Journal of Economic Psychology 82 (2021) 102347

Contents lists available at ScienceDirect

Journal of Economic Psychology

journal homepage: www.elsevier.com/locate/joep

Simple guilt and cooperation[☆]

Ronald Peeters[a,*], Marc Vorsatz[b]

[a] Department of Economics, University of Otago, PO Box 56, Dunedin 9054, New Zealand
[b] Departamento de Análisis Económico, Universidad Nacional de Educación a Distancia, Calle Senda del Rey 11, 28040 Madrid, Spain

ARTICLE INFO

JEL classification:
C72
C91

Keywords:
Psychological game theory
Guilt
Prisoner's dilemma

ABSTRACT

We introduce simple guilt into a generic prisoner's dilemma (PD) game and solve for the equilibria of the resulting psychological game. It is shown that for all guilt parameters, it is a pure strategy equilibrium that both players defect. But if the guilt parameter surpasses a threshold, a mixed strategy equilibrium and a pure strategy equilibrium in which both players cooperate emerge. We implement three payoff constellations of the PD game in a laboratory experiment and find in line with our equilibrium analysis that first- and second-order beliefs are highly correlated and that the probability of cooperation depends positively on these beliefs. Maximum likelihood estimations of a model of noisy introspection reveal that experimental data is best fitted with positive guilt levels and that omission of guilt results in a substantial increase in the noise parameters.

1. Introduction

The observation that individual (expected) payoff maximization may lead to a socially undesirable (Pareto inefficient) outcome is the key insight of the Prisoner's Dilemma (PD) game. But by now it is also well-established that a non-negligible fraction of subjects participating in laboratory experiments decides to cooperate in the PD game even though they should not do so from a purely materialistic point of view (see, Chaudhuri, 2011, for an overview). Rationalizations of this behavior include other regarding preferences—among which we would like to highlight models of altruism (cf. Andreoni, 1990, inequality aversion (cf. Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999) and preferences for efficiency (cf. Engelmann & Strobel, 2004)—, intentions/reciprocity (cf. Cox et al., 2007; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006; Rabin, 1993), and emotions (cf. Eisenberg, 2000; Elster, 1998).

The literature in social psychology (cf. Baumeister et al., 1994) emphasizes the role of guilt for the maintenance, protection, and strengthening of interpersonal relationships. This emotion motivates individuals in particular to exhibit pro-social behavior. In the economic literature, Battigalli and Dufwenberg (2007, 2009) define simple guilt as the degree by which player i suffers from letting another player j down towards her payoff expectation. Since the payoff expectations of player j depend on her first-order beliefs about the strategy of player i, the expected let-down of player i towards player j is related to i's second-order beliefs. That is, the utility function of the players depend on second-order beliefs. Evidence on the prevalence of guilt motives in experimental settings include Charness and Dufwenberg (2006) who study trust games with pre-play communication, Miettinen and Suetens (2008) who

[☆] We are grateful to Amber Heijne, Martin Strobel, Peter Werner, and the two anonymous reviewers for their valuable advices. Financial support from the Spanish Ministry of Economics and Competitiveness, Spain, through project PGC2018-096977-B-I00, is gratefully acknowledged. All materials related to this study (z-tree programs, raw data, Stata-scripts, and Mathematica programs) are publicly available on the Open Science Framework (OSF) under the link https://osf.io/dqjmy/?view_only=c44d40abf12b4836a221778e133dbcd1.

[*] Corresponding author.
E-mail addresses: ronald.peeters@otago.ac.nz (R. Peeters), mvorsatz@cee.uned.es (M. Vorsatz).

https://doi.org/10.1016/j.joep.2020.102347
Received 19 February 2020; Received in revised form 15 November 2020; Accepted 17 November 2020
Available online 21 November 2020

consider a PD game with voluntary pre-play communication that also introduces a penalty for unilateral defectors, Dufwenberg et al. (2011) who focus on framing effects in public good games, Battigalli et al. (2013) who consider games of strategic information transmission, Bracht and Regner (2013) and Bellemare et al. (2018, 2019) who analyze binary trust and dictator games, Dhami et al. (2019) who theoretically relate reciprocity, simple guilt, and intentions in a public goods game to each other and establish experimentally, using the strategy-method, that second-order beliefs have a significant effect on actions, and Patel and Smith (2019) who study the provision of public goods. Our paper aims at contributing to this literature by interpreting the experimental data as the outcome of a mixed strategy equilibrium of the psychological game (cf. Geanakoplos et al., 1989) induced by simple guilt and by determining, for various payoff constellations, the degree of guilt aversion that is consistent with the experimental data.

In our theoretical analysis, we introduce simple guilt – precisely as conceptualized by Battigalli and Dufwenberg (2007) – into a symmetric PD game and solve for the equilibria of the resulting psychological game. The crucial consequence of introducing psychological costs in the form of simple guilt into the utility function is that player i lets player j down by a strictly positive amount only if she expects player j her to contribute with a strictly positive probability, but she finally decides to defect. That is, psychological costs can only be positive if a player defects; never if she cooperates. This insight leads to the following equilibrium specification (Proposition 1).

(a) Defection for both players remains a pure strategy equilibrium for all values of the guilt parameter. The idea is that if player i is sure that player j thinks that player i defects with probability 1, then there is no psychological cost of defection and the standard analysis applies.
(b) For sufficiently high guilt parameters, the pure strategy profile in which both players cooperate can be sustained as an equilibrium. The reason is that the benefits from reducing psychological costs to 0 (by cooperating instead of defecting) more than offset the associated loss in material payoffs.
(c) For guilt parameters that surpass the threshold, there is also a mixed strategy equilibrium.
(d) There is no asymmetric equilibrium in pure or mixed strategies.

In our experiment, we consider three different payoff configurations that allow us to assess the robustness of our results. Games are played one-shot. It is our main objective to interpret experimental behavior as the mixed strategy equilibrium and derive from there the degree of guilt that is consistent with the data. To do that, observe that in the mixed strategy equilibrium, first- and second-order beliefs coincide with the probability that a player cooperates (equilibrium beliefs are correct) and therefore, we do not ask subjects only about their actions, but also elicit their beliefs at the individual level in an incentive compatible way. For first-order beliefs we use the Quadratic Scoring Rule; for second-order beliefs we apply the Interval Scoring Rule.

We find for all three payoff variations of the PD game that there is a high correlation between first- and second-order beliefs and that the cooperation rate is lower than the average first- and average second-order belief. In fact, depending on the payoff configuration, cooperation rates are between 0.23 and 0.26, while the first- and second-order beliefs range from 0.33 to about 0.40 (Result 1). This contradicts the mixed strategy equilibrium hypothesis on two grounds. First, both average first- and average second-order beliefs are not consistent with the observed cooperation rates (beliefs are significantly higher). Second, the cooperation rate does not differ significantly between treatments, which is only consistent with the equilibrium comparative statics if guilt varies between the three payoff variations. There is evidence of context-dependent guilt in the literature, for example Bellemare et al. (2018), but assuming a non-constant guilt has the methodological drawback that few restrictions are imposed so that a wide variety of behavior (cooperation rates) can be sustained by the model. The theoretical analysis also reveals that for a given guilt parameter, there is a positive dependence of the cooperation rate on first- and second-order beliefs. Probit estimations confirm this theoretical prediction (Result 2). Finally, motivated by the aforementioned contradictions, we estimate the model of noisy introspection introduced by Goeree and Holt (2004) as an alternative to the mixed strategy equilibrium hypothesis. The most important finding in this respect is that the estimate of the guilt parameter is for all three payoff specifications substantially bounded away from zero, which suggests that some part of the behavior is better explained by guilt than by bounded rationality. This interpretation is further strengthened if we compare the estimation results of the noisy introspection model with guilt to that without guilt. It turns out that the estimates of the noise parameters are larger for the model without guilt, that is, guilt functions as a partial substitute for these noise parameters (Result 3).

2. A model of simple guilt in the prisoner's dilemma

There are two players $i \in \{1,2\}$ who have to decide simultaneously and independently between "cooperating" (C) and "defecting" (D). That is, the strategy space of player i is equal to $S_i = \{C, D\}$. Let $s_i \in S_i$ be a particular strategy for player i. We denote generic strategy profiles by $s = (s_1, s_2)$. Material payoffs are as depicted in the bi-matrix below, where $c > a > d > b$ and $a + d > b + c$. Following standard conventions, player 1 selects rows and player 2 selects columns. Also, in each particular cell of the bi-matrix, the first number corresponds to the material payoff of player 1 and the second number to the material payoff of player 2. For example, the material payoff of player 1 at profile $s = (C, D)$ is $\pi_1(C, D) = b$.

$$
\begin{array}{c|c|c|}
 & C & D \\
\hline
C & a, a & b, c \\
\hline
D & c, b & d, d \\
\hline
\end{array}
$$

Let α_i be the first-order belief of player i that the other player j chooses strategy $s_j = C$. The expected payoffs $\pi_i(s_i \mid \alpha_i)$ of player i from playing strategy s_i are then given by

$$\pi_i(C \mid \alpha_i) = \alpha_i\, a + (1-\alpha_i)\, b \qquad \text{and} \qquad \pi_i(D \mid \alpha_i) = \alpha_i\, c + (1-\alpha_i)\, d.$$

Let $G_{s_i}(s_j, \alpha_j)$ be the amount by which player i lets player j down towards her payoff expectations at the strategy profile $s = (s_i, s_j)$ given that player j holds the first-order belief α_j. We assume that

$$\begin{aligned} G_D(C, \alpha_j) &= \max\{0\,;\, \pi_j(C \mid \alpha_j) - b\} = \alpha_j\,(a-b), \\ G_D(D, \alpha_j) &= \max\{0\,;\, \pi_i(D \mid \alpha_j) - d\} = \alpha_j\,(c-d), \\ G_C(C, \alpha_j) &= \max\{0\,;\, \pi_i(C \mid \alpha_j) - a\} = 0, \end{aligned}$$

and

$$G_C(D, \alpha_j) = \max\{0\,;\, \pi_j(D \mid \alpha_j) - c\} = 0.$$

Replacing player j's first-order belief about player i's play (α_j) by player i's second-order belief about player j's belief about player i's play (β_i), we obtain player i's expectation about how much player j feels being let down towards her payoff expectations at profile s:

$$\begin{aligned} \widetilde{G}_D(C, \beta_i) &= \beta_i\,(a-b), \\ \widetilde{G}_D(D, \beta_i) &= \beta_i\,(c-d), \end{aligned}$$

and

$$\widetilde{G}_C(C, \beta_i) = \widetilde{G}_C(D, \beta_i) = 0.$$

Now, let

$$U_i(s_i \mid \alpha_i, \beta_i) = \alpha_i\,[\,\pi_i(s_i, C) - \theta \cdot \widetilde{G}_{s_i}(C, \beta_i)\,] + (1-\alpha_i)\,[\,\pi_i(s_i, D) - \theta \cdot \widetilde{G}_{s_i}(D, \beta_i)\,]$$

be the expected utility of player i from playing s_i when her first-order belief is equal to α_i and her second-order belief is equal to β_i. Here $\theta \geq 0$ captures a player's sensitivity towards letting down the other player, which is assumed to be homogeneous across players. Then,

$$U_i(C \mid \alpha_i) = \alpha_i\, a + (1-\alpha_i)\, b$$

and

$$\begin{aligned} U_i(D \mid \alpha_i, \beta_i) \;&=\; \alpha_i\,[\,c - \theta \cdot \widetilde{G}_D(\beta_i, C)\,] + (1-\alpha_i)\,[\,d - \theta \cdot \widetilde{G}_D(\beta_i, D)\,] \\ &=\; \alpha_i\,[\,c - \theta \cdot \beta_i\,(a-b)\,] + (1-\alpha_i)\,[\,d - \theta \cdot \beta_i\,(c-d)\,] \\ &=\; \alpha_i\, c + (1-\alpha_i)\, d - \theta \cdot \beta_i\,[\,\alpha_i\,(a-b) + (1-\alpha_i)\,(c-d)\,]. \end{aligned}$$

Everything else equal, $U_i(D \mid \alpha_i, \beta_i)$ is decreasing in θ and in β_i.

We are going to analyze pure and mixed strategy equilibria, so let $\sigma_i \in \Sigma_i = [0,1]$ be a mixed strategy for player i, where σ_i denotes the probability that player i chooses strategy $s_i = C$. The expected utility of player i from strategy σ_i is then given by

$$U_i(\sigma_i \mid \alpha_i, \beta_i) = \sigma_i\, U_i(C \mid \alpha_i, \beta_i) + (1-\sigma_i)\, U_i(D \mid \alpha_i, \beta_i).$$

Finally, note that the *psychological prisoner's dilemma game* is completely described by the set of players $N = \{1,2\}$, the players' strategy spaces $\Sigma \equiv \Sigma_1 \times \Sigma_2 = [0,1]^2$, and their expected utilities $U_i(\sigma_i \mid \alpha_i, \beta_i)$ induced by their first- and second-order beliefs.

The *psychological equilibrium*, as defined by Geanakoplos et al. (1989), consists of two parts. First, equilibrium beliefs have to be correct. In our case, this means that player i's first-order belief α_i coincides with the optimal mixed strategy σ_j^* of the other player j and that player i's second-order belief β_i coincides with the first-order belief α_j of the other player j, which, in turn, must be equal to σ_i^*. Second, at the equilibrium strategy profile $\sigma^* = (\sigma_1^*, \sigma_2^*)$, players maximize expected utilities given their beliefs, that is, for all $i \in \{1,2\}$ and all $\sigma_i \in \Sigma_i$, $U_i(\sigma_i^* \mid \alpha_i, \beta_i) \geq U_i(\sigma_i \mid \alpha_i, \beta_i)$. Combining the conditions we can say that the strategy profile σ^* is an equilibrium of the psychological prisoner's dilemma game if for all players $i \in \{1,2\}$ and all strategies $\sigma_i \in \Sigma_i$, $U_i(\sigma_i^* \mid \sigma_j^*, \sigma_i^*) \geq U_i(\sigma_i \mid \sigma_j^*, \sigma_i^*)$.

We find that the psychological prisoner's dilemma game exhibits the following equilibrium structure. First, for all $\theta \geq 0$, it is an equilibrium that both players defect. While this is the unique equilibrium with purely selfish players, additional equilibria might emerge in the psychological prisoner's dilemma game when players feel guilt. In fact, if $\theta \geq \overline{\theta} \equiv \frac{c-a}{a-b}$, then there are two additional equilibria: one equilibrium in pure strategies in which both players cooperate and another equilibrium in mixed strategies.[1,2]

[1] There exist parameter configurations for which two symmetric mixed Nash equilibria may exist in addition to the defective equilibrium. Fig. 3 in the appendix provides an example of such a parameter configuration (for which $a + d < b + c$ is a necessary condition). Since we do not use such configurations in our experiment, we abstain from a further specification of these in Proposition 1.

[2] The occurrence of a mixed strategy equilibrium is not unique to the presence of simple guilt, as the same feature can be obtained with other standard extensions of the assumption that individuals are purely materialistic as well.

Table 1
Parameter configurations used in the experiment, with payoffs expressed in ECUs.

Configuration	*a*	*b*	*c*	*d*
PD1	10	1	12	6
PD2	10	3	12	6
PD3	10	1	14	6

Proposition 1. *The equilibrium structure of the psychological prisoner's dilemma game is as follows:*

(a) For all $\theta \geq 0$, the strategy profile $s^* = (D, D)$ is an equilibrium in pure strategies.
(b) For all $\theta \geq \overline{\theta}$, the strategy profile $s^* = (C, C)$ is an equilibrium in pure strategies.
(c) For all $\theta \geq \overline{\theta}$, the strategy profile where both players cooperate with probability

$$\sigma^* = \frac{-[a+d-b-c+\theta(c-d)] + \sqrt{[a+d-b-c+\theta(c-d)]^2 + 4\theta(a+d-b-c)(d-b)}}{2\theta(a+d-b-c)}$$

is the unique symmetric equilibrium in mixed strategies.
(d) There are no asymmetric equilibria.

Proof. See the Appendix. ■

3. Laboratory experiment

Since the prevalence and the intensity of guilt is not guaranteed to be insensitive to minor changes in context or incentives (Bellemare et al., 2018; Dufwenberg et al., 2011; Khalmetski, 2016), we consider the prisoner's dilemma using the three different parameter configurations as presented in Table 1. All three parameter configurations satisfy the assumptions $c > a > d > b$ and $a + d > b + c$ that we imposed in our theoretical analysis. Moreover, $2a > b + c > 2d$, such that the three variations are consistent in terms of efficiency ranking over outcomes. Relative to the PD1 configuration, it is less risky for the players to cooperate in the PD2 configuration, in the sense that the sucker payoff that is obtained in case the opponent did not cooperate is less detrimental for her payoff. In the PD3 configuration, players are more tempted to defect, relative to the PD1 configuration, when they believe the opponent will cooperate.

3.1. Design and procedures

In our experiment, we elicit via one decision screen for each player: (1) her action choice, (2) her belief about the opponent cooperating, and (3) her belief about the opponent's belief about her own cooperation decision.[3] We opted for this procedure because it is more consistent with the equilibrium notion, according to which beliefs and actions form simultaneously, than a sequential approach in which one asks first about actions and afterwards, on a different computer screen, about beliefs. The game was neutrally framed by avoiding the labels "cooperation" and "defection" and using the labels "Action X" and "Action Y" instead. The first-order and second-order beliefs were elicited in an incentive compatible way. For first-order beliefs, that concern beliefs over binary decisions, we use the Quadratic Scoring Rule (QSR; see Offerman et al., 2009); for second-order beliefs, that concern beliefs over a continuum of possible first-order beliefs, we apply the Interval Scoring Rule (ISR; see Schlag & van der Weele, 2009).

To elicit first-order beliefs (henceforth, denoted by *FOB*), we ask how likely a subject regards the event that the other player will choose the cooperative action (Action X). To answer this question, subjects are provided with a slider that contains as grid points all numbers from 0 up to 100 and a triangular pointer that can be moved over the grid. The extreme values 0 and 100 correspond to the extreme beliefs "totally unlikely" and "totally likely" respectively. The answer $z \in [0, 1]$ yields a payoff of 10 ECU with probability $2z - z^2$ in case the opponent indeed chooses the cooperative action and with probability $1 - z^2$ in case the opponent defects. While moving the triangular pointer over the grid, the percentages in each of the two potential cases are displayed on screen in real time so that participants are at any time aware of the consequences of their choices.

For second-order beliefs (henceforth, *SOB*), the same type of slider is used, but instead of one value, two values x and y have to be chosen. These two values indicate the lower- and upper-bound of the interval that participants believe to contain the value z chosen by their opponent when asked about her first-order belief. In case the value z indeed happens to be contained in the interval $[x, y]$, the participant gets a payoff of 10 ECU with probability $(1 - (y - x))^2$ and nothing for sure in case the value z is outside the interval $[x, y]$. Hence, the probability to receive the 10 ECU in case of a correct guess is decreasing in the length of the chosen interval.

Note that, like e.g. Dufwenberg and Gneezy (2000), Charness and Dufwenberg (2006), Vanberg (2008), Peeters et al. (2015), and Danilov et al. (2019), we use self-reported first- and second-order beliefs in order to investigate guilt in the prisoner's dilemma.

[3] A screenshot of the description of the game as displayed throughout the experiment can be found in Figure B.1 of the online appendix. Screenshots for the decisions (1)–(3) can be found in Figure B.2. Results are disclosed to the participants on a screen as in Figure B.3 and Figure B.4.

As argued by Bellemare et al. (2017) using self-reported second-order beliefs leaves these beliefs more 'endogenous' in comparison to alternative methods where second-order beliefs are induced either directly by communicating the self-reported first-order beliefs of the other player (cf. Ellingsen et al., 2010) or via a strategy method where action choices are made for any possible first-order belief the other player may hold (cf. Bellemare et al., 2018; Dhami et al., 2019; Khalmetski et al., 2015). By inducing beliefs, a signal about the other player's thoughts about how to play the game are communicated. Observed cooperative behavior may then be less unconditional than what can be obtained by self-reported beliefs where no signal about the other player's thoughts are provided. Interestingly in this regard, Danilov et al. (2019) show that "if the agent is a norm complier but not guilt averse then under the uncertainty about the social norm she reacts to disclosed expectations of others affected by her choice even though she does not care about these other's expectations *per se*". Further, we elicit beliefs about the other player's behavior rather than about population behavior (as is done in e.g. Ridinger & McBride, 2016), because, conceptually, it agrees more to our interpretation of the psychological game.

Finally, for each subject, one of the three decisions was independently chosen for actual payment, with ECUs being exchanged in Euros on a one-to-one basis. The feedback screen revealed the decisions of both participants in a pair, the payoff-relevant decision, and the final payoff in Euros. Subjects knew from the beginning that feedback about actions and beliefs will be provided at the end of the experiment. Before the results screen was presented, we asked the participants for the least amount of compensation (in ECU) for which they are willing to switch to the other action. The revealed values can be interpreted as the participants' willingness to pay to avoid feeling guilty, as estimated in Bellemare et al. (2011). In order to avoid deception, we did not implement an actual switch of action, and accordingly did not provide incentives for a truthful revelation (e.g. by means of a BDM mechanism). It was made explicit to the participants that this question was hypothetical (see Figure B.3 in the online appendix).

In the post-experimental questionnaire, we elicit information on the participants' gender, risk attitude, and propensity to experience guilt. The participants' risk attitudes are elicited, as suggested in Dohmen et al. (2011), by asking them to answer the question "How do you see yourself: Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" by ticking a box on a scale from 0 to 10, where the value 0 means "not at all willing to take risks" and the value 10 means "very willing to take risks". To elicit their propensity towards the self-conscious feelings of guilt, we use the Guilt and Shame Proneness Scale (GASP) developed by Cohen et al. (2011). The GASP contains two guilt subscales that assess negative behavior-evaluations (NBEs) and repair action tendencies following private transgressions. The former subscale captures feeling bad about how one acted; the latter captures action tendencies (i.e., behavior or behavioral intentions) focused on correcting or compensating for transgression (such as for having violated a social norm). We consider the Guilt-NBE subscale most relevant in the context of the present situation. For this subscale participants have to answer the following four questions: (1) "After realizing you have received too much change at a store, you decide to keep it because the salesclerk doesn't notice. What is the likelihood that you would feel uncomfortable about keeping the money?", (2) "You secretly commit a felony. What is the likelihood that you would feel remorse about breaking the law?", (3) "At a coworker's housewarming party, you spill red wine on their new cream-colored carpet. You cover the stain with a chair so that nobody notices your mess. What is the likelihood that you would feel that the way you acted was pathetic?", and (4) "You lie to people but they never find out about it. What is the likelihood that you would feel terrible about the lies you told?". Answers are given on a 7-point Likert scale, where the value 1 means "very unlikely" and the value 7 means "very likely", and their final score on this subscale is the average response given.

The experiments were conducted in the experimental laboratory at Maastricht University in March 2017. We recruited undergraduate students from various disciplines via ORSEE (Greiner, 2015). Participants operated in one of three possible payoff configuration (PD1, PD2 or PD3). All interactions took place anonymously via computer clients that were connected to a central server. The experiments were programmed in z-Tree (Fischbacher, 2007). In total 278 students participated in the experiment: 90 in PD1, 92 in PD2 and 96 in PD3.[4] Instructions are provided in Section A of the online appendix.[5]

3.2. Hypotheses

For each of the three variations of the prisoner's dilemma, Fig. 1 depicts the set of all symmetric equilibria as a function of the guilt parameter θ. It can be observed that in the mixed strategy equilibrium, lower cooperation rates go together with more guilt. This may a priori be counter-intuitive, but has a relatively simple explanation. The expected utility from defecting

$$U_i(D \mid \alpha_i, \beta_i) \;=\; \alpha_i\, c + (1-\alpha_i)\, d - \theta \cdot \beta_i\, [\, \alpha_i\, (a-b) + (1-\alpha_i)\,(c-d)\,]$$

depends on $\theta \cdot \beta_i$, which shows that second-order beliefs and guilt intensity are substitutes. Then, since a higher cooperation rate implies higher equilibrium second-order beliefs, more cooperation reduces the guilt parameter in this equilibrium.

[4] These sample sizes were aimed for based on the sample sizes used in Peeters et al. (2015). A typical session lasted about 40 min and the average payoff was about 10.28 Euros.

[5] All experiments were conducted with the informed consent of healthy adult subjects who were free to withdraw from participation at any time. Only individuals who voluntarily entered the experiment recruiting database were invited, and informed consent was indicated by electronic acceptance of an invitation to attend an experimental session. The experiments were conducted following the peer-approved procedures established by Maastricht University's Behavioral and Experimental Economics Laboratory (BEElab). Our study was approved by the BEElab at a public ethics review and project proposal meeting that is mandatory for all scholars wishing to use the BEElab facilities.

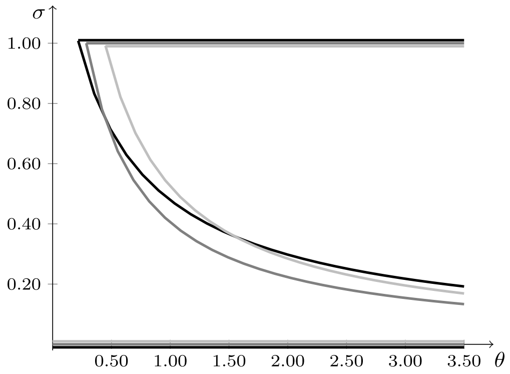

Fig. 1. The set of symmetric equilibria as a function of the parameter θ for the parameter configurations used in the experiment (black: PD1, gray: PD2, lightgray: PD3).

Table 2
Summary statistics: means and standard deviations.

	PD1	PD2	PD3
Gender (1 = Male)	0.4333 (0.4983)	0.4565 (0.5008)	0.4583 (0.5009)
Risk attitude (0–10)	6.3222 (1.9593)	6.2065 (2.0410)	6.3438 (2.0917)
Guilt-NBE (1–7)	5.0000 (1.2196)	4.9592 (1.2136)	4.8776 (1.2242)
Cooperation	0.2667 (0.4447)	0.2391 (0.4289)	0.2292 (0.4225)
First-order belief	0.4057 (0.2213)	0.3987 (0.2886)	0.3338 (0.2582)
Second-order belief	0.3995 (0.2008)	0.4051 (0.2671)	0.3911 (0.2522)

Under the assumption that guilt is the same in all three games, treatment comparisons are directly obtained from Fig. 1. It can be observed that the black line, which corresponds to PD1, can be both above and below of both the gray line (PD2) and the lightgray line (PD3). That is, depending on the guilt parameter, the cooperation rate in PD1 can be higher or lower than those in PD2 and PD3. Also, the cooperation rate is higher in PD3 than in PD2, independently of θ.

Hypothesis 1. Given θ, there is always more cooperation in the unique mixed strategy equilibrium in PD3 than in PD2. Depending on θ, there is more or less cooperation in PD1 than in either of the other two games.

4. Results

In our data analysis, we proceed as follows. The summary statistics in Section 4.1 show that there are no important treatment effects, neither for the cooperation rate nor for the beliefs. This contradicts Hypothesis 1. Another point that contradicts the mixed strategy equilibrium interpretation is that beliefs are not consistent with the observed cooperation rates. In the next step of our analysis, we apply regression techniques to analyze guilt motives. Given a guilt parameter θ and given a first-order belief α_i, $U_i(D \mid \alpha_i, \beta_i)$ is decreasing in β_i, while the expected utility from cooperating $U_i(C \mid \alpha_i)$ is independent of β_i. Subjects with higher second-order beliefs should have thus more incentive to cooperate everything else fixed. Probit regressions show that the cooperation rate is indeed increasing in the beliefs (Section 4.2). In Section 4.3, we estimate the model of noisy introspection of Goeree and Holt (2004) with guilt as an alternative to the mixed strategy equilibrium hypothesis. We find that the guilt parameter plays in all three treatments a considerable role. Finally, we show that the economic guilt expressed through choices in the PD game does not relate with the psychological guilt obtained from the ex-post questionnaire (Section 4.4).

4.1. Summary statistics

Table 2 presents summary statistics on participant characteristics and their decisions in the experiment. Mann–Whitney tests do not indicate any significant differences in the participants' characteristics between the three games concerning gender, risk-attitude, and guilt as measured by Guilt-NBE in GASP ($p > .41$ in all cases).[6] This means that eventual differences in results across games should be attributed to the variation in the incentives provided by the game parameters (including θ) rather than potential subject pool biases.

For the between treatments comparisons of cooperation rates and beliefs Mann–Whitney tests are applied. The cooperation rates differ slightly between game variations, with 26.67% cooperation in PD1, 23.91% cooperation in PD2, and 22.92% cooperation in

[6] Reported p-values are two-sided throughout.

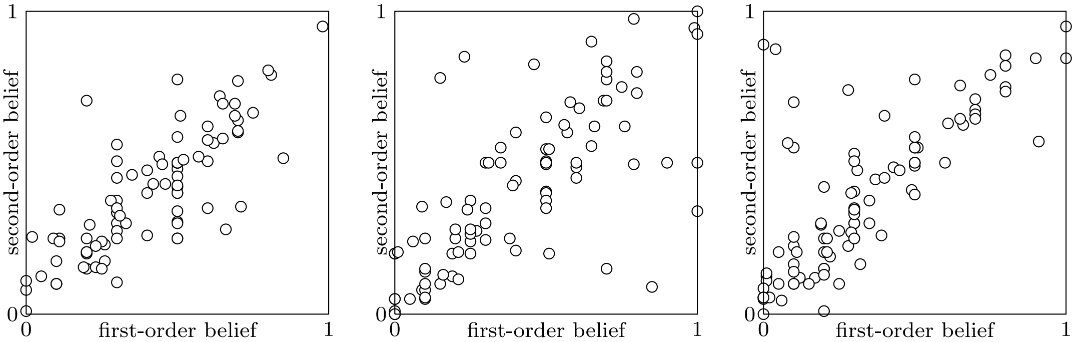

Fig. 2. Combination of elicited first- and second-order beliefs in subsequently PD1 (left), PD2 (middle) and PD3 (right).

PD3, but are not significantly different across the three games ($p > .55$ in all cases). The average first-order belief of 0.3987 in PD2 is not significantly different from those in the other two games ($p > .15$ in both cases), but the first-order belief of 0.4057 in PD1 is significantly larger than the 0.3338 in PD3 ($p = .0146$). Finally, there are no significant differences in the average midpoints of the reported second-order belief intervals across game variations ($p > .54$ in all cases). Hence, apart from first-order beliefs being different between PD1 and PD3, there are no significant differences across game variations.[7]

Comparing participants' choices within game variation, we find that average cooperation rates are substantially below the average first- and second-order beliefs, and these difference are significant according to Wilcoxon tests ($p < .005$ in both cases). Fig. 2 presents scatter plots of the combinations of first-order beliefs and the midpoints of second-order belief intervals for all subjects in the three different game configurations. For each game variation the two beliefs show a high level of correlation, with the correlation coefficients being 0.8013 for PD1, 0.7335 for PD2, and 0.7540 for PD3. Wilcoxon tests show that for PD1 and PD2, there are no significant differences between reported first- and second-order beliefs ($p > .23$ in both cases), but the midpoints of the reported second-order belief intervals are significantly above the first-order beliefs in PD3 ($p = .0038$). Overall, 229 of the 278 subjects (82.4%) reported a first-order belief that is within the reported interval for the second-order belief.

Result 1. *There are no substantial differences in choices and beliefs across game variations. Cooperation rates are lower than first- and second-order beliefs. First- and second-order beliefs correlate highly.*

Result 1 is at odds with mixed strategy equilibrium behavior. First, average beliefs are not consistent with the observed cooperation rates. And second, since the cooperation rate is the same in all three treatments, the mixed strategy equilibria of the three games in Fig. 1 should cross in the very same θ (which they do not). There are at least three explanations for this incongruity. One possibility is that the we are not able to identify treatment differences due to a lack of power. This cannot be ruled out because for a one-sided type I error of $\alpha = 0.05$ and a power of $1 - \beta = 0.8$, we can detect an effect size (difference in cooperation rates) of about 0.2. This known drawback of many laboratory experiments shows the importance of meta studies. Another possibility is that guilt is context-dependent, as has been recently reported in Bellemare et al. (2018). And, finally, subjects may not be fully rational. We address the two latter points in Section 4.3.

4.2. Beliefs and cooperation

In this subsection, we analyze whether the cooperation decision depends on the first- and second-order beliefs of the subjects. From the theoretical model we can see that

$$\frac{\partial U_i(C)}{\partial \alpha_i} - \frac{\partial U_i(D)}{\partial \alpha_i} = (1 + \theta \beta_i) \cdot (a + d - b - c) > 0,$$

which suggests a positive dependence of first-order beliefs on the rate of cooperation. Moreover, as we have already indicated before, for a fixed guilt parameter θ and a fixed first-order belief α_i,

$$\frac{\partial U_i(C)}{\partial \beta_i} - \frac{\partial U_i(D)}{\partial \beta_i} = 0 + \theta [\alpha_i (a - b) + (1 - \alpha_i)(c - d)] \geq 0.$$

Our model therefore also predicts a positive correlation between the cooperation decision and the second-order beliefs.

We perform probit regressions to test the above-mentioned hypotheses. Columns (1), (2), (4) and (5) in Table 3, with beliefs separately included, reveal that, consistent with our two hypotheses, the marginal effects of first- and second-order beliefs on cooperation rates are positive and significant in all game variations. However, when first- and second-order beliefs are jointly included as regressors, in Columns (3) and (6), we find that in PD1 and PD2 only the second-order beliefs are significantly correlated with cooperative behavior, while in PD3 only the first-order beliefs are significantly correlated.

[7] See Section C of the online appendix for differences in beliefs within and across game variations among the cooperators and the defectors.

Table 3
Probit regressions for the dependency of cooperation choices on first- and second-order beliefs. Marginals (eyex) are reported.

	Prisoner's dilemma 1					
	(1)	(2)	(3)	(4)	(5)	(6)
First-order belief	1.6339***		0.6517	1.6883***		0.6123
Second-order belief		2.3196***	1.8729**		2.4606***	2.0168**
Gender				−0.3124	−0.5351	−0.4905
Risk attitude				0.6157	0.5792	0.6116
Guilt-NBE				0.6670	0.4820	0.6069
Pseudo R-squared	0.2334	0.3138	0.3261	0.2632	0.3495	0.3591
	Prisoner's dilemma 2					
	(1)	(2)	(3)	(4)	(5)	(6)
First-order belief	1.3931***		0.8238	1.4110***		0.8497
Second-order belief		2.1631***	1.8534**		2.1112***	1.8384**
Gender				0.1342	0.0211	0.0589
Risk attitude				0.5608	0.4085	−0.0775
Guilt-NBE				1.5083	0.3441	0.6613
Pseudo R-squared	0.2934	0.4133	0.4478	0.3140	0.4152	0.4493
	Prisoner's dilemma 3					
	(1)	(2)	(3)	(4)	(5)	(6)
First-order belief	1.1713***		0.8599**	1.1848***		0.9540*
Second-order belief		1.3437***	0.6348		1.2822***	0.4310
Gender				−0.2961	−0.1934	−0.2329
Risk attitude				1.3278	1.2498	1.1889
Guilt-NBE				−1.0364	−0.5071	−0.8529
Pseudo R-squared	0.2786	0.2214	0.2952	0.3193	0.2537	0.3242

***$p < .001$.
**$p < .01$.
*$p < .05$.

Result 2. *Subjects with higher first- and second-order beliefs are more likely to cooperate. In PD1 and PD2 (where first- and second-order beliefs are not significantly different), second-order beliefs are more explanatory, while in PD3 (where second-order beliefs are larger than the first-order beliefs), first-order beliefs are more explanatory.*

4.3. Noisy introspection

Since it is well established that it is difficult for subjects participating in economic experiments to apply equilibrium arguments, or they may have at least doubts whether their co-players are capable of applying them, it is possible that a model of boundedly rational choice captures subjects' behavior better than the mixed strategy equilibrium. We focus here on the noisy introspection model of Goeree and Holt (2004) according to which the probability that a player cooperates is given by the expression

$$\sigma = \frac{\exp(U(C \mid \alpha)/\mu_0)}{\exp(U(C \mid \alpha)/\mu_0) + \exp(U(D \mid \alpha, \beta)/\mu_0)},$$

where $\mu_0 \geq 0$ is a "noise parameter" (to be estimated). The choice probability of one player naturally becomes the first-order belief of the other player, however with increasing mistakes. In particular,

$$\alpha = \frac{\exp(U(C \mid \beta)/\mu_1)}{\exp(U(C \mid \beta)/\mu_1) + \exp(U(D \mid \beta, \gamma)/\mu_1)},$$

where γ indicates the third-order belief and $\mu_1 = t \cdot \mu_0$. The parameter $t \geq 1$ (to be estimated) specifies the degree by which the noise increases. All higher-order beliefs are then generated by iterating on t, that is, for all $k \geq 2$, $\mu_k = t^k \cdot \mu_0$.

The model of noisy introspection not only provides flexibility to accommodate subjects' behavior, it also encompasses other well-known specifications of bounded rationality. For example, it reduces to the logit equilibrium of McKelvey and Palfrey (1995) when $t = 1$ (beliefs are then consistent with the choice probability), whereas a particular instance of level-k behavior is obtained if there is a k such that $\mu_{k'} = 0$ for all $k' \leq k$ and $\mu_{k'} = \infty$ for all $k' > k$. In difference with the literature on level-k behavior, which usually assigns subjects to levels on an individual basis and thereby obtains a non-degenerate distribution of levels, here all mass is necessarily assigned to a unique level.

In order to estimate the model, let $K \geq 2$. Take $p_{K-1}(K) = p_K(K) = 0.5$ and for all $k = K-2, \ldots, 0$, let

$$p_k(K) = \frac{\exp(U(C \mid p_{k+1})/\mu_k)}{\exp(U(C \mid p_{k+1})/\mu_k) + \exp(U(D \mid p_{k+1}, p_{k+2})/\mu_k)},$$

Table 4

Maximum likelihood estimations. Means and standard deviations are bootstrapped using 1000 samples of 60 observations. For models (1) and (3), fifteen iterations ($K = 15$) are used to solve for the cooperation rate and first- and second-order belief using the noisy introspection model; for model (2), fifty iterations ($K = 50$) are used, since the lower value of t decreases the speed of convergence of the iterative process. Using two-sample t-tests, we find that all parameter comparisons (between treatments for a given model and between the different models for a given treatment) are highly significant ($p < .001$ in all cases). The estimated cooperation rates and first- and second-order beliefs in the bottom of the table result from the noisy introspection model using the estimated means of the parameters (applying the same number of iterations as mentioned above).

Game	PD1			PD2			PD3		
Model	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
θ	0.8309	0.9166		0.6977	0.3841		0.5803	1.0467	
	(0.0400)	(0.0645)		(0.0348)	(0.2745)		(0.3875)	(0.0273)	
μ_0	1.3735	2.6918	4.0814	0.6194	2.8795	2.3821	2.4052	2.4418	3.9666
	(0.3029)	(0.2623)	(0.6879)	(0.1646)	(1.0427)	(0.3544)	(1.3188)	(0.1117)	(0.6169)
t	1.3806		1.8180	1.5318		2.1144	1.4783		1.6838
	(0.1151)		(0.2394)	(0.1619)		(0.3506)	(0.1795)		(0.1748)
$\widehat{\sigma}$	0.2792	0.3440	0.2794	0.2488	0.3485	0.2494	0.2389	0.3072	0.2357
$\widehat{\alpha}$	0.3798	0.3447	0.3779	0.3764	0.3485	0.3756	0.3295	0.3090	0.3343
$\widehat{\beta}$	0.4312	0.3456	0.4335	0.4359	0.3485	0.4409	0.3897	0.3110	0.4000

where $U(C \mid \alpha) = \alpha\, a + (1-\alpha)\, b$, $U(D \mid \alpha, \beta) = \alpha\, c + (1-\alpha)\, d - \theta \cdot \beta\, [\, \alpha\, (a-b) + (1-\alpha)\, (c-d)\,]$, and for all $k = 1, \ldots, K-2$, $\mu_k = t^k \mu_0$ with $\mu_0 > 0$ and $t \geq 1$. For $k \leq K-2$, p_k is the probability that a player with first-order belief p_{k+1}, second-order belief p_{k+2}, and guilt parameter θ cooperates when playing in accordance with the logit response model with noise μ_k. For each (θ, μ_0, t), this gives us the following outcome for the cooperation rate and first- and second-order beliefs:

$$\widehat{\sigma} = \lim_{K\to\infty} p_0(K), \qquad \widehat{\alpha} = \lim_{K\to\infty} p_1(K), \qquad \text{and} \qquad \widehat{\beta} = \lim_{K\to\infty} p_2(K).$$

For $\theta = 0$, this is precisely the model of noisy introspection of Goeree and Holt (2004) that allows (with $t > 1$) for the noise in the players' reasoning process to be increasing in the order of the belief reasoned about. What we add, via our specification of $U(D \mid \alpha, \beta)$, is to allow for players to be sensitive to feelings of guilt.

We maximize the log-likelihood function

$$\log\, L = \widetilde{\sigma}\log(\widehat{\sigma}) + (1-\widetilde{\sigma})\log(1-\widehat{\sigma}) + \widetilde{\alpha}\log(\widehat{\alpha}) + (1-\widetilde{\alpha})\log(1-\widehat{\alpha}) + \widetilde{\beta}\log(\widehat{\beta}) + (1-\widetilde{\beta})\log(1-\widehat{\beta}),$$

where $\widetilde{\sigma}$, $\widetilde{\alpha}$ and $\widetilde{\beta}$ refer to the empirically observed cooperation rates and first- and second-order beliefs. The typical approach when estimating (θ, μ_0, t) is by concentrating only on actions. Since we have extracted first- and second-order beliefs in an incentive compatible way, it is however natural to maximize the likelihood under the complete string of observed data instead of restricting our attention to only actions. Table 4 presents the results. In column (1), we jointly estimate $\theta \geq 0$, $\mu_0 > 0$ and $t \geq 1$. Columns (2) and (3) are restricted models. We set $t = 1$ in column (2), which corresponds to the logit version of the quantal response equilibrium. In column (3), $\theta = 0$ in order highlight the impact of guilt.

First, we obtain from Table 4 the following insights regarding the different models for a given game configuration—that is, when comparing the different columns for PD1, the different columns for PD2, and the different columns for PD3. (*i*) Most importantly, the estimates of θ in models (1) is substantially bounded away from 0, providing some evidence for the presence of guilt.[8] Moreover, imposing $\theta = 0$ has in all three treatments the consequence that the estimates for μ_0 and t increase significantly. Hence, more noise (bounded rationality) is needed to explain the data when θ is assumed away. (*ii*) The estimate for t in models (1) is in all three treatments larger than 1, indicating that noise increases in the beliefs. Further, fixing $t = 1$ causes the estimates for θ and μ_0 to increase in PD1 and PD3. This is different for PD2, where the estimate for θ decreases in exchange for a substantial increase in μ_0.

Second, regarding treatment comparisons, and concentrating on the main model (1), the estimated θ is largest in PD1 (0.8309), second largest in PD2 (0.6977), and smallest in PD3 (0.5803). The numeric estimates are not too different across games, but since treatment differences are significant, one has nevertheless to be careful when assuming guilt to be context-independent. It is finally worth noting that the relative ordering of game variations is different for all three parameters. While the estimated μ_0 is largest in PD3, second largest in PD1, and smallest in PD2, the estimated t is largest in PD2, second largest in PD3, and smallest in PD1. It therefore seems that θ, μ_0, and t are (imperfect) substitutes of each other.[9]

Finally, with respect to the estimated cooperation rate $\widehat{\sigma}$ and the estimated first- and second-order beliefs $\widehat{\alpha}$ and $\widehat{\beta}$, it is straightforward that the models (2) cannot provide a good fit of the experimental data as $t = 1$ implies that beliefs are consistent with the cooperation rate. And our previous discussions emphasized that the observed average first- and second-order beliefs are higher than the observed cooperation rates. Another experimental finding was that first- and second-order beliefs are consistent. However $\mu_0 > 0$ and $t > 1$ naturally leads to $\alpha \neq \beta$, so that the noisy introspection model fails to incorporate that feature. The

[8] Out of the 1000 samples, a value of θ equal to zero was never returned for PD1 and PD2, and 123 times for PD3.

[9] Related to our Hypothesis 1, Patel and Smith (2019) estimate subjects' sensitivity to guilt by mirroring average population choices against the (completely) mixed strategy equilibrium. The average cooperation rates of 0.2667 in PD1, 0.2391 in PD2, and 0.2292 in PD3 are consistent with the mixed strategy equilibrium interpretation with guilt parameters 2.32 in PD1, 1.85 in PD2 and 2.52 in PD3, respectively.

estimated cooperation rates and the estimated beliefs differ only slightly between models (1) and (3), the main difference here is, as we already indicated before, that the inclusion of θ reduces the estimated values of μ_0 and t.

Result 3. *The estimates of the guilt parameter θ are bounded away from 0 and vary slightly but significantly between treatments.*

4.4. Economic versus psychological guilt

Compared to economists, (social) psychologists have a longer tradition in studying guilt and adopt a slightly broader definition by referring to it as an emotional state associated with the violation of an intrinsic moral standard, which is not necessarily related to expectations/beliefs of others. The most widely used tool to measure proneness to guilt is the Test of Self-Conscious Affect-3 (TOSCA-3) by Tangey et al. (2000). Bellemare et al. (2019) finds that guilt elicited via TOSCA-3 correlates highly with guilt sensitivity measured within a framework of psychological game theory in the context of binary trust and dictator games. Although TOSCA-3 asks respondents for their reactions to hypothetical real-life situations that concern both evaluative and behavioral responses to transgressions, the test does not differentiate between them. The Guilt and Shame Proneness scale (GASP) by Cohen et al. (2011), however, allows distinguishing negative behavior-evaluations (Guilt-NBE) from action oriented repair responses. Bracht and Regner (2013) finds that the Guilt-NBE component of GASP is predictive for pro-social behavior in the context of a binary trust game.

In Section 4.2, we have already seen that guilt as measured by the Guilt-NBE component of GASP is not predictive for cooperative behavior in the context of the prisoner's dilemma. In this subsection we investigate whether Guilt-NBE correlates with guilt when measured at the individual level. Where Bracht and Regner (2013) elicit Guilt-NBE one week before the experimental session running the trust game, we elicit it after subjects received feedback on the outcome of the prisoner's dilemma game. Despite this crucial difference, the distribution of subjects' responses to the four Guilt-NBE questions are quite similar.[10] Also, we do not find a significant difference in how cooperators and defectors answer the questions (Mann–Whitney: $p = .5772$).

Instead of eliciting second-order beliefs directly, Bellemare et al. (2019) asked subjects to make decisions for various potential first-order beliefs of the other player. This allows them to estimate lower- and upper-bounds on the guilt-sensitivity parameter on the individual level. To estimate guilt-sensitivity on the individual level in our experiment, we make use of the answers to the hypothetical BDM question. In the hypothetical BDM, subjects indicated the amount B_i they would need to receive so that they are indifferent between cooperation and defecting. For cooperators this gives us the equation

$$U_i(C \mid \alpha_i) = U_i(D \mid \alpha_i, \beta_i) + B_i,$$

from which we obtain

$$\theta_i = \frac{[\,\alpha_i\,(c-a) + (1-\alpha_i)\,(d-b)\,] + B_i}{\beta_i\,[\,\alpha_i\,(a-b) + (1-\alpha_i)\,(c-d)\,]}.$$

For defectors this gives us the equation

$$U_i(D \mid \alpha_i, \beta_i) = U_i(C \mid \alpha_i) + B_i,$$

from which we obtain

$$\theta_i = \frac{[\,\alpha_i\,(c-a) + (1-\alpha_i)\,(d-b)\,] - B_i}{\beta_i\,[\,\alpha_i\,(a-b) + (1-\alpha_i)\,(c-d)\,]}.$$

Using the reported values for α_i, β_i and B_i, we obtain for each subject their individual estimate of the guilt sensitivity parameter θ_i. The average (standard deviation) θ_i for cooperators is 3.10 (1.39) in PD1, 2.14 (0.69) in PD2 and 3.52 (3.47) in PD3.[11] Regarding the across game variation comparison, we find with the help of Mann–Whitney tests that this estimated guilt level is significantly smaller in PD2 than in PD1 and PD3 ($p = .0040$ and $p = .0290$) and that is no significant difference between PD3 and PD1 ($p = .6285$).

Pearson's correlation tests reveal that there is no significant relation between individual guilt levels and average responses to the Guilt-NBE questions ($p = .7871$), neither for any of the individual questions ($p > .13$).[12] We can image five reasons for why, unlike Bellemare et al. (2019), we do not find a significant correlation. First, the questions postulated in the guilt questionnaire are more oriented to social norms in general, rather than taking into account the beliefs others may have about one's behavior in the framed circumstances – a small difference in how guilt is conceptualized by economists and psychologists. Second, we elicited participants' first- and second-order beliefs simultaneously and on the individual level rather than, as is perhaps more common, on the population level.[13] Third, related to Footnote 4.4, there might be an inaccuracy in the estimated guilt-sensitivity levels.

[10] Our subjects did not respond to the four Guilt-NBE questions of GASP only, but to all 16 GASP questions. Notably, also the distributions of the answers to the remaining questions related to guilt-repair and shame look quite similar.

[11] The reason to focus on the cooperators is that for four defectors the guilt parameter cannot be determined as they report a second-order belief of zero, and that for 193 of the remaining 206 defectors we find a negative guilt parameter. The latter is caused by subjects providing (sometimes unreasonably) high bids in the BDM.

[12] Participants' answers to the four questions positively correlate highly significantly in pairwise comparisons ($p < .006$).

[13] The reason is that if we would elicit first-order beliefs on the population level, then for the second-order beliefs we would have to ask participants either (1) about their co-player's belief about the population, or (2) about the population's belief about the population. In either case, the second-order belief elicited does not apply to own behavior, which we consider crucial when studying the concept of simple guilt.

The estimated guilt-levels are non-negative for the cooperators, however many subjects reported high hypothetical bids, such that care needs to be taken when interpreting these and their transformation into θ. Fourth, players make decisions simultaneously in our setting. Fifth, and related to the first and fourth reason, psychological and economic guilt might simply respond differently in normatively different contexts. Regarding the prevalence of these five reasons, it seems plausible that a combination of these is causing our insignificance result. In fact, it is unlikely that the third reason is the only one. While Guilt-NBE is predictive for trustworthiness (Bracht & Regner, 2013), it can be seen from Table 3 that it is not for cooperative behavior.

5. Concluding discussion

In this paper, we have shown theoretically that cooperation in the prisoner's dilemma game can be sustained in equilibrium if players are guilt averse. While defection always remains a pure strategy equilibrium of the psychological game induced by guilt aversion, both a pure strategy equilibrium in which players cooperate and a mixed strategy equilibrium appear whenever players are sufficiently guilt averse (Proposition 1). The results of our laboratory experiment do not support the treatment comparisons derived from the mixed strategy equilibrium in Hypothesis 1. While, for all three game specifications, first- and second-order beliefs are highly correlated (Result 1) and the action depends on these beliefs in the way suggested (Result 2), the cooperation rates do not differ between treatments. The latter contradicts the assumption that guilt is context-independent. A second violation of the mixed strategy equilibrium is that the observed cooperation rates are not consistent with the stated first- and second-order beliefs. To explore whether actions are partly driven by boundedly rational behavior, we then estimate the noisy introspection model of Goeree and Holt (2004) including guilt for our data. We believe this to be the main innovation of this paper. There is evidence of guilt aversion in all three treatments because the estimates of the guilt parameter are sufficiently bounded away from zero. Moreover, the estimates of the two noise parameters are larger in the model without guilt than in the models with guilt, which hints at guilt being a partial substitute for these model parameters (Result 3).

As indicated in the Introduction, many other models are able to explain cooperation in the PD game, even the existence of a mixed strategy equilibrium. Consequentialistic models assume utilities to solely depend on the final payoffs; in particular, actions and beliefs do not enter the utility function directly. Models that fit this category include pure altruism (Andreoni, 1990), inequality aversion (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999), preferences for efficiency (Engelmann & Strobel, 2004), or, more generally, distributional preferences (Charness & Rabin, 2002). To elaborate, in the framework of Bolton and Ockenfels (2000), where equal disutility is assumed for proportional payoffs deviating from 0.5 on either side, cooperation can occur in equilibrium (either as a pure or as result of a mixed strategy) when this disutility is large enough. Consideration of the framework of Fehr and Schmidt (1999), where the disutility for unequal payoffs differs between being on the high or the low end, learns that players can cooperate in equilibrium if the players' disutility for receiving a higher payoff than the opponent is large enough, irrespective of the disutility that players may perceive for receiving a lower payoff. In both these models, there is no particular role of first- or second-order beliefs, apart from these being consistent with choices in equilibrium. In the framework of Charness and Rabin (2002), cooperation can result in equilibrium if the combination of the Rawlsian and the efficiency factor are sufficiently salient in the players' utility function.

In procedural models, agents value other aspects of the outcome than the final payoffs, such as the justness of actions taken and the (imputed) intentions. Well-known models in this category include the models by Rabin (1993), Dufwenberg and Kirchsteiger (2004), Cox et al. (2007), and López-Pérez (2008). For instance, in the specification of Rabin (1993) the own 'kindness' towards the other player and the 'expected kindness' of the other towards the self enter the player's utility function, where kindness is a function of a player's action and first-order belief and the expected kindness is a function of a player's first- and second-order beliefs. While for the particular specification of the expected kindness functions proposed in the appendix of Rabin (1993) the expected kindness is a function of the first-order belief only, in general, this framework can explain cooperation choices to be dependent on second-order beliefs. López-Pérez (2008) studies extensive-form games assuming that there is an ex-ante norm, on which all agents agree, of how players should behave in each moment they are called to take an action. Agents have an aversion to breaking this norm, that is, an aversion to deviating from the pre-specified game path. Incorporating this idea into the simultaneous-move PD game and under the assumption that there is a cooperation norm, agents only get a disutility if they are unilateral defectors. Observe that our direct application of simple guilt implies that agents get a disutility not only when they are unilateral defectors, but whenever they defect.

One can conclude from our theoretical and experimental analysis that guilt aversion is another reason why people may choose to cooperate in the prisoner's dilemma game. One way to analyze in the future the prevalence of the different drivers of cooperation would be by estimating a model of noisy introspection for different motives and compare their explanatory power.

Appendix. Proofs

Proof of Proposition 1

(a) To see that the strategy profile $s^* = (D, D)$ is an equilibrium in pure strategies for all $\theta \geq 0$ simply note that $U_i(D \mid 0, 0) = d > b = U_i(C \mid 0, 0)$.

(b) We show that $s^* = (C, C)$ is an equilibrium in pure strategies whenever $\theta \geq \overline{\theta}$. Each player gets $U_i(C \mid 1, 1) = a$ from cooperating. If a player deviates and defects, her payoff is $U_i(D \mid 1, 1) = c - \theta(a - b)$. Hence, $U_i(C \mid 1, 1) \geq U_i(D \mid 1, 1)$ as long as $a \geq c - \theta(a - b)$. This equation solves for $\theta \geq \frac{c-a}{a-b}$.

(c) A $\sigma \in (0,1)$ constitutes a symmetric equilibrium in mixed strategies if and only if $U_i(C \mid \sigma,\sigma) = U_i(D \mid \sigma,\sigma)$. That is,

$$\sigma\, a + (1-\sigma)\, b = \sigma\, c + (1-\sigma)\, d - \theta \cdot \sigma\, [\, \sigma\, (a-b) + (1-\sigma)\, (c-d)\,].$$

We see that for $\theta = 0$ this renders a solution that cannot be an equilibrium: $\sigma = \frac{d-b}{a+d-b-c} > 1$. We assume henceforth that $\theta > 0$. Rewriting the previous equation we obtain that

$$\theta\, (a+d-b-c)\, \sigma^2 + [\, a+d-b-c+\theta\, (c-d)\,]\, \sigma - (d-b) = 0.$$

Consequently, the two possible solutions to this quadratic equation are

$$\sigma^*_{1,2} = \frac{-\,[\, a+d-b-c+\theta\, (c-d)\,] \pm \sqrt{[\, a+d-b-c+\theta\, (c-d)\,]^2 + 4\,\theta\, (a+d-b-c)\, (d-b)}}{2\,\theta\, (a+d-b-c)}.$$

From $a+d-b-c > 0$ and $c > a > d > b$, we can conclude that both solutions are real and that the smallest solution is negative and the largest solution positive. Hence, the smallest solution cannot be an equilibrium. For the largest solution to be an equilibrium, we have to show that its value is less than or equal to 1. That is, we have to show that

$$\sqrt{[\, a+d-b-c+\theta\, (c-d)\,]^2 + 4\,\theta\, (a+d-b-c)\, (d-b)} \le [\, a+d-b-c+\theta\, (c-d)\,] + 2\,\theta\, (a+d-b-c).$$

Since the expressions on both sides of this inequality are positive, this inequality is satisfied if and only if

$$4\,\theta\, (a+d-b-c)\, (d-b) \le 2\, [\, a+d-b-c+\theta\, (c-d)\,]\, 2\,\theta\, (a+d-b-c) + 4\,\theta^2\, (a+d-b-c)^2,$$

or

$$d-b \le [\, a+d-b-c+\theta\, (c-d)\,] + \theta\, (a+d-b-c).$$

This inequality holds if and only if $\theta \ge \frac{c-a}{a-b}$.

(d) A strategy profile where one player plays D while the other plays C with positive probability (with beliefs being consistent with this play) cannot be an equilibrium, since D is the unique best-response to D. Moreover, a strategy profile where one player plays C while the other plays D with positive probability (with beliefs being consistent with this play) can also not be an equilibrium. First, C is the unique best-response to C if $\theta \ge \overline{\theta}$. Second, for $\theta < \overline{\theta}$, while D is the unique best-response to C, C is not a best-response to D in return. Therefore, the only possibility to have asymmetric equilibria, is them to be in completely mixed strategies.

Suppose player j cooperates with probability σ_j, and player i has beliefs $\alpha_i = \sigma_j$ and β_i. Player i is indifferent between playing C and D if and only if

$$U(C \mid \sigma_j, \beta_i) \;=\; U(D \mid \sigma_j, \beta_i).$$

From this we find that player j leaves player i indifferent between these two actions by choosing[14]

$$\sigma_j \;=\; 1 - \frac{(a-c) + \theta\, \beta_i\, (a-b)}{(1+\theta\, \beta_i)\, (a+d-b-c)}.$$

Similarly, we find that player i leaves player j indifferent between C and D by playing

$$\sigma_i \;=\; 1 - \frac{(a-c) + \theta\, \beta_j\, (a-b)}{(1+\theta\, \beta_j)\, (a+d-b-c)}.$$

Equilibrium conditions require $\beta_i = \sigma_i$ and $\beta_j = \sigma_j$, such that we obtain the system of equations

$$\sigma_i \;=\; 1 - \frac{(a-c) + \theta\, \sigma_j\, (a-b)}{(1+\theta\, \sigma_j)\, (a+d-b-c)} \tag{1}$$

and

$$\sigma_j \;=\; 1 - \frac{(a-c) + \theta\, \sigma_i\, (a-b)}{(1+\theta\, \sigma_i)\, (a+d-b-c)} \tag{2}$$

to be satisfied in an equilibrium. Inverting Eq. (1), we obtain

$$\sigma_j \;=\; \frac{(d-b) - (a+d-b-c)\, \sigma_i}{\theta\, (c-d) + \theta\, (a+d-b-c)\, \sigma_i}. \tag{3}$$

The derivatives of the right hand-sides of Eqs. (2) and (3) to σ_i are

$$-\frac{\theta\, (c-b)\, (a+d-b-c)}{[\, (a+d-b-c) + \theta\, (a+d-b-c)\, \sigma_i\,]^2}$$

and

$$-\frac{\theta\, (c-b)\, (a+d-b-c)}{[\, \theta\, (c-d) + \theta\, (a+d-b-c)\, \sigma_i\,]^2},$$

[14] Note that we ignore, for the moment, the possibility for this solution to be outside the interval $(0,1)$.

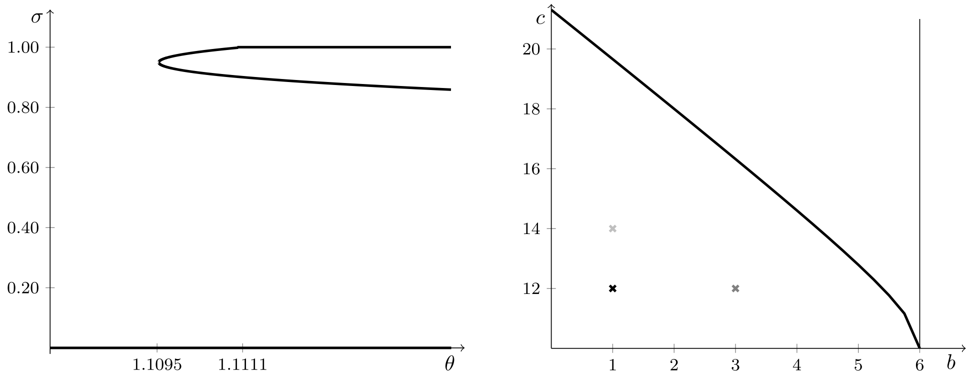

Fig. 3. The left graph plots the set of symmetric equilibria for the situation $(a, b, c, d) = (10, 1, 20, 6)$ where $a + d < b + c$ and for which two mixed strategy equilibria exist for values of θ in $(1.1095, 1.1111)$. For $a = 10$ and $b = 6$ (as in all three variations used in the experiment), for all pairs (b, c) below the curve in the right graph there does not exist a θ for which there are two mixed strategy equilibria. The three crosses mark the pairs used in the experiment (black: PD1, gray: PD2, lightgray: PD3).

respectively. From $c > b$ and $a+d > b+c$, it follows that both these derivative are negative, implying that both right hand-sides are downward sloping. Moreover, we see that the only difference between the slopes are the terms that are constant with respect to σ_i in the denominator. Since all terms in the derivative are positive, we find that, one of the curves is steeper than the other, at all $\sigma_i > 0$. This means that the two curves can cross at most once on the positive domain, implying that we can have at most one (feasible) solution to Eqs. (1) and (2) on the positive domain, and hence at most one mixed strategy equilibrium. By symmetry of the game, asymmetric equilibria always come in pairs; that is, if (σ', σ'') is an equilibrium, then also (σ'', σ') is an equilibrium. Hence, the only possible equilibrium is symmetric, which is the equilibrium identified in part (c).

Multiple mixed equilibria

Please see Fig. 3.

Appendix. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.joep.2020.102347.

References

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

Battigalli, P., Charness, G., & Dufwenberg, M. (2013). Deception: The roles of guilt. *Journal of Economic Behaviour and Organization*, *93*, 227–232.

Battigalli, P., & Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, *97*(2), 170–176.

Battigalli, P., & Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, *144*(1), 1–35.

Baumeister, R., Stillwell, A., & Heatherton, T. (1994). Guilt: An interpersonal approach. *Psychological Bulletin*, *115*(2), 243–267.

Bellemare, C., Sebald, A., & Strobel, M. (2011). Measuring the willingness to pay to avoid guilt: Estimation using equilibrium and stated belief models. *Journal of Applied Econometrics*, *26*(3), 437–453.

Bellemare, C., Sebald, A., & Suetens, S. (2017). A note on testing guilt aversion. *Games and Economic Behavior*, *102*, 233–239.

Bellemare, C., Sebald, A., & Suetens, S. (2018). Heterogeneous guilt sensitivities and incentive effects. *Experimental Economics*, *21*(2), 316–336.

Bellemare, C., Sebald, A., & Suetens, S. (2019). Guilt aversion in economics and psychology. *Journal of Economic Psychology*, *73*, 52–59.

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

Bracht, J., & Regner, T. (2013). Moral emotions and partnerships. *Journal of Economic Psychology*, *39*, 313–326.

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

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

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, *14*(1), 47–83.

Cohen, T., Wolf, S., Panter, A., & Insko, C. (2011). Introducing the GASP scale: A new measure of guilt and shame proneness. *Journal of Personality and Social Psychology*, *100*(5), 947–966.

Cox, J., Friedman, D., & Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, *59*(1), 17–45.

Danilov, A., Khalmetski, K., & Sliwka, D. (2019). Descriptive norms and guilt aversion. Working paper.

Dhami, S., Wei, M., & al Nowaihi, A. (2019). Public goods games and psychological utility: Theory and evidence. *Journal of Economic Behaviour and Organization*, *167*, 361–390.

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. (2011). Individual risk attitudes: Measurement, determinants and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550.

Dufwenberg, M., Gächter, S., & Hennig-Schmidt, H. (2011). The framing of games and the psychology of play. *Games and Economic Behavior*, *73*(2), 459–478.

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

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

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology*, *51*, 665–697.

Ellingsen, T., Johannesson, M., Tjøtta, S., & Torsvik, G. (2010). Testing guilt aversion. *Games and Economic Behavior*, *68*(1), 95–107.

Elster, J. (1998). Emotions in economic theory. *Journal of Economic Literature*, *36*(1), 47–74.

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

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

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

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

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*(1), 60–79.

Goeree, J. K., & Holt, C. A. (2004). A model of noisy introspection. *Games and Economic Behavior*, *46*, 365–382.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, *97*, 110–119.

Khalmetski, K., Ockenfels, A., & Werner, P. (2015). Surprising gifts: Theory and laboratory evidence. *Journal of Economic Theory*, *159*(A), 163–208.

López-Pérez, R. (2008). Aversion to norm-breaking: A model. *Games and Economic Behavior*, *64*, 237–267.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, *10*(1), 6–38.

Miettinen, T., & Suetens, S. (2008). Communication and guilt in a prisoner's dilemma. *Journal of Conflict Resolution*, *52*(6), 945–960.

Offerman, T., Sonnemans, J., van de Kuilen, G., & Wakker, P. (2009). A truth-serum for non-Bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, *76*(4), 1461–1489.

Patel, A., & Smith, A. (2019). Guilt and participation. *Journal of Economic Behaviour and Organization*, *167*, 279–295.

Peeters, R., Vorsatz, M., & Walzl, M. (2015). Beliefs and truth-telling: A laboratory experiment. *Journal of Economic Behaviour and Organization*, *113*, 1–12.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *83*(5), 1281–1302.

Ridinger, G., & McBride, M. (2016). Theory of mind ability and cooperation in the prisoners dilemma. Working paper.

Schlag, K., & van der Weele, J. (2009). Efficient interval scoring rules. Working paper.

Tangey, J. P., Dearing, R. L., Wagner, P. E., & Gramzow, R. (2000). *The test of self-conscious affect-3 (TOSCA-3)*. Fairfax: George Mason University.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica*, *76*(6), 1476–1480.