices at *all* decision nodes in the game protocol. In the symmetrized SPD, each player thus has 8 pure strategies; in the first-mover role to choose between "cooperate" ($C$) and "defect" ($D$), and in the second-mover role to choose between $C$ and $D$, both if the first-mover played $C$ and if the first-mover played $D$.

In the two other game protocols, each player likewise has 4 pure strategies: a binary choice as first-mover, and a binary choice as second-mover. In the TGs, for the first-mover role each subject selects between investing ($I$) or not ($N$). Each subject also specifies whether (s)he would like to give back something ($G$) or keep everything ($K$) if the first-mover invested. In the UGs each subject specifies if (s)he would choose an equal ($E$) or an unequal ($U$) division in the first-mover role, and whether (s)he would accept ($A$) or refuse ($F$) an unequal proposal.

### 3.2 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. 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):

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

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:

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

Comparing the second terms in these utility functions one sees that if $i$ is an altruist, then $i$ cares about the other individual's monetary payoff $(mx+w-y)/2$ (which depends on the other's strategy $y$), while if $i$ is driven by Kantian morality (s)he does not care about the other's monetary payoff, but instead cares about the monetary payoff $(mx+w-x)/2$, which would result if both players were to use $i$'s strategy $x$.

However, this induces identical trade-offs for altruists and Kantian moralists, as shown by the derivatives with respect to own strategy $x$:

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

and

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

Hence, for generic parameter values, whether an altruist or a Kantian moralist, the individual either gives the whole endowment or nothing at all (this is due to the assumed risk neutrality). Moreover, 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 $\sigma_i$ exceeds $1/m$ while the Kantian moralist gives everything if $\kappa_i$ exceeds $1/m$.[6] Therefore, we would be unable to separate altruism from a Kantian concern using dictator games.[7]

By contrast, by 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. To see this, consider the Ultimatum Game protocol, as in Figure 1c. When symmetrically randomized, in this game a behavior strategy is a vector, $x = (x_1, x_2) \in [0,1]^2$, where $x_1$ is the probability with which the player proposes an equal sharing, and $x_2$ the probability with which he accepts an unequal sharing. Then a risk-neutral subject $i$ obtains the following expected utility from playing $x = (x_1, x_2)$ when he believes that the opponent will play $\hat{y} = (\hat{y}_1, \hat{y}_2)$ (the randomization factor $1/2$ has been

[6] 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).

[7] Many experiments use allocation tasks of the following sort. 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 $T > P > S$. It can be verified that 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 ($\alpha_i \cdot \kappa_i > 0$).

omitted):

$$\begin{aligned} u_i(x,\hat{y}) \;=\;& (1-\kappa_i)[x_1R+(1-x_1)\hat{y}_2T+(1-x_1)(1-\hat{y}_2)S \qquad (8)\\ &+\hat{y}_1R+(1-\hat{y}_1)x_2P+(1-\hat{y}_1)(1-x_2)S]\\ &-[\alpha_i(1-\hat{y}_1)x_2+\beta_i(1-x_1)\hat{y}_2](T-P)\\ &+\kappa_i[x_1R+(1-x_1)x_2T+(1-x_1)(1-x_2)S\\ &+x_1R+(1-x_1)x_2P+(1-x_1)(1-x_2)S]. \end{aligned}$$

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

$$\begin{aligned} \frac{\partial u_i(x,\hat{y})}{\partial x_1} \;=\;& (1-\kappa_i)[R-\hat{y}_2T-(1-\hat{y}_2)S]+\beta_i\cdot\hat{y}_2(T-P) \qquad (9)\\ &+\kappa_i\cdot[2(R-S)-x_2(T+P-2S)] \end{aligned}$$

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

These expressions reveal the key difference between 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 ($\alpha_i = \beta_i = 0$). 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 $\kappa_i \neq 0$). The expressions (9) and (10) further show that estimation of the preference parameters requires information of the subjects' beliefs about the opponent's play information that we elicit from the subjects.

While we provide an equally detailed analysis of the other two game protocols in Appendix A1, here we discuss the effect of Kantian morality in these protocols in the light of simple examples. First, consider a 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 $\kappa_i$, 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. In other words, the first-mover choice by 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 as second-mover at all nodes (information that we collect in the experiment). This example highlights an important consequence of Kantian morality: a subject's preferences over moves off the path of a strategy pair $(x, y)$ may influence his or her decisions on its path. This differs sharply from altruism, inequity aversion or spite, since such individual's first-mover choice depends only on her belief about her opponent's ensuing second-mover choice.

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.

### 3.3 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). Each of the game protocols had different monetary payoffs, which are listed in Table 1. 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.

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 A3. Sessions took around 1 hour and subjects earned between €10.50 and €26.90 with an average of €18.80. The experimental design and main analyses were pre-registered.[8]

# 4 Statistical analysis

The econometric strategy consists in producing both individual and aggregate estimates of the parameters in the utility function specified in (1). In

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

the statistical data analysis to follow we will add yet another term to the expression in (1), an idiosyncratic random variable that we take to be Gumbel distributed. Assuming that all such "noise" terms are statistically independent with the same mode, and with a subject-specific variance, this render all subject's choice probabilities on the familiar logit form (McFadden, 1974), each with a subject-specific noise parameter. When analyzing the choice data in terms of this random utility model, we will use the subjects' stated beliefs about other subjects' strategy choice. Hence, no equilibrium assumption is imposed.

## 4.1 Individual preferences

For each subject $i$, we estimate the individual's social and moral preference parameters $\alpha_i$, $\beta_i$, and $\kappa_i$ 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. As a robustness check, we also carry out the analysis under the alternative assumption that all subjects are risk neutral (all $r_i = 0$), see Section 5.3. 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 $\varepsilon_{ix_i}$ 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 $\varepsilon_{ix_i}$ 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):

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

where $\lambda_i > 0$ is a "noise" parameter, which is estimated alongside the preference parameters in $\theta_i$, and $X_g$ denotes the set of pure strategies in game protocol $g$. The smaller the parameter $\lambda_i$ 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 $\lambda_i$ for each individual $i$. Then, the probability density function can be written as:

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

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.

## 4.2 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. (2018). 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 $\theta_k = (\alpha_k, \beta_k, \kappa_k)$, the CRRA parameter $r_k$, and the noise parameter $\lambda_k$. The log-likelihood is given by:

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

where $\phi_k$ 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). As part of the EM algorithm, we estimate the posterior probabilities $\tau_{i,k}$ that individual $i$ belongs to type $k$ by:

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

# 5 Results

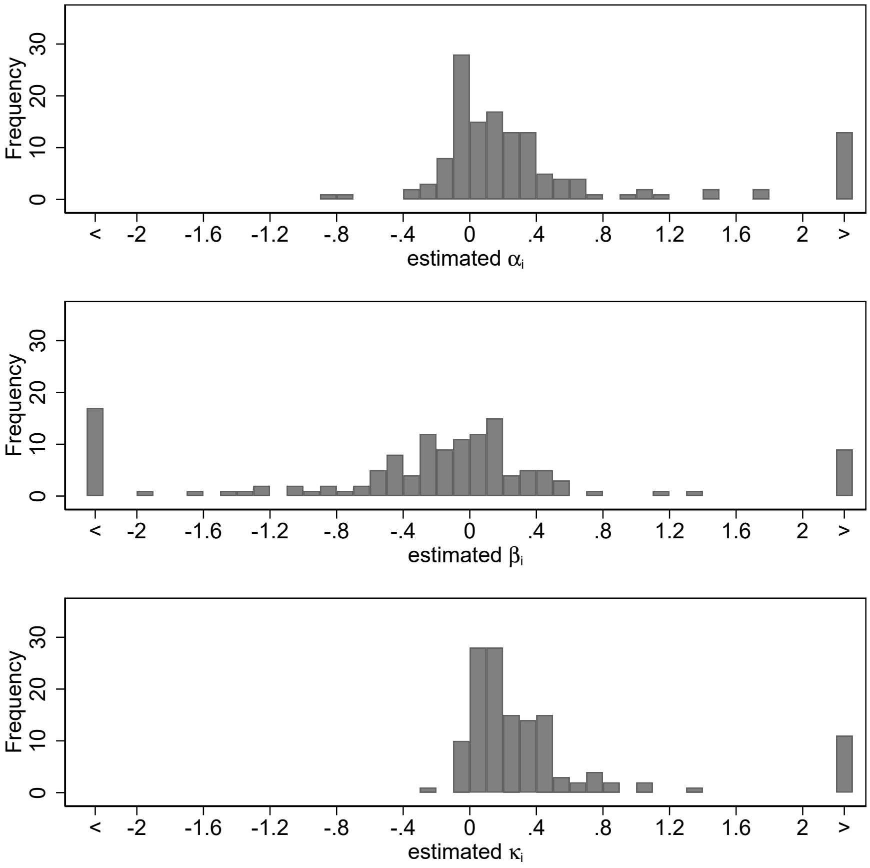
## 5.1 Individual preferences

Figure 2 shows the marginal distributions of the estimated individual preference parameters $\alpha_i$, $\beta_i$, and $\kappa_i$.[9] For all three parameters, we observe considerable heterogeneity. Most estimates of $\alpha_i$ and $\kappa_i$ are positive and signed-ranks tests confirm that the parameter distributions are located to the right of zero ($p < 0.001$ for both $\alpha_i$ and $\kappa_i$ estimates). By contrast, most estimates of $\beta_i$ 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 ($\alpha_i > 0, \beta_i < 0$).

The distributions in Figure 2 indicate that there is substantial heterogeneity in social preferences and Kantian morality concerns. For most subjects, the parameter estimates are of reasonable size. Indeed, mean and median estimates for the three parameters have absolute values below one half. However, for some subjects we obtain very large estimates of $\alpha_i$, $\beta_i$, and/or $\kappa_i$ (in absolute value). 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. In Appendix A2 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.

[9] In Table A.1 in Appendix A2, we present an overview of the actions and beliefs for each game protocol. Table A.2 in Appendix A2 presents all decisions in the risk elicitation task.

Figure 2: Distributions of individual parameter estimates



*Note:* All estimates of $\alpha_i$, $\beta_i$ and $\kappa_i$ 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 2: Individual parameter estimates

| Parameter | Median | Mean | S.D. | Min | Max |
|---|---|---|---|---|---|
| $\alpha_i$ | 0.14 | 0.19 | 0.38 | $-0.89$ | 1.75 |
| $\beta_i$ | -0.06 | $-0.14$ | 0.51 | $-1.97$ | 1.37 |
| $\kappa_i$ | 0.18 | 0.24 | 0.22 | $-0.10$ | 1.10 |

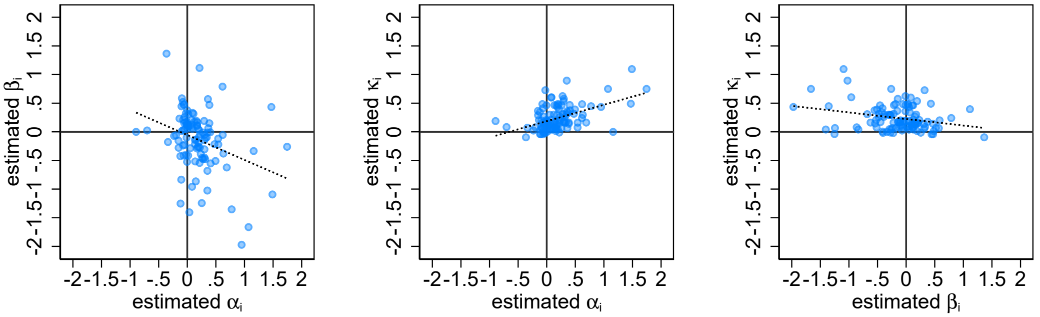
*Notes:* Table based on the 109 subjects for whom the $\alpha_i$, $\beta_i$ and $\kappa_i$ estimates have absolute value below 2. Table A.3 shows a similar table based on all 136 subjects.

Table 2, which shows summary statistics for the parameter estimates, provides further support for the pattern observed in Figure 2. Median and mean estimates are positive for $\alpha_i$ and $\kappa_i$, but negative for $\beta_i$. Moreover, the relatively large standard deviations indicate that there is considerable heterogeneity in social preferences and Kantian morality. [10]

Figure 3 illustrates the pairwise correlations between the three preference parameter estimates. The left panel of Figure 3 shows that the estimates for $\alpha_i$ and $\beta_i$ 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 $\beta_i < 0$, 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 3 reveals a strong and positive correlation between $\alpha_i$ and $\kappa_i$ 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. (2018) call it, "behindness aversion," with Kantian

[10] For these estimates we used the risk elicitation task to determine $r_i$. However, as a robustness test we also estimate $r_i$ alongside the preference parameters ($\alpha_i$, $\beta_i$, $\kappa_i$). Doing so 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 $\alpha_i$, $\rho = 0.566, p < 0.001$, $n = 109$ for $\beta_i$, and $\rho = 0.606, p < 0.001$ for $\kappa_i$) although the correlation between the imposed and estimated $r_i$ values is weak ($\rho = 0.069, p = 0.478$). The estimates of $\alpha_i$, $\beta_i$ and $\kappa_i$ are not systematically smaller or larger using either method (signed-rank tests, $p = 0.198$, $n = 109$ for $\alpha_i$, $p = 0.228$, $n = 109$ for $\beta_i$, and $p = 0.388$, $n = 109$ for $\kappa_i$).

Figure 3: 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 $\alpha_i$, $\beta_i$ and $\kappa_i$ estimates have absolute value below 2.

morality. For the estimates of $\beta_i$ and $\kappa_i$ we find a negative correlation (Spearman's $\rho = -0.173$, $p = 0.071$, $n = 109$).

## 5.2 Aggregate estimations

Table 3 presents the estimates of the finite mixture models for one, two and three types. In each of the models, we assume that there is a fixed number of "preference types" in the population and we estimate parameters for each type, where individuals are endogenously assigned to one of the types (see section 4.2 for details). To distinguish these estimates from the individual ones, we use an index $k$ to designate the type.

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 |
| $\alpha_k$ | 0.15 | 0.07 | 0.28 | 0.06 | 0.11 | 0.28 |
| | (0.02) | (0.02) | (0.07) | (0.05) | (0.05) | (0.07) |
| $\beta_k$ | 0.03 | 0.13 | $-0.34$ | 0.11 | 0.19 | $-0.37$ |
| | (0.03) | (0.03) | (0.17) | (0.12) | (0.07) | (0.14) |
| $\kappa_k$ | 0.19 | 0.21 | 0.19 | 0.18 | 0.26 | 0.20 |
| | (0.01) | (0.02) | (0.03) | (0.03) | (0.03) | (0.03) |
| $\lambda_k$ | 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) |
| $\phi_k$ | 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:* Standard errors in parentheses. Table A.4 in Appendix A2 shows the estimates of a 4-type model. Table based on our ‘core sample’ of 109 subjects. Table A.5 in Appendix A2 shows estimates based on the full sample.

When assuming only one type, that is, a representative agent, we obtain the estimates $\alpha_0 = 0.15$, $\beta_0 = 0.03$, and $\kappa_0 = 0.19$, where the index 0 stands for the representative agent. In other words, the representative agent dislikes disadvantageous inequity, is virtually indifferent with respect to advantageous inequality, and has a positive degree of Kantian morality. Moreover, the representative agent's degree of risk aversion is estimated to be $r_0 = 0.88$. We note that if we approximate $r_0$ to be 1, then the monetary utility becomes logarithmic, which permits the expected utility in (1) for an individual $i$ to be written in the following form:

$$u_i(x,y) \ = \ \mathbb{E}_{\eta(x,y)}[v_i(\gamma)] + \ \kappa_0 \mathbb{E}_{\eta(x,x)}[\ln m_i(\gamma)], \tag{15}$$

where

$$v_i(\gamma) = (1-\kappa_0)\ln m_i(\gamma) + \alpha_0 \ln\left(\min\left\{1, \frac{m_i(\gamma)}{m_j(\gamma)}\right\}\right) + \beta_0 \ln\left(\min\left\{1, \frac{m_j(\gamma)}{m_i(\gamma)}\right\}\right). \tag{16}$$

The Bernoulli function value $v_i(\gamma)$ has the form of a Cobb-Douglas function. It attaches weight to each of three "goods"; own monetary payoff, the ratio between own monetary payoff and that of the opponent, when behind, and between the opponent's and own monetary payoff when ahead. According to our estimates, it is thus as if the representative agent places roughly 70% weight on the utility of own payoff, about 13% weight on behindness aversion, no weight on being ahead, and roughly 17% weight on the Kantian morality.[11] These individuals thus exhibit Kantian morality and behindness aversion.

As can be seen in Table 3, in the two multi-type models all types exhibit, like the representative agent, both behindness aversion ($\alpha_k > 0$) and Kantian morality ($\kappa_k > 0$), the latter being of the same order of magnitude as the representative agent. Unlike the representative agent, however, none of the types in these models is indifferent to the other's monetary utility when ahead: while some types dislike being ahead, other types like it.

More specifically, when assuming two types, the most common type (Type 1) exhibits inequity aversion, with parameter estimates $\alpha_1 = 0.07$ and $\beta_1 =$

[11] More precisely, normalizing to unity the sum of the coefficients in front of the four logarithmic terms in (15) and (16), one obtains $(1-\kappa_0)/((1-\kappa_0)+\alpha_0+\beta_0+\kappa_0) = 0.70$ etc.

0.13, and a degree of Kantian morality $\kappa_1 = 0.21$. This type's risk aversion is close to that of the representative agent ($r_1 = 0.86$), but the behindness aversion is weaker, and, unlike the representative agent, this type also dislikes advantageous inequality. Hence, this type combines inequity aversion with Kantian morality. This type represents about 63% of the subjects. The other type, Type 2, exhibits a combination of strong *spite* ("negative altruism") and Kantian morality, with $\alpha_2 = 0.28$, $\beta_2 = -0.34$, and $\kappa_2 = 0.19$. With $r_2 = 1.07$, this type is similar in terms of risk aversion to the other type. While social preferences and Kantian morality thus play major roles for both types, their main concern is their own monetary utility. In this sense, pure self-interest is still the main driver. The finding that a sizeable share of the subjects (here 37%) 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., 2019).

