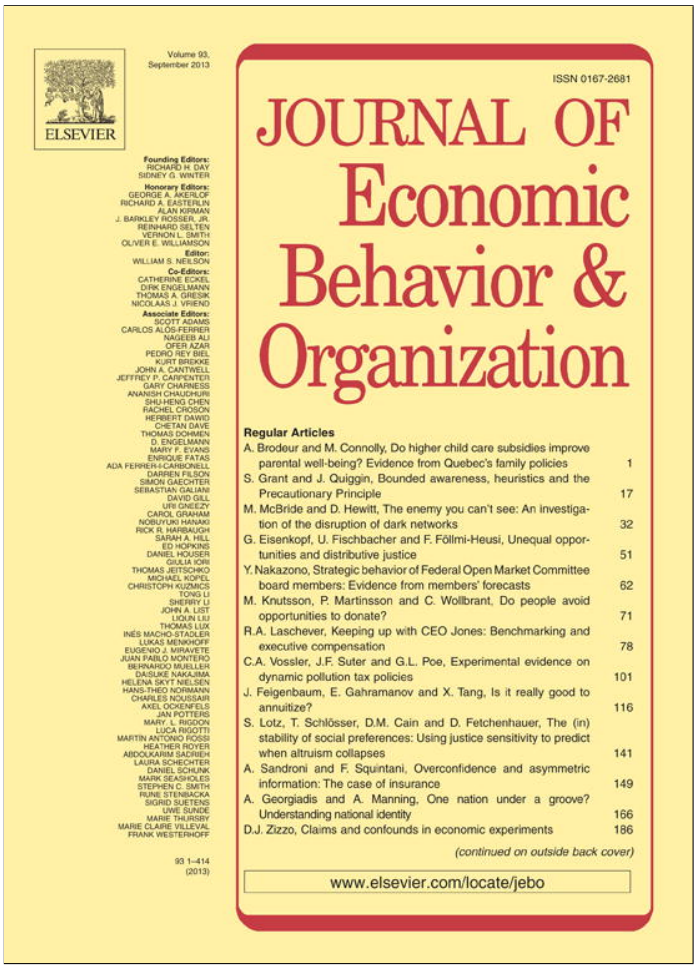

Journal of Economic Behavior & Organization 93 (2013) 227–232

Contents lists available at ScienceDirect

Journal of Economic Behavior & Organization

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

Deception: The role of guilt

Pierpaolo Battigalli [a], Gary Charness [b], Martin Dufwenberg [c,d,e,*]

[a] Department of Economics and IGIER, Bocconi University, via Rontgen 1, Milan, Italy
[b] Department of Economics, University of California, 2127 North Hall, Santa Barbara, CA 93106-9210, USA
[c] Department of Economics, University of Arizona, Tucson, AZ 85721-0108, USA
[d] Department of Economics, University of Gothenburg, Box 640, SE-40530 Gothenburg, Sweden
[e] Department of Decision Sciences and IGIER, Bocconi University, via Rontgen 1, Milan, Italy

ARTICLE INFO

Article history:
Received 14 March 2013
Accepted 18 March 2013
Available online 4 April 2013

JEL classification:
C72
C91
D03

Keywords:
Deception
Guilt aversion
Experiments
Psychological games

ABSTRACT

Evidence suggests that whether or not people dislike lying is situation-dependent. We argue that the theory of simple guilt can accommodate this well.

1. Introduction

Gneezy (2005) reports intriguing experimental evidence indicating that people do not like to lie. His subjects deceive primarily if they thereby gain a lot, or impose little loss. Through a carefully spun web of treatments (presented below) he highlights ways in which some seemingly plausible models of motivation (e.g. distributional preferences, or a fixed cost of lying) fall short of capturing the central tendencies of the data.

In other situations people habitually lie without remorse. We suggest that examples can be drawn from used car sales, promises made by politicians, tax returns sent to the IRS, testimony in traffic courts (under oath!), and game shows like Survivor. These examples are confounded though; people may dislike lying per se and yet lie because of countervailing benefits. But no such confound can touch the following example taken from the world of poker. It concerns chit-chat amongst players between deals (not regular bluffs). In his book *Bad Beats and Lucky Draws*, Phil Hellmuth, Jr. (2005, p. 34) describes a Texas Hold'Em game in which he held 10♥-6♥. He ended up not having to show his cards. Another player (Johnny Chan) said: "I thought you had a pair of sevens and a flush draw." Hellmuth responded: "Nope, actually I had the 10♦-J♦." This is a lie of commission! One might take Hellmuth to be a type with an unusually limited aversion to lying. But that is

* Corresponding author at: Department of Economics, University of Arizona, Tucson, AZ 85721-0108, USA. Tel.: +1 520 626 1540; fax: +1 520 621 8450.
E-mail addresses: pierpaolo.battigalli@unibocconi.it (P. Battigalli), charness@econ.ucsb.edu (G. Charness), martind@eller.arizona.edu (M. Dufwenberg).

0167-2681/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jebo.2013.03.033

Table 1
Payoffs used in CTSR game

Treatment	Option	Payoff to	
		Player 1	Player 2
1	A	5$	6$
	B	6$	5$
2	A	5$	15$
	B	6$	5$
3	A	5$	15$
	B	15$	5$

not the case. He writes: "Although I never lie outside of poker, to me, lying about what you just had in a poker hand is part of bluffing. Why give someone a 'free read' on your play?"[1]

We argue that Battigalli and Dufwenberg's (2007) (cf. Battigalli and Dufwenberg, 2009,[2] Geanakoplos et al., 1989) theory of simple guilt can explain the central tendencies of Gneezy's data, while accommodating other situations where people do not suffer when they lie. Section 2 recalls Gneezy's results, Section 3 introduces guilt, Section 4 describes the fit with data, and Section 5 concludes.

2. Gneezy's experiment

Gneezy studies a two-player "cheap talk sender-receiver" (CTSR) game. There are two options, A and B. Only player 1 is informed of the involved monetary consequences, and then sends one of two messages to player 2:

Message A: "Option A will earn you more money than option B."
Message B: "Option B will earn you more money than option A."

Player 2 must choose between options A and B after getting 1's message. The monetary consequences, known to 1 but not to 2, vary across three treatments as described in Table 1:

Message A tells the truth; message B is a lie. Message B was chosen in, respectively, 36%, 17%, and 52% of the cases in treatments 1, 2, and 3.

In order to determine if these results reflect aversion to lying (as opposed to preferences over distributions of payoffs) Gneezy employs three dictator treatments, where player 1 chooses between options A and B and player 2 has no choice. For the CTSR games, Gneezy reports evidence (p. 386) that player 2 followed 1's message in about 80% of the cases, and player 1 expected the message to be followed in about 80% of the cases. To allow comparability, in the dictator games the probability of executing 1's choice was 80% with the dollar consequences as seen in Table 1. If lying were painless, one would expect the frequency of option B choices in the dictator game to match the frequency of message B choices in the CTSR games.

That did not happen. Option B was chosen in, respectively, 66%, 42%, and 91% of the cases; each number is significantly higher than the corresponding one in the CTSR treatments. Gneezy concludes: "it is not only care for others that motivate behavior, but also aversion to lying" (p. 388). However, