CEU CENTRAL EUROPEAN UNIVERSITY

DEPARTMENT OF ECONOMICS
WORKING PAPERS

economics.ceu.edu

Beliefs About People's Prosociality
Eliciting predictions in dictator games

by

András Molnár[1]

Christophe Heintz[2]

2016/1

[1] Department of Economics, Central European University,Budapest, Hungary; Carnegie Mellon University, Pittsburgh, PA 15213, United States, molnar_andras@phd.ceu.edu
[2] Department of Economics, Central European University,Budapest, Hungary

Abstract

One of the most pervasive economic decisions that people have to take is whether to enter an economic interaction. A rational decision process takes into account the probability that the partner will act in a favorable way, making the interaction or the cooperative activity beneficial. Do people actually decide upon such predictions? If yes, are these predictions accurate? We describe a novel experimental method for eliciting participants' implicit beliefs about their partners' prosociality: In a modified dictator game, receivers are offered to forgo what the dictator shall transfer and take a sure amount instead. We then infer receivers' subjective probabilities that the dictator makes a prosocial decision. Our results show that people do form prior beliefs about others' actions based on others' incentives, and that they decide whether to enter an interaction based on these beliefs. People know that others have prosocial as well as selfish preferences, yet the prior beliefs about others' prosocial choices is biased: First, participants underestimate others' prosociality. Second, their predictions about others' choice correlate with their own choice, reflecting a consensus effect. We also find a systematic difference between implicit and explicit predictions of others' choices: Implicit beliefs reflect more trust towards others than explicit statements.

JEL: C91, D63, D84

Keywords: belief, consensus effect, prosociality, dictator game, preditcion

1. Introduction

Holding true beliefs about others' prosociality can be very advantageous: It allows one to know when to engage in beneficial collective actions, and whether one can spare the cost of monitoring, enforcing and other actions meant to influence or constrain others' choice. However, naively believing that others are more prosocial than they actually are can lead to detrimental decisions: Engaging in risky economic interactions and being exploited by opportunistic others. Results from experimental economic games have shown that there is a significant variation in prosociality and cooperativeness within populations (Henrich et al., 2005). This interpersonal variation is often linked to the widely accepted claim that some people have stronger other-regarding preferences than others (e.g. Kahneman et al., 1986; Charness and Rabin, 2002). An adequate belief about others' prosociality should therefore represent accurately the probability that a potential partner will act prosocially and adequately evaluate the risk of entering an interaction.

Examples of behavioral strategies that involve beliefs about others' cooperative or prosocial dispositions are conditional cooperation and partner selection (Willer et al., 2010). Conditional cooperation—cooperate only if you believe that your partner will cooperate as well—is observed to be a robust strategy in human societies (Axelrod and Hamilton, 1981; Boyd and Richerson, 1988; Fehr and Fischbacher, 2004). Partner selection also is a decision based on the beliefs that one has about potential partners. The decision process is such that given any number of partners, one chooses the partner who has the highest probability to cooperate. Partner selection has been argued to provide the social environment for the evolution of the preference for fairness in humans (e.g. Baumard et al., 2013), but this social environment must include people who are able to form accurate beliefs about others' cooperative dispositions. Thus, strategic decision processes require people to estimate others' cooperativeness or prosociality.

Several theories have been proposed about how people predict others' future choice after interacting with them (e.g. Knoepfle et al., 2009; Camerer, 2003, chapter 6). These models describe how people form beliefs about their partner's choice in a given situation on the basis of information about their partner's past behavior in similar situations. Statistical learning algorithms such as Bayesian updating can model this process (e.g. Belot et al., 2012). However, the range of application of these models is limited because they do not apply to novel interactions where people have no knowledge of the history of the partner for the very same type of decision problem.

Such models lack the ability to specify how contextual information is used to make predictions. In particular, we hypothesize that people know that their partners' decisions depend on their incentives, which is not expressed by current models. This is a common sense hypothesis that is further supported by the psychology of social cognition: Humans ascribe beliefs and desires to others, which makes them able to predict others' actions (these types of inferences are made even at a very young age, see, e.g., Woodward 1998).

There are experimental studies examining beliefs and expectations in economic games (Dufwenberg and Gneezy, 2000; Fetchenhauer and Dunning, 2010; Iriberri and Rey-Biel, 2013), but they involve situations where people had information about their partners' past behavior or social norms constrained possible actions.

By contrast, we describe a simple economic experiment that reveals the prior beliefs that people have when they interact with new and anonymous partners. More precisely, we elicit and estimate the beliefs that people have about others' propensity to make

prosocial choices even though they have no information about past behavior. In addition, we ask participants to make explicit estimations about others' choices and are thus able to compare beliefs with these explicit self-reports.

In general, we find that people take into account the incentives that others might have. These incentives include not only others' material payoffs but also the incentives that are related to others' social preferences: When people predict others' choices, they predict that others will sometimes sacrifice their own gains to benefit their partner. Furthermore, our data suggest that one's own other-regarding preferences and cognitive biases such as the better-than-average effect can influence these beliefs. Finally, we find that there are systematic differences between beliefs and explicit estimations: When asked explicitly, people tend to significantly overestimate others' selfish motives.

Hypotheses.
Take a decision context where the following three conditions are satisfied:

1. The partner's choice has consequences on the payoff of the predicting agent.
2. The predicting agent has full knowledge of the partner's choice situation.
3. The predicting agent does not know the partner's history of decisions in similar contexts.

We hypothesize that people decide whether to enter economic interactions on the basis of beliefs about potential partners' intentions (**H1**). On the one hand, this is a common sense assumption: We certainly have beliefs whether an employee will work hard, an employer provide good working conditions, a client pay in time or a furnisher provide the requested goods. On the other hand, this is a strong prediction because it implies that people will compute and rely on implicit probabilistic beliefs about partners' intentions. In our experiment, we assess whether ascribing such beliefs to participants is the best way to explain the observed patterns of behavior. Second, we hypothesize that people take into account their partners' incentives when making their predictions. In particular, we hypothesize that people believe that their partners have prosocial preferences as well as preference for monetary gains (**H2**). Our third hypothesis relates to the content of these prior beliefs: We hypothesize that people perform relatively well when they estimate the probability of selfish or generous choices, which is to say that their beliefs accurately reflect the observed behavior (**H3**). We also hypothesize that people's beliefs are consistent with their reported estimations, there is no systematic difference between them (**H4**). Because predictions often depend on own decision mechanisms as well, we hypothesize that beliefs about others' choice correlate with own choice: A consensus effect emerges between own and predicted choices (**H5**).

2. Methods

2.1. Participants and power analysis

We conducted an a priori power analysis to estimate the minimum sample size required to detect a moderate effect size (Cohen's $d = 0.5$) at a high power level (.95). The sample size that satisfies the above criteria is 52, therefore we decided to stop collecting data after the number of participants in both roles (allocator and recipient) reached 52.

117 participants (58 allocators and 59 recipients[1], 66 female, mean age: 25.3 years) participated in 8 experimental sessions. We intended to have 16 participants in each session, but some of the signed-up participants did not show up, therefore the number of participants per session ranged between 12 and 16. All sessions were conducted in January 2014 in the computer labs of xxx. We recruited the participants through the online Research Participation System of xxx, and they were mostly (but not exclusively) students at xxx. The participant pool was more heterogeneous than in similar studies: We had participants from 35 different countries, covering several academic areas, mainly social sciences and humanities.

2.2. Materials and protocol

We implemented a computer-based two-person experimental game that was a modified version of the dictator game (Forsythe et al., 1994). The experiment was programmed in the software z-Tree (Fischbacher, 2007). At the beginning of a session we randomly assigned participants to Group 1 or Group 2. Then we seated the groups in two adjacent computer rooms.

Before the computer-based main task, we conducted a pen-and-paper risk aversion attitude test (Eckel and Grossman, 2008). In this task we asked the participants to choose among several bets in order to assess their attitude to risk.

After the risk aversion assessment task we assigned Group 1 participants to Role A and Group 2 participants to Role B. Role A had the opportunity to select between two allocations for A and B. For instance, A could choose either 700 units for herself and 200 for her partner or 600 for herself and 600 for her partner. While such alternatives were presented to A, B had—unknowingly to A—the opportunity to accept a sure amount instead of getting the amount allocated by A. If B selected this sure amount, A still received her payoff according to her choice. Importantly, A did not know about B's alternative sure amount option, and B knew that A did not know: Thus B simply had to predict the choice that A would make between her two options. We illustrate a schematic representation of the protocol in Figure 1.

It was necessary to implement a design with asymmetric information so that B's choice would only based on her belief about A's preference between two types of monetary distribution. When the outside option is common knowledge, A's choice can be based on her belief that B will or should take the outside option. This situation leads A to withdraw her concern for B and therefore not express her prosocial preferences. This is indeed what we observed in a pilot study we conducted with 20 participants to check the effect of common knowledge about B's outside option. The results clearly show that A chose the selfish options in the vast majority of cases, regardless of the alternative (total welfare maximizing) option. This result is in line with other findings showing that people are quick to find reasons not to make altruistic choices. More precisely, people withdraw their concern for others as soon as they find that these others are not justified to expect a prosocial choice from them (Heintz et al., 2015). Interestingly, people in role B did predict A's choice and often chose the outside option (see online Supplementary).

[1]Our protocol allowed us to have unequal numbers of allocators and recipients, see Materials and Protocol.

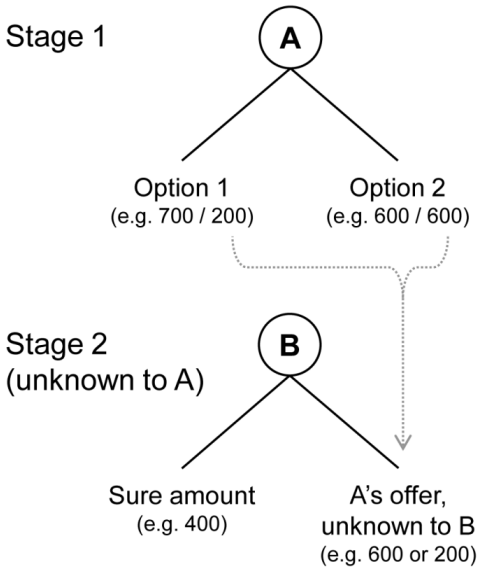

Figure 1: Structure of the main task.
Stage 1: the allocator (A) chooses between two allocation options for herself and the recipient (B). Stage 2 (unknown to A): B can blindly accept A's choice or B can accept a sure amount that is specified in each round. If B accepts the sure amount A still receives her share according to her choice.

Implementing a protocol with asymmetric knowledge about the outside option for B enabled us to make sure that concern for others was at stake and to better control what social preferences were expressed.[2]

Participants in Role B had to make 80 decisions, divided into four blocks with a short break between them. During one block, the two possible allocations among which A had to choose were always the same, but the sure amount that was presented as an option to B varied. This sure amount was always between the higher and the lower transfers that B could receive from A. We randomized the order of the sure amounts within each block: This enabled us to rule out the possibility that participants made consistent choices (i.e. chose the sure amount only above a certain value) only due to the sequence of choices rather than due to stable beliefs about others' dispositions. We also counterbalanced the order of options and blocks across participants. The possible allocations that could be chosen by A and the sure amounts that were presented to B are summarized in Table 1.

We did not simply ask participants to state their minimal acceptable sure amount as in the procedure implemented by Becker et al. (1964).[3] Our procedure did not require participants to understand the complex bidding mechanism implemented in the Becker-

[2]Note that the protocol did not involve any kind of deception. First, every decision that was made by A could affect real payoffs. Second, each A played as B after she completed the first part of the task, thus every participant became aware of the initial information asymmetry. Finally, we debriefed participants after the experiment about the goal and hypotheses of the study, and explained why the information asymmetry was necessary.

[3]In this procedure participants state their minimal acceptable sure amount. The sure amount is then randomly determined. If it is higher than the participant's stated minimal acceptable sure amount, then

Table 1: Payoff structures and sure amounts across blocks

	A's payoff		B's payoff		Sure amount		
Block	Option 1	Option 2	Option 1	Option 2	Min	Max	Increment
6/3 v 5/7	600	500	300	700	300	680	20
7/2 v 6/6	700	600	200	600	200	580	20
5/2 v 6/6	500	600	200	600	200	580	20
5/1 v 1/5	500	100	100	500	100	480	20

Note: A chooses Option 1 or 2, and B either accepts A's choice or takes the sure amount.

DeGroot-Marschak procedure and it allowed us to gather twenty decisions of B for each allocation problem that A faced. These twenty decisions made our analysis less error-prone and enabled us to check the consistency of participants' choice.

Participants in Role A had to select between the above options, but instead of asking them to make the same decisions 20 times, they had to make only one decision for each block. We then created a distribution of these choices and randomly sampled offers for each B across the 80 rounds.

We analyzed participants' beliefs about others' choice when these others faced four different allocation problems. Thus, we could investigate whether people's beliefs about others' choices are sensitive to the specific incentives that these others face. Block 1 created a situation where participants had to choose between own payoff maximization and social welfare maximization (e.g. Charness and Rabin, 2002). Block 2 added a further social incentive to this dilemma: inequality aversion (e.g. Fehr and Schmidt, 1999). Block 3 served as a baseline condition, where we expected that the vast majority would select Option 2, but we allowed for the possibility to make a spiteful choice, which has been observed in economic games (e.g. Levine, 1998). Block 4 was a fixed-sum allocation problem, where we expected that the vast majority would select Option 1. By comparing differences in beliefs between blocks, we determined whether participants ascribed different types of other-regarding dispositions to others.

By varying the sure amounts within blocks, we elicited B's certainty equivalents for A's offer. Therefore, we could estimate each B's subjective belief about A's choice. For instance, in Block 1, if a participant always prefers the sure amount when it is above 360 units, then she believes that her partner will most likely choose Option 1 (a gain of 300 for her) rather than Option 2 (a gain of 700 for her). The expected value of the transfer will be close to 300 and a sure amount of 360 or higher will yield a higher expected payoff for B. If a participant does not accept sure amounts below 600 in the same context, then she believes that her partner will most likely choose Option 2 (a gain of 700 for her) and it is more favorable to accept this offer. In our analysis we estimated each individual's sure amount cut-point (the certainty equivalent of the offer) for each block, below which they accepted the offer and above which they accepted the sure amount. This cut-point enabled us to specify participants' belief about the subjective probabilities of others making selfish or generous choices.

the participant gets the sure amount. If it is smaller, then the participant gets the result of the other alternative, in our case, this would be the amount selected by A.

In order to ensure that participants in Role B understood the consequences of their own and their partner's possible actions, they had to answer a set of nine control questions before the actual experiment. Participants made less than one error on average ($M = 0.55$), and the majority (78%) solved this task without any error, suggesting that participants understood the instructions sufficiently.

After A completed the first task (four rounds where they had to decide between two allocations), they were asked to complete a second task. We presented them with a decision task that was identical to the task done by B: They had to choose whether to accept the transfer of another participant or to take a sure amount. However, instead of interacting in real-time with another participant, they were reacting to decisions made by participants in previous sessions (for the first session, they reacted to the choices made by others who belonged to their own group).

After participants finished the 84 or 80 rounds,[4] we asked them to estimate the number of people (out of 100) who picked the first option in each of the blocks. We decided to elicit natural frequencies instead of standard probabilities because it has been documented that people can represent natural frequencies better than standard probabilities during decision-making (Gigerenzer and Hoffrage, 1995). We provided no feedback about partners' actual choices during the experiment: Participants saw their own payoff and their partners' choices only after the last round and after the explicit estimation task. This process ruled out any learning effects, so that only the prior beliefs could inform actions and estimations.

Although incentivized belief elicitation is a widely used technique in experimental economics, we did not incentivize this estimation task in order to avoid any possible hedging effect due to risk aversion. Hedging can lead to biased predictions if the agent has a financial stake in the predicted event itself (Armantier and Treich, 2013).

After the estimation task, we asked the participants to fill out a short personality test that was a mix of the 20-item MACH-IV inventory (Christie and Geis, 1970) and the 7-item Interpersonal Reactivity Index inventory (Perspective Taking items; Davis 1983). Finally, we collected demographic data (age, gender, citizenship, and academic area) and recorded whether participants had attended any course in game theory or behavioral economics.

Each session lasted about 75 minutes. We selected 4 out of 80 rounds (and additional 2 out of 4 for Group 1) for actual payment in cash.

2.3. Measurements and analysis

We recorded decisions in the risk attitude task and in each round of the main protocol. The latter decisions were binary choices (A: Option 1 or 2, B: accepting A's transfer or accepting sure amount). We also recorded explicit estimations about the number of other participants (out of 100) who chose Option 1.

We analyzed choices of participants in role B as follows. We specified a sure amount for the twenty choices that a participant had to make in a given block, such that:

[4] Participants in Group 1 had to make 4 (Task 1) + 80 (Task 2) = 84 decisions, while participants in Group 2 had to make 80 decisions (Task 1).

$$U(\text{sure amount}) = p_{\text{high transfer}} \times U(\text{high transfer}) + (1 - p_{\text{high transfer}}) \times U(\text{low transfer}) \quad (1)$$

Is there a unique sure amount S_c for which the above equation holds? That is the case if the participant preferred to receive her partner's transfer for any choice when the sure amount was lower than S_c , and she preferred to take the sure amount for any choice when the sure amount was higher than or equal to S_c . Based on this assumption but allowing for error, we determined the most probable value of S_c for a given participant in a given block with respect to two principles:

1. If there is such a value, then it minimizes the number of inconsistent choices. An individual made an inconsistent choice if she chose the partner's transfer when the sure amount was higher than or equal to S_c , or if she chose the sure amount when it was lower than S_c .
2. Isolated inconsistent choices contain less information about subjective valuation than inconsistent choices that are close to the potential cut-point. In other words, when an inconsistent choice occurred far from the cut-point, it was most likely a genuine mistake (e.g. due to inattention), and reveals little about the participant's belief about her partner's choice.

We implemented these two principles as follows. Let d_x denote a participant's decision at a given sure amount level x (ranging from 0 to 19).[5] The binary variable $d_x = 0$ if the participant accepted her partner's transfer and $d_x = 1$ if she took the sure amount. Let $WE(c)$ be the function that attributes a weighted error score for each potential cut-point:

$$WE\,(c) = \sum_{x>c} \frac{|d_x - 1|}{(x-c)^2} + \sum_{x<c} \frac{d_x}{(x-c)^2} + (d_c - 1) \quad (2)$$

The cut-point c_0 is the c that has the smallest weighted error WE(c):

$$c_0 = WE^{-1}\left(\min_{c \in I} [WE(c)]\right) \quad (3)$$

A participant's subjective valuation S_c about the risky choice in a given block is the ${c_0}^{\text{th}}$ sure amount level (e.g. if $c_0 = 3$ in Block 1, then $S_c = 300{+}3{\times}20 = 360$). We used an inverse quadratic distance weighting for errors, because we wanted to underweight outlier errors according to principle (2). While this method included an arbitrary component, we still obtained all of our important results when we used other reasonable methods to determine the cut-points (see online Supplementary).

After calculating the cut-points for each participant in each block (four cut-points per participant), we estimated the variable of main interest: B's subjective belief about the probability that her partner A chooses Option 1 ($p_{\text{high transfer}}$). The cut-point S_c expresses the monetary amount for which B's choice is indifferent between the sure amount and A's transfer. We therefore have:

[5] $x = 0$ when the sure amount is equal to the lower payoff for B in the allocation; $x = 19$ when the sure amount is equal to the higher payoff for B in the allocation minus 20 units.

$$p_{\text{high transfer}} = \frac{U(S_c) - U(\text{low transfer})}{U(\text{high transfer}) - U(\text{low transfer})} \tag{4}$$

In order to calculate the subjective probability, we had to translate monetary gains into utility. For risk neutral people, it is sufficient to assume that $U(c) = c$. But we can only make this direct calculation from the cut-points to probabilities if we assume perfect risk-neutrality, which is not a valid assumption for such interactions. Evidence suggests that the vast majority of people are more or less risk-averse in economic games (Dave et al., 2010; Eckel and Grossman, 2008; Holt and Laury, 2002). Therefore, we adjusted the utility function in view of the risk aversion coefficient that we measured in the first task (see online Supplementary for details). In the further sections we refer to these probabilities as inferred implicit beliefs of participants about others' behavior.

For within-individual analyses, we performed one sample t-tests, when we compared explicit estimation or beliefs to observed behavior (i.e. A's decisions) within blocks, and we used paired samples t-tests, when we compared estimations to beliefs within or between blocks. For between-individual comparison we performed independent samples t-tests or one-way ANOVAs, depending on the number of levels of the independent variable. Statistical analysis was performed in SPSS (v.22). All differences reported in the main text are significant, $p = .05$, two-tailed, and have a medium or large effect size (Cohen's $d > 0.3$), if not noted otherwise.

3. Results

First, we discuss participants' choices in Role A. These choices clearly reveal that participants have prosocial preferences. The proportion of participants who selected Option 1 is illustrated in Figure 2 (white bars). Participants did not always maximize their own monetary payoff when they could increase social welfare. In Blocks 1 and 2 we replicated the results by Charness and Rabin (2002). Blocks 3 and 4 served as baselines in our study and we found no surprising results here: In Block 3 only 4 participants (7%) selected Option 1, whereas in Block 4 52 participants (90%) did.

3.1. People rely on probabilistic prior beliefs when they enter interactions

In our experiment, B faced the risk that A would choose the option that is less favorable to them.[6] Did participants compile this risk at all when they made a strategic decision? We assume that a participant did compile the risk of interaction if she systematically chose the sure option above a certain amount and the partner's transfer below this amount. An inconsistent choice is either choosing the transfer when the sure option is above the cut-point as defined above (equation 2), or taking the sure option when it is below this cut-point. It is highly improbable that a person choosing randomly makes

[6]Participants in role B came from Group 1 (who had previously played in role A) and Group 2 (who had not previously played in role A). There is no significant difference in any of the possible pair-wise comparisons between these two groups (implicit or explicit, all $p > .1$) except for the implicit beliefs in Block 4, $t(116) = 2.160$, $p = .033$, Cohen's $d = 0.404$, 95% CI $[0.01, 0.20]$. Therefore, in the following sections we report the combined results of Group 1 and 2, unless noted otherwise.

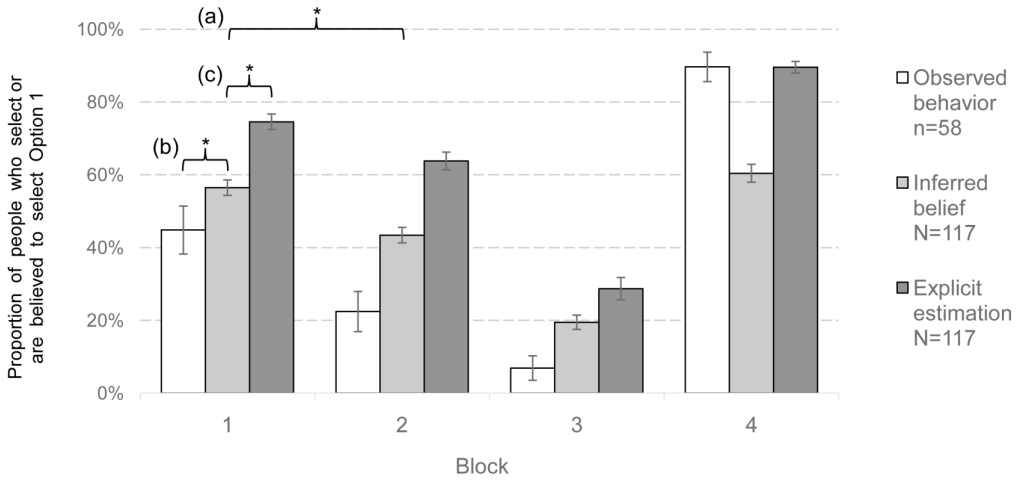

Figure 2: Means of observed behavior (white), inferred beliefs (light gray), and explicit estimations (dark gray) about the proportion of people who select or are believed to select Option 1 across blocks. Error bars represent standard errors.
There are three main effects:
(a) We can observe significant differences between the blocks.
(b) There are systematic differences between observed behavior and implicit beliefs: People implicitly overestimate the proportion of selfishly behaving people, except for Block 4.
(c) There are systematic differences between explicit estimations and implicit beliefs: People overestimate the proportion of selfishly behaving people in their explicit estimations even more.
* stands for $p < .01$, two-tailed.

no inconsistent choice.[7] Therefore, if the number of inconsistent choices is close to zero, then the participants' choices are based on a determinate valuation of the risky choice. This is indeed what we found. Participants were quite consistent across rounds: On average, participants made less than 1 inconsistent choice out of 20 decisions per block, $M(SD) = 0.73(0.98)$. We can therefore assert that our participants had a clear valuation of the risky choice and acted on stable probabilistic beliefs about the likelihood that their partners would select the option that was favorable to them.

Below we question how these beliefs were formed. In particular, were these probabilistic beliefs formed about the possible intentions of their partner, or were they independent from others' intentions? We could measure this potential dependency because A had different incentives for Option 1 and 2 across blocks.

3.2. People are sensitive to others' incentives

Results clearly indicate that participants in Role B believed that their partners would behave differently across blocks. More precisely, they knew that the probability that their partner would transfer the high amount (Option 2) would change when the stakes for the partner changed. For instance, in Block 1, participants in Role A could prefer Option 2 because it meant sacrificing little of the benefits from Option 1 and increasing the transfer to their partner by more than 100%. Thus, participants concerned with social welfare chose Option 2, and participants concerned only with their own material benefit selected Option 1. By contrast, in Block 3, participants in Role A who were only concerned with their own material benefit also chose to transfer the high amount to their partner.

All of the possible cross-block comparisons yield a significant difference with a medium or large effect size, except for the comparison of implicit beliefs between Blocks 1 and 4, $t(116) = 2.113$, $p = .035$, Cohen's $d = 0.197$, 95% CI [0.00, 0.08]. Since the partner's monetary incentive (the allocator's monetary gain for each option) and social incentive (the recipient's monetary gain for each option) were the only difference between blocks, the fact that recipients behaved differently across blocks is best explained by the hypothesis that they were sensitive to their partners' incentives and predicted their partners' behavior based on these assumed selfish and social preferences. This sensitivity to others' incentives allowed participants to adapt their behavior to the behavior of their partner. Participants assumed that others had prosocial preferences; otherwise, they would have predicted that their partner would always choose Option 1 in all blocks except for Block 3. This was clearly not the case. Moreover, the sensitivity was adequate: The probabilistic beliefs varied in the same direction as the population's actual distribution of choices (cf. light gray and white bars in Figure 2).

At the individual level, we could measure sensitivity by the variance of individual beliefs across blocks. Participants in our experiment demonstrated different sensitivity to their partners' incentives: Some of them barely changed predictions about their partners' behavior across blocks. This low level of sensitivity might be explained by the

[7] The exact probability is $p = .00036$. Explanation: the number of possible decision sets is 220. There are 21 perfectly consistent decisions sets without any errors. If the cut-point is at 0, there are 19 different decision sets with one error. If the cut-point is at 1 or 20, there are 18 different decision sets with one error. If the cut-point is between 2 and 19, there are 17 different decision sets with one error. That gives a sum of $21 + 19 + 18 \times 2 + 17 \times 18 = 382$ different decision sets with one or no error. The probability or randomly sampling a choice set with less than two errors among all possible choice sets is $p = 382/220 = .00036$. Likewise, the probability of making only one mistake is very low.

inability to take others' incentives into account. However, there is no significant correlation between the self-reported perspective taking ability (score maintained in the IRI-PT questionnaire) and the sensitivity to different contexts ($p > .8$). We cannot conclude that individual differences in sensitivity are due to differences in ability to take others' perspectives. It is possible that the lack of sensitivity in our experimental setting resulted from some participants' inability to identify the experimental context as strategic and their incomplete understanding of the experimental task. Some might have doubted that they were interacting with humans due to the nature of the anonymous computerized task. More controversially, our measure might have grasped the ability to take others' perspective better than the IRI-PT questionnaire.

3.3. Prior beliefs are not accurate

Although B were able to predict the variation in their partners' behavior across blocks, their sensitivity to their partners' incentives did not warrant the formation of true beliefs (cf. light gray and white bars in Figure 2). Participants' beliefs were only qualitatively consistent with the actual behavior. For all blocks but Block 4, people estimated that the proportion of others who chose the selfish option is much higher than it really was (cf. dark gray and white bars in Figure 2). When asked explicitly, more than one third of the participants (45/117) estimated that the majority of A would always choose the own payoff maximizing option and the majority (97/117) estimated that others would more likely select the own payoff maximizing option in at least 3 blocks. The discrepancy was the most striking in Block 2: On average, people expected that 64% would select 700/200 rather than 600/600, while in reality only 22% made this choice. This result shows, first, that the common priors elicited in these situations did not adequately represent what people actually chose. And second, that, as Fetchenhauer and Dunning (2009) already noted: "People underestimate the degree to which other people follow fairness norms in economic games, such as in the dictator or ultimatum games" (p. 265).

3.4. Explicit estimations are more pessimistic than implicit beliefs

There is a remarkable difference between implicit beliefs driving choices and their corresponding explicit estimations. Apart from the observation that explicit and implicit beliefs correlated positively (Blocks 1-4: $r(115) = .34, .44, .45, .40$, respectively, all $p < .001$), explicit estimations tended to underestimate others' prosociality even more than implicit beliefs, reflecting more pessimistic predictions.

3.5. Consensus effect between predictions and choices

Consistent with the results in economic research on consensus effects (e.g. Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Bicchieri and Xiao, 2007; Reuben et al., 2009), we also found a strong consensus effect in Blocks 1 and 2 (Blocks 3 and 4 could not be analyzed because the vast majority chose the same option in these blocks). The correlation between own choices and beliefs about others' choices in Blocks 1 and 2 was significant, $r(56) = .393$, $p = .002$, and $r(56) = .504$, $p < .001$, respectively. In both blocks, there was a significant difference between the beliefs of participants who acted selfishly and who acted prosocially (cf. dark and light gray in Figure 3). These results are also consistent with the recent finding that one's own preferences influence

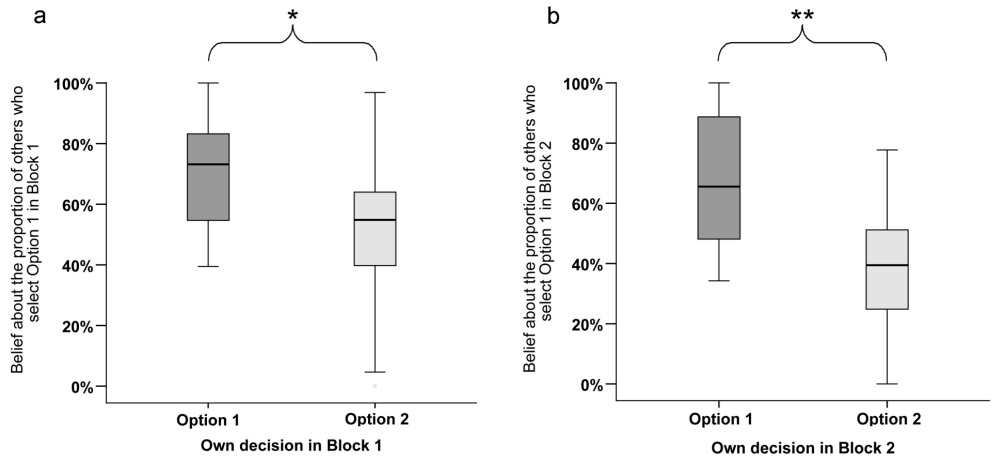

Figure 3: Beliefs about others' choice in Block 1 (a) and Block 2 (b), sorted by own choice. Data are illustrated on Tukey boxplots: Boxes denote the interquartile ranges (IQR), the horizontal lines within the boxes denote the medians, and error bars denote the highest and lowest values within 1.5 IQR. There is a main effect of consensus in both blocks: Participants who picked Option 1 (dark gray) believed that a higher proportion of others would also select Option 1, compared to participants who selected Option 2 (light gray). Furthermore, the individual beliefs are highly distributed, indicating that participants did not simply project their own choices to others. * stands for $p = .002$. ** stands for $p < .001$

one's beliefs about others' behavior in social dilemmas, even after controlling for feedback about others' behavior (Blanco et al., 2014).

Figure 3 also illustrates the distribution of individual implicit beliefs in Blocks 1 and 2. We can observe an interesting pattern if we compare the range of distributions of beliefs in Blocks 1 and 2. People who selected generous options (light gray in Figure 3) were more heterogeneous in their beliefs and their distribution of beliefs covered almost the whole range of possible beliefs. By contrast, people who selected selfish options (dark gray) almost never believed that others would more likely select the generous option than the selfish one.

3.6. Summary of results

Our analysis showed that people made decisions in strategic interactions on the basis of implicit probabilistic beliefs (R1) that took into account others' prosociality and the contextual incentives that others faced (R2). Meanwhile, beliefs about others' choices were off the target: People tended to overestimate the number of others that go for own payoff maximization and have no concern for social welfare (R3). This overestimation of selfishness was even more salient in explicit estimations (R4). We also reported significant consensus effects: People relied on their own (actual or hypothetical) choice when they predicted others' (R5).

4. Discussion

Why did some people, even those who selected the generous option, believe that others were most likely to select the selfish option? And why did people, on average,

underestimate others' prosociality in the current experimental context? Are people on average, systematically biased towards the assumption that others are more selfish than themselves? In the following section, we discuss different accounts that can explain such findings.

First, it is plausible that participants' choices did not only express probabilistic beliefs about others' choice but also a disutility from being the victim of someone else's intentional choice. Bohnet and Zeckhauser (2004) use the term betrayal aversion for the finding that "Individuals are much more willing to take risks when the outcome is due to chance than when it depends on whether another player proves trustworthy" (p. 479). This means that the participants' low willingness to engage in risky social interactions contributed to their apparent skepticism. However, in this case the explicit estimations should have been more optimistic than the implicit beliefs, and we observed the opposite. Betrayal aversion can therefore explain our results only if we supplement it with some plausible ad hoc hypotheses explaining explicit beliefs. For instance, it could be argued that explicit beliefs are mainly post-hoc rationalizations, strongly influenced by the saliency of betrayal.

An alternative explanation for generally pessimistic beliefs is that people have a sampling bias in their social learning process. Because people avoid interacting with others who they presume to be too selfish, they do not gather information about them. Therefore, people cannot correct their beliefs when their presumption is false: They cannot learn that some of the people they categorize as too selfish are in fact prosocial. By contrast, people willingly interact with others those they consider prosocial. Eventually, people learn that some of them, in fact, behave selfishly. Thus, initial positive beliefs can be falsified, but initial negative beliefs cannot. This asymmetry in feedback can lead people to form incorrect beliefs that others are, on average, more selfish than they actually are (Fetchenhauer and Dunning, 2010).

Finally, the apparent underestimation of others' prosociality might stem from the overestimation of one's own prosociality. People might believe that they have stronger prosocial preferences than others, which is consistent with the widely reported better-than-average effect (Alicke et al., 1995; Epley and Dunning, 2000; Larrick et al., 2007). If the majority of people think that they are more generous than others, then, on average, there will be a systematic overestimation of selfishness in beliefs.

Another interesting pattern in our data is the remarkable difference between implicit beliefs and explicit estimations about others' behavior. It has been reported that explicitly stated estimations might not reflect true beliefs, and that people often fail to best-respond their own stated beliefs (Costa-Gomes and Weizsäcker, 2008). However, the systematic differences suggest that there is more than the mere inability to best-respond beliefs: If this is the case, we should not expect systematic differences but a random discrepancy. The findings about systematic differences between beliefs and estimations are consistent with the results described by Fetchenhauer and Dunning (2009), who found that people systematically underestimate the trustworthiness of others explicitly, but are more optimistic on the behavioral level. What might be the underlying process that leads to such differences between explicit and implicit estimations? One possible account is that when people make explicit estimations, they rely more on reflective and rational thinking, which is associated more with selfish behavior compared to the automatic and intuitive thinking (Rand et al., 2012; Zaki and Mitchell, 2013). However, there is no general agreement in the literature about this relation. There are authors who argue for

the opposite: People are intuitively selfish and prosociality requires reflective thinking (DeWall et al., 2008; Steinbeis et al., 2012).

It might also be the case that the observed discrepancy between implicit and explicit beliefs is not really due to the difference between the processes of belief formation, but due to a slight difference in the framing of the question. It has been documented that people are more optimistic in their judgments when they have to evaluate the generosity of one individual drawn from a target population compared to when they have to estimate the generosity of the same population in general (Belot et al., 2012; Critcher and Dunning, 2013). Since we elicited implicit beliefs in interactions with one individual at a time and asked for explicit estimations about a hypothetical population of 100 people, the remarkable differences between implicit and explicit beliefs might be explained at least partly by the above framing effect.

5. Conclusions

The recent experimental economic literature and social theory has mainly focused on social preferences as crucial aspects of human sociality, which allow our rich social and economic life (e.g. Fehr and Fischbacher, 2003; Gintis et al., 2003; Baumard et al., 2013). On one hand, people should be able to attribute such preferences to other agents: The mere existence of social preferences would not lead us far if they were not combined with the knowledge that people have them. On the other hand, one cannot blindly trust others' prosociality, one must be vigilant and avoid disadvantageous economic interactions. Acquiring and processing information that helps to infer others' social dispositions is therefore extremely advantageous. If one doubts that a given partner will act in one's favor then one will benefit by not interacting at all, by selecting another partner, or by making a contract that ensures that the partner will behave fairly.

In this paper we provide evidence that people form probabilistic beliefs about others' behavior before entering a risky economic interaction, and that they know others have other-regarding preferences. The probabilistic belief that one's partner will choose a beneficial option is formed in view of the partner's material and social incentives.

We have shown that people assume that their partners have prosocial dispositions, even before having information about their personality traits or individual history. However, we observe that these prior beliefs are off the target: People systematically underestimate the power of prosocial dispositions and overestimate the probability of selfish acts. This underestimation of others' prosociality is most likely an effect of systematic self-deception (Alicke et al., 1995): People think of themselves as more prosocial than others (Epley and Dunning, 2000).

In everyday situations people usually have access to partner-related information like history, reputation, or personality traits, and they are usually more familiar with the context. We can therefore predict that they usually are more accurate in their predictions of their partners' social choices. Accurate estimation of others' motives might depend on several other aspects of an interaction such as social norms involved in the context, commitment, or group belonging. For instance, the social norm of entitlement[8] might

[8]That is, people recognize that others are entitled to receive a certain income or to possess certain rights because they have earned them by providing an effort, see Hoffman and Spitzer (1985).

lead to more accurate and homogeneous beliefs about others' behavior. Such additional factors might create strong expectations about others' behavior, which, in turn, might ground behavior itself (Bicchieri and Xiao, 2007). How social norms affect prior beliefs formation and how such effects interact with partners' history or perceived personal traits are potential questions.

References

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., and Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology*, 68(5):804.

Armantier, O. and Treich, N. (2013). Eliciting beliefs: Proper scoring rules, incentives, stakes and hedging. *European Economic Review*, 62:17–40.

Axelrod, R. and Hamilton, W. (1981). The evolution of cooperation. *Science*, 211(4489):1390–1396.

Baumard, N., André, J.-B., and Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, 36(01):59–78.

Becker, G. M., Degroot, M. H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science*, 9(3):226–232.

Belot, M., Bhaskar, V., and Van De Ven, J. (2012). Can observers predict trustworthiness? *Review of Economics and Statistics*, 94(1):246–259.

Bicchieri, C. and Xiao, E. (2007). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22:191–208.

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

Bohnet, I. and Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization*, 55(4):467–484.

Boyd, R. and Richerson, P. J. (1988). The evolution of reciprocity in sizable groups. *Journal of Theoretical Biology*, 132(3):337–356.

Camerer, C. F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

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

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

Christie, R. and Geis, F. (1970). Scale construction. *Studies in machiavellianism*, pages 10–34.

Costa-Gomes, M. A. and Weizsäcker, G. (2008). Stated beliefs and play in normal-form games. *The Review of Economic Studies*, 75(3):729–762.

Critcher, C. R. and Dunning, D. (2013). Predicting persons' versus a person's goodness: Behavioral forecasts diverge for individuals versus populations. *Journal of Personality and Social Psychology*, 104(1):28–44.

Dave, C., Eckel, C. C., Johnson, C. A., and Rojas, C. (2010). Eliciting risk preferences: When is simple better? *Journal of Risk and Uncertainty*, 41(3):219–243.

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1):113–126.

DeWall, C. N., Baumeister, R. F., Gailliot, M. T., and Maner, J. K. (2008). Depletion makes the heart grow less helpful: Helping as a function of self-regulatory energy and genetic relatedness. *Personality and Social Psychology Bulletin*, 34(12):1653–1662.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30(2):163–182.

Eckel, C. C. and Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, 68(1):1–17.

Epley, N. and Dunning, D. (2000). Feeling "holier than thou": Are self-serving assessments produced by errors in self- or social prediction? *Journal of Personality and Social Psychology*, 79(6):861–875.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785–791.

Fehr, E. and Fischbacher, U. (2004). Social norms and human cooperation. *Trends in Cognitive Sciences*, 8(4):185–190.

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

Fetchenhauer, D. and Dunning, D. (2009). Do people trust too much or too little? *Journal of Economic Psychology*, 30(3):263–276.

Fetchenhauer, D. and Dunning, D. (2010). Why so cynical? Asymmetric feedback underlies misguided skepticism regarding the trustworthiness of others. *Psychological Science*, 21(2):189–193.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Forsythe, R., Horowitz, J. L., Savin, N., and Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6(3):347–369.

Gigerenzer, G. and Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102(4):684–704.

Gintis, H., Bowles, S., Boyd, R., and Fehr, E. (2003). Explaining altruistic behavior in humans. *Evolution and Human Behavior*, 24(3):153–172.

Heintz, C., Celse, J., Giardini, F., and Max, S. (2015). Facing expectations: Those that we prefer to fulfil and those that we disregard. *Judgment and Decision Making*, 10(5):442–455.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M., Marlowe, F. W., Patton, J. Q., and Tracer, D. (2005). Economic man in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, 28(06):795–815; discussion 815–55.

Hoffman, E. and Spitzer, M. L. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distributive justice. *The Journal of Legal Studies*, 14(2):259.

Holt, C. A. and Laury, S. (2002). Risk aversion and incentive effects. *The American Economic Review*, 92(5):1644–1655.

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

Kahneman, D., Knetsch, J. L., and Thaler, R. (1986). Fairness as a constraint on profit seeking: Entitlements in the market. *The American Economic Review*, 76(4):728–741.

Knoepfle, D. T., Tao-yi Wang, J., and Camerer, C. F. (2009). Studying learning in games using eye-tracking. *Journal of the European Economic Association*, 7(2-3):388–398.

Larrick, R. P., Burson, K. a., and Soll, J. B. (2007). Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not). *Organizational Behavior and Human Decision Processes*, 102(1):76–94.

Levine, D. K. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics*, 1(3):593–622.

Rand, D. G., Greene, J. D., and Nowak, M. a. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416):427–430.

Reuben, E., Sapienza, P., and Zingales, L. (2009). Is mistrust self-fulfilling? *Economics Letters*, 104(2):89–91.

Steinbeis, N., Bernhardt, B. C., and Singer, T. (2012). Impulse control and underlying functions of the left DLPFC mediate age-related and age-independent individual differences in strategic social behavior. *Neuron*, 73(5):1040–1051.

Willer, R., Feinberg, M., Irwin, K., Schultz, M., and Simpson, B. (2010). *Handbook of the Sociology of Morality*. Handbooks of Sociology and Social Research. Springer New York, New York, NY.

Woodward, A. (1998). Infants selectively encode the goal object of an actors reach. *Cognition*, 69, 134.

Zaki, J. and Mitchell, J. P. (2013). Intuitive prosociality. *Current Directions in Psychological Science*, 22(6):466–470.