When assuming three types, for all types we again estimate a positive Kantian morality parameter $\kappa_k$. 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 (39%). The new Type 2 (31%) is close to the previous Type 1, combining inequity aversion with (relatively strong) Kantian morality. It represents around 31% of the population. Type 1 is very close to *Homo moralis*. It combines slight inequity aversion ($\alpha_1$ and $\beta_1$ are not significantly different from zero at the 5% level), with Kantian morality. Note that Type 1 is also less risk averse than the other types. 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. Again, pure self-interest remains the main motive behind choices. Note that we do not observe types who are best described by pure self-interest. This is in line with the findings by Bruhin et al. (2018). Importantly, we observe relatively little heterogeneity in estimates of the morality parameter $\kappa_k$. In all cases, $\kappa_k$ is around 0.2, showing that most people are well described by having Kantian morality concerns.

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 (McLachlan et al., 2019). 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

$$ICL = -2\ln L + d\ln N + EN(\boldsymbol{\tau}), \tag{17}$$

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 on our sample. The last term in (17) is the entropy

$$EN(\boldsymbol{\tau}) = -\sum_{k=1}^{K}\sum_{i=1}^{N}\tau_{i,k}\ln\tau_{i,k}, \tag{18}$$

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.

Bruhin et al. (2018) use the 'normalized entropy criterion' (NEC, Celeux & Soromenho, 1996), which is defined as:

$$NEC = \frac{EN(\boldsymbol{\tau})}{\ln L(K) - \ln L(1)}, \tag{19}$$

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 2-types model and the ICL selects the 3-types model. Table A.4 in Appendix A2 shows estimates and goodness-of-fit metrics for a 4-types model. The 4-types model has a higher NEC and a (slightly) lower ICL than the 2-types and 3-types models in 3. Note that marginal improvement in the ICL score is largest when going from the representative agent to the 2-types model. So, assuming two types instead of a representative agent brings us a long way in capturing the heterogeneity in the population.

## 5.3 Risk neutrality

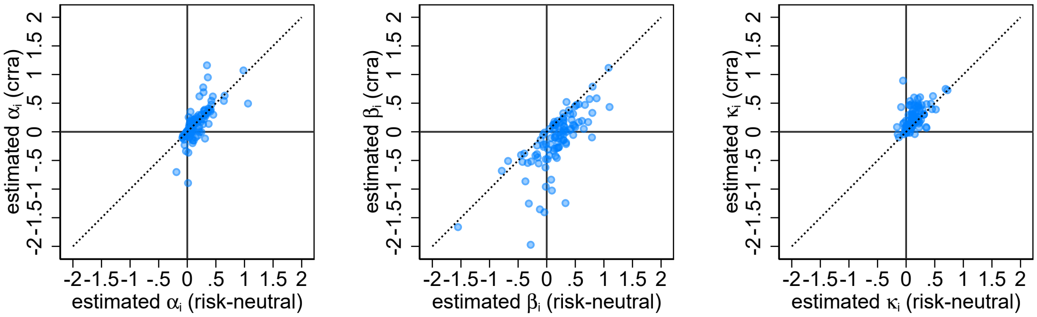
In the main analysis, we estimated the parameters in the expected utility function (1) with monetary utilities of the CRRA form as in (2) and (3), for pre-estimated degrees of risk aversion. As a robustness check, we also estimate the parameters in (1) under the alternative assumption that all subjects are risk neutral (i.e., $r_i = 0$ for all subjects $i$).

Figure 4 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 $\alpha_i$ (when behind) the Spearman rank correlation is $\rho = 0.802$. For the inequity parameter $\beta_i$ (when ahead) it is $\rho = 0.774$, and for the Kantian morality parameter $\kappa_i$ it is $\rho = 0.627$ (all three rank correlations hold for $p < 0.001$, $n = 109$).

The middle panel in Figure 4 also shows that the $\beta_i$ estimates are much higher under risk neutrality than under CRRA. Indeed, for 94 out of 109 subjects, the risk-neutral estimate is lower than the CRRA estimate (signed-rank test, $p < 0.001$).[12] By contrast, the risk-neutral estimates of $\kappa_i$ (80 out of 109, signed-rank test: $p < 0.001$) and $\alpha_i$ (64 out of 109, signed-rank test: $p = 0.068$)

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

Figure 4: 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.

are lower for most subjects than under CRRA.[13] 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 the estimates of the parameters $\alpha_k$ and $\kappa_k$ are not much affected. For all types in Tables 3 and 4, $\alpha_k$ and $\kappa_k$ are positive, under both risk hypotheses, with the Kantian morality parameter values somewhat lower under risk neutrality than under CRRA. In line with the individual parameter estimates, the finite mixture estimates of the parameters $\beta$ tend to be much higher under risk neutrality than under CRRA. Moreover, under risk-neutrality, all estimates of $\beta_k$ are non-negative, in contrast to the CRRA estimates, where we observed $\beta_k < 0$ for some types $k$.[14]

[13] Most risk-neutral estimates of $\kappa_i$ (96 out of 109) and $\alpha_i$ (92 out of 109) are positive (signed-rank tests, $p < 0.001$)

[14] One can easily see how assuming risk neutrality would bias estimates of $\beta_k$. Take for example the UG protocol. Both risk aversion and 'aheadness aversion' ($\beta_i > 0$) would induce one to choose $E$ over $U$.

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 |
| $\alpha_k$ | 0.17 | 0.19 | 0.13 | 0.18 | 0.19 | 0.01 |
| | (0.02) | (0.02) | (0.02) | (0.03) | (0.02) | (0.04) |
| $\beta_k$ | 0.25 | 0.00 | 0.36 | 0.26 | 0.00 | 0.50 |
| | (0.03) | (0.04) | (0.05) | (0.06) | (0.03) | (0.07) |
| $\kappa_k$ | 0.10 | 0.11 | 0.11 | 0.11 | 0.10 | 0.14 |
| | (0.01) | (0.01) | (0.02) | (0.02) | (0.01) | (0.05) |
| $\lambda_k$ | 7.62 | 4.01 | 8.98 | 9.29 | 3.79 | 6.92 |
| | (0.60) | (0.51) | (0.95) | (1.17) | (0.36) | (0.79) |
| $\phi_k$ | 1.00 | 0.36 | 0.64 | 0.48 | 0.34 | 0.18 |
| | (-) | (0.07) | (0.07) | (0.06) | (0.05) | (0.06) |
| $\ln L$ | -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:* Standard errors in parentheses. Table based on our ‘core sample’ of 109 subjects.

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 a lower ICL score than the risk-neutral model. For the 3-types model, for example, the ICL score under the CRRA assumption is quite a bit lower than under risk neutrality (4339.3 versus 4515.6), showing that the CRRA model considerably improves the fit over the risk-neutrality model.

## 5.4 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.

### 5.4.1 Individual estimations

We conduct likelihood-ratio tests to see if adding the Kantian morality parameter $\kappa_i$ to a model with only the two social preference parameters $\alpha_i$ and $\beta_i$ improves the fit. The likelihood-ratio tests reveal that adding $\kappa_i$ 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 $\alpha_i$ to $(\beta_i, \kappa_i)$, or $\beta_i$ to $(\alpha_i, \kappa_i)$, improves the fit at the 5% level for 20 and 26 individuals, respectively (at the 10% level, for 25 $(\alpha_i)$ and 37 $(\beta_i)$ 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:

$$BIC = -2\ln(L) + d\ln(18), \tag{20}$$

and

$$AIC = -2\ln(L) + 2d, \tag{21}$$

where $\ln(18)$ in (20) comes from the 18 observations per subject. Since $\ln 18 \approx 2.89 > 2$, BIC gives a heavier penalty per parameter than AIC.

Table 5 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 $\kappa_i$ in combination with $\alpha_i$ and/or $\beta_i$ have the lowest BIC scores. In sum, for 34 subjects (31.2%), the model with the lowest BIC score includes $\kappa_i$. In comparison, $\alpha_i$ and $\beta_i$ are included in the model with the lowest BIC score for 23 subjects (21.1%) and 35 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 $\kappa_i$ for 48 subjects (or 44.0%). Again, a larger number of subjects than for $\alpha_i$ (31 subjects, or 28.4%) and also slightly more subjects than $\beta_i$ (40 subjects, or 36.7%).

### 5.4.2 Aggregate estimations

We also evaluate the value added of Kantian morality for the finite mixture estimations. Table A.6 in Appendix A2 shows estimates for finite mixture models with only $\alpha_k$ and $\beta_k$ (i.e. where $\kappa_i = 0$). For any given number of types, these fixed mixture estimates give higher ICL scores than the model including Kantian morality, indicating that fixed mixture estimates that include the parameter $\kappa_i$ provide a better fit.

### 5.4.3 Reciprocity vs. Kantian morality

We finally compare the value added of the Kantian morality parameter, to the value if one were to instead of Kantian morality add reciprocity. For this

Table 5: Best individual fit

| | BIC | | AIC | |
|---|---|---|---|---|
| Parameters | Frequency | Percentage | Frequency | Percentage |
| $\alpha_i, \beta_i, \gamma_i$ | 2 | 1.8 | 6 | 5.5 |
| $\alpha_i, \beta_i$ | 7 | 6.4 | 6 | 5.5 |
| $\alpha_i, \gamma_i$ | 5 | 4.6 | 8 | 7.3 |
| $\beta_i, \gamma_i$ | 4 | 3.7 | 10 | 9.2 |
| $\alpha_i$ | 9 | 8.3 | 11 | 10.1 |
| $\beta_i$ | 22 | 20.2 | 18 | 16.5 |
| $\gamma_i$ | 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.

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

$$
\begin{aligned}
u_i(x,y) \quad = \quad & \sum_{\gamma} \eta_{(x,y)}(\gamma) \cdot \pi_i(\gamma) \\
& - \alpha_i \cdot \sum_{\gamma} \eta_{(x,y)}(\gamma) \cdot \max\left\{0, \pi_{ij}(\gamma) - \pi_i(\gamma)\right\} \\
& - \beta_i \cdot \sum_{\gamma} \eta_{(x,y)}(\gamma) \cdot \max\left\{0, \pi_i(\gamma) - \pi_{ij}(\gamma)\right\} \\
& - \delta_i \cdot q \cdot \sum_{\gamma} \eta_{(x,y)}(\gamma) \cdot \max\left\{0, \pi_{ij}(\gamma) - \pi_i(\gamma)\right\},
\end{aligned}
\tag{22}
$$

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 $\delta_i$ 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.7 in Appendix A2 we provide the results of finite mixture models based on (22). The 3-types model has the lowest ICL score among the reciprocity models. Based on the ICL score, the 3-types reciprocity model performs better than the mixture models with only $\alpha_k$ and $\beta_k$ (see A.6). This shows that, adding reciprocity improves the fit of the model. Importantly however, the 3-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.

# 6 Concluding discussion

In this paper, we report results from a laboratory experiment designed to evaluate the explanatory power of Kantian morality in standard strategic interactions. To distinguish Kantian morality from other social concerns, we posit a general utility function that nests several much studied preference classes, such as pure self-interest, altruism, spite, 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 2-types model suggests that a bit more than two thirds of the subjects display a combination of inequity aversion with Kantian morality, and the remaining third a combination of Kantian morality and behindness aversion (and indifference towards advantageous inequity). Within the 3-types model, the inequity averse and Kantian moral type still represents a little less than one third of the subjects. However, now there is another type which displays only Kantian morality, while the remaining 40% of the subjects appear to combine spite with Kantian morality. Quite remarkably, all the preference types—both the representative agent and the preference types within the 2-types and the 3-types model—have an estimated Kantian morality parameter $\kappa_k$ of around 0.2.

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. (2019) show that preferences that combine Kantian morality with either altruism or spite are favored by evolution by natural selection.[15]

Compared with other experimental studies with structural preference estimations, our results agree with those of Bruhin et al. (2018) 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. (2018) 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

[15] This result does not contradict that of Alger and Weibull (2013), which is shown by Alger et al. (2019) 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. (2019) 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).

by Miettinen et al. (2019).

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 intergroup conflict are expected to impact preferences guiding behavior in strategic interactions (Alger et al., 2019; 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).

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. (2019). Evolution of preferences in structured populations: genes, guns, and culture. *Journal of Economic Theory*, forthcoming.

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.

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., & 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.

Binmore, K. (1998). *Just playing: Game theory and the social contract vol. 2*. Cambridge, MA: MIT Press.

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

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

Bruhin, A., Fehr, E., & Schunk, D. (2018). The many faces of human sociality: Uncovering the distribution and stability of social preferences. *Journal of the European Economic Association*, *17*(4), 1025–1069.

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

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.

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. (2019). *Revealed preferences in a sequential prisoners' dilemma: a horse-race between six utility functions* (revised version of CESifo Working Paper Series No. 6358, 2017).

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*, 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 3.2), 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).[16] 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):

$$\begin{aligned} u_i(x,y) \;=\; & (1-\kappa_i)[x_1[y_2R+(1-y_2)S]+(1-x_1)P] \\ & +(1-\kappa_i)[y_1[x_2R+(1-x_2)T]+(1-y_1)P] \\ & +\kappa_i\{x_1[x_2R+(1-x_2)S]+(1-x_1)P\} \\ & +\kappa_i\{x_1[x_2R+(1-x_2)T]+(1-x_1)P\} \\ & -[\alpha_i x_1(1-y_2)+\beta_i y_1(1-x_2)](T-S). \end{aligned} \tag{23}$$

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

$$\frac{\partial u_i(x,\hat{y})}{\partial x_1} = (1-\kappa_i)[S-P+\hat{y}_2(R-S)]+\kappa_i[x_2(2R-S-T)+S+T]-\alpha_i(1-\hat{y}_2)(T-S), \tag{24}$$

and

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

The social preference parameters $\alpha_i$ and $\beta_i$ represent consequentialistic motives: they give weight to the monetary payoff consequences given what the

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

subject believes about the opponent's actual play. By contrast, the Kantian morality parameter $\kappa_i$ 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.

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):

$$\begin{aligned} u_i(x,y) \quad = \quad & (1-\kappa_i)[x_1 y_2 R + x_1(1-y_2)S + (1-x_1)y_3 T + (1-x_1)(1-y_3)P] \\ & + (1-\kappa_i)[y_1 x_2 R + y_1(1-x_2)T + (1-y_1)x_3 S + (1-y_1)(1-x_3)P] \\ & + \kappa_i [x_1 x_2 R + x_1(1-x_2)S + (1-x_1)x_3 T + (1-x_1)(1-x_3)P] \\ & + \kappa_i [x_1 x_2 R + x_1(1-x_2)T + (1-x_1)x_3 S + (1-x_1)(1-x_3)P] \\ & - \alpha_i [x_1(1-y_2) + (1-y_1)x_3](T-S) \\ & - \beta_i [(1-x_1)y_3 + y_1(1-x_2)](T-S). \end{aligned} \tag{26}$$

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

$$\begin{aligned} \frac{\partial u_i(x,\hat{y})}{\partial x_1} \quad = \quad & (1-\kappa_i)[S-P+\hat{y}_2(R-S)-\hat{y}_3(T-P)] \\ & + \kappa_i[x_2(2R-S-T)+(1-x_3)(S+T-2P)] \\ & + \beta_i \hat{y}_3(T-S) - \alpha_i(1-\hat{y}_2)(T-S), \end{aligned} \tag{27}$$

$$\frac{\partial u_i(x,\hat{y})}{\partial x_2} = (1-\kappa_i)\hat{y}_1(R-T) + \kappa_i x_1(2R-S-T) + \beta_i \hat{y}_1(T-S), \tag{28}$$

and

$$\frac{\partial u_i(x,\hat{y})}{\partial x_3} = (1-\kappa_i)(1-\hat{y}_1)(S-P) + \kappa_i(1-x_1)(T+S-2P) - \alpha_i(1-\hat{y}_1)(T-S). \tag{29}$$

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 Additional tables

Table A.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 |
| 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 | . |
| 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 | . |

*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.

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 A3 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 |
|---|---|---|---|---|---|
| $\alpha_i$ | 0.17 | 599.72 | 5938.54 | −0.89 | 68 186.74 |
| $\beta_i$ | −0.11 | 50.30 | 697.75 | −496.78 | 8105.22 |
| $\kappa_i$ | 0.20 | 189.62 | 2200.83 | −0.29 | 25 666.71 |

*Notes:* Table based on estimates from all 136 subjects.

Table A.4: The 4-types model

| | Type 1 | Type 2 | Type 3 | Type 4 |
|---|---|---|---|---|
| $\alpha_k$ | 0.04 | 0.15 | 0.25 | 0.09 |
| | (0.08) | (0.19) | (0.05) | (0.02) |
| $\beta_k$ | 0.30 | $-0.15$ | $-0.28$ | 0.19 |
| | (0.07) | (0.24) | (0.14) | (0.05) |
| $\kappa_k$ | 0.25 | 0.26 | 0.17 | 0.15 |
| | (0.07) | (0.05) | (0.03) | (0.02) |
| $\lambda_k$ | 0.45 | 0.01 | 0.19 | 12.67 |
| | (0.10) | (0.01) | (0.09) | (0.02) |
| $\rho_k$ | 0.79 | 2.11 | 0.89 | $-0.13$ |
| | (0.11) | (0.36) | (0.21) | (0.06) |
| $\phi_k$ | 0.19 | 0.23 | 0.33 | 0.24 |
| | (0.06) | (0.05) | (0.05) | (0.06) |
| $\ln L$ | | -2099.9 | | |
| $EN(\boldsymbol{\tau})$ | | 20.64 | | |
| ICL | | 4328.3 | | |
| NEC | | 0.088 | | |

*Notes:* Standard errors in parentheses. 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.5: 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 |
| $\alpha_k$ | 0.15 | 0.08 | 0.31 | 0.08 | 0.06 | 0.21 |
| | (0.02) | (0.02) | (0.12) | (0.14) | (0.04) | (0.12) |
| $\beta_k$ | 0.01 | 0.12 | $-0.63$ | $-0.17$ | 0.29 | $-0.31$ |
| | (0.03) | (0.03) | (0.40) | (0.24) | (0.06) | (0.36) |
| $\kappa_k$ | 0.20 | 0.23 | 0.19 | 0.20 | 0.26 | 0.15 |
| | (0.01) | (0.02) | (0.05) | (0.03) | (0.05) | (0.06) |
| $\lambda_k$ | 0.31 | 0.35 | 0.06 | 0.04 | 1.20 | 0.28 |
| | (0.10) | (0.13) | (0.08) | (0.21) | (0.38) | (0.11) |
| $r_k$ | 0.94 | 0.94 | 1.39 | 1.68 | 0.51 | 0.77 |
| | (0.10) | (0.11) | (0.36) | (0.75) | (0.15) | (0.39) |
| $\phi_k$ | 1.00 | 0.60 | 0.40 | 0.42 | 0.25 | 0.33 |
| | (-) | (0.04) | (0.04) | (0.08) | (0.07) | (0.04) |
| $\ln L$ | -2897.5 | -2633.9 | | -2583.5 | | |
| $EN(\boldsymbol{\tau})$ | 0.00 | 5.17 | | 16.65 | | |
| ICL | 5819.6 | 5327.1 | | 5267.2 | | |
| NEC | - | 0.020 | | 0.053 | | |

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

Table A.6: 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 |
| $\alpha_k$ | 0.00 | −0.05 | 0.11 | −0.07 | 0.12 | −0.01 |
| | (0.01) | (0.02) | (0.06) | (0.07) | (0.06) | (0.05) |
| $\beta_k$ | 0.22 | 0.35 | −0.44 | 0.19 | −0.43 | 0.52 |
| | (0.04) | (0.03) | (0.16) | (0.22) | (0.20) | (0.04) |
| $\lambda_k$ | 0.76 | 1.46 | 0.04 | 1.54 | 0.05 | 1.56 |
| | (0.24) | (0.57) | (0.09) | (0.90) | (0.09) | (0.70) |
| $r_k$ | 0.66 | 0.48 | 1.45 | 0.48 | 1.41 | 0.42 |
| | (0.11) | (0.10) | (0.33) | (0.73) | (0.40) | (0.15) |
| $\phi_k$ | 1.00 | 0.63 | 0.37 | 0.35 | 0.36 | 0.29 |
| | (-) | (0.05) | (0.05) | (0.06) | (0.06) | (0.06) |
| $\ln L$ | -2414.0 | -2229.1 | | -2198.7 | | |
| $EN(\boldsymbol{\tau})$ | 0.00 | 4.41 | | 16.40 | | |
| ICL | 4846.8 | 4504.9 | | 4479.4 | | |
| NEC | - | 0.024 | | 0.074 | | |

*Notes:* Standard errors in parentheses. Based on our ‘core sample’ of 109 subjects.

Table A.7: Estimates at the aggregate level (reciprocity)

| | 1 type | 2 types | | 3 types | | |
|---|---|---|---|---|---|---|
| | Rep. agent | Type 1 | Type 2 | Type 1 | Type 2 | Type 3 |
| $\alpha_k$ | $-0.08$ | $-0.19$ | 0.11 | $-0.13$ | 0.13 | $-0.49$ |
| | (0.03) | (0.03) | (0.08) | (0.02) | (0.02) | (0.04) |
| $\beta_k$ | 0.17 | 0.27 | $-0.49$ | 0.46 | 0.03 | $-0.32$ |
| | (0.04) | (0.04) | (0.17) | (0.02) | (0.02) | (0.03) |
| $\delta_k$ | 0.17 | 0.27 | 0.02 | 0.13 | $-0.12$ | 1.01 |
| | (0.04) | (0.05) | (0.06) | (0.04) | (0.02) | (0.07) |
| $\lambda_k$ | 0.37 | 0.36 | 0.04 | 1.73 | 2.10 | 0.00 |
| | (0.15) | (0.26) | (0.07) | (0.02) | (0.06) | (0.00) |
| $r_k$ | 0.87 | 0.90 | 1.51 | 0.40 | 0.15 | 2.74 |
| | (0.13) | (0.15) | (0.32) | (0.02) | (0.02) | (0.04) |
| $\phi_k$ | 1.00 | 0.64 | 0.36 | 0.35 | 0.31 | 0.34 |
| | (-) | (0.05) | (0.05) | (0.05) | (0.04) | (0.06) |
| $\ln L$ | -2393.9 | -2197.6 | | -2149.7 | | |
| $EN(\boldsymbol{\tau})$ | 0.00 | 3.52 | | 15.57 | | |
| ICL | 4811.2 | 4450.3 | | 4394.7 | | |
| NEC | - | 0.018 | | 0.064 | | |

*Notes:* Standard errors in parentheses. Based on our ‘core sample’ of 109 subjects.

# Appendix A3 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.

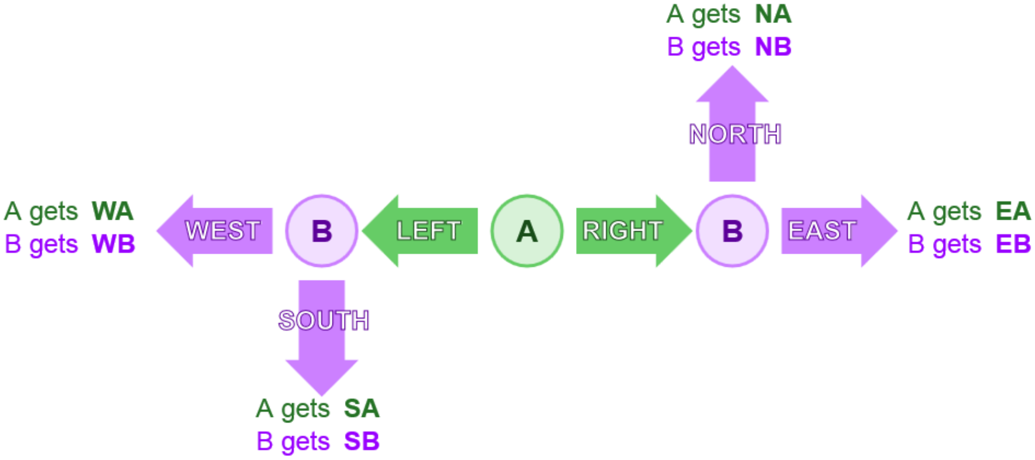
If A chooses LEFT, person B has to choose between WEST or SOUTH. If person A chooses RIGHT, person B has to choose between NORTH and EAST.

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

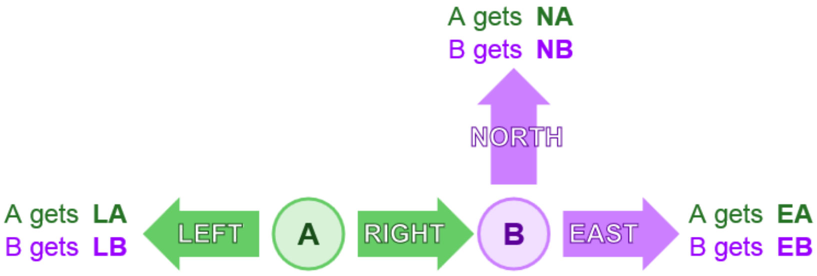
In the decision situation shown in Figure 2, person A first chooses LEFT or RIGHT. If A chooses LEFT, person B has no choice to make. If A chooses RIGHT, B has to choose between NORTH and EAST.

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



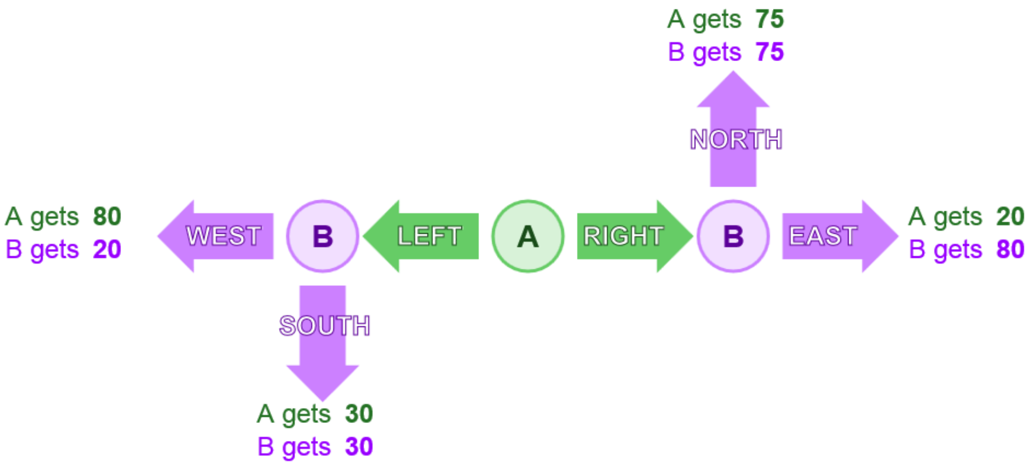
**Example**

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.

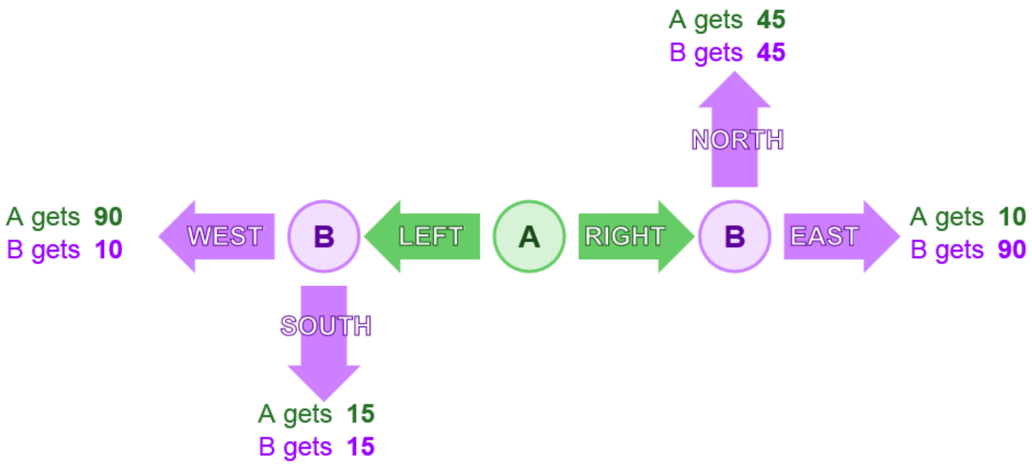
### Decision screens

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:

○ are always the same

○ are sometimes the same

○ are always different

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

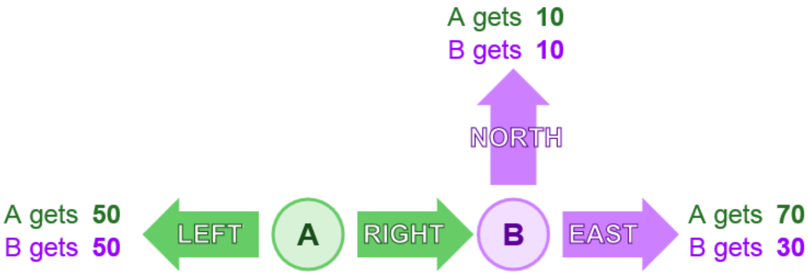
Suppose A chooses LEFT and B chooses SOUTH and EAST. How much would A and B earn?

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**

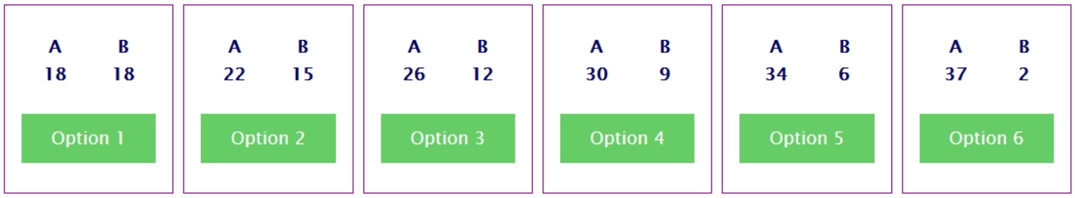
You have reached the end of the instructions. You can still go back by using the menu above. If you are ready, click on ’continue’ below. If you need help, please raise your hand.

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.



| A | B | A | B | A | B | A | B | A | B | A | B |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 18 | 18 | 22 | 15 | 26 | 12 | 30 | 9 | 34 | 6 | 37 | 2 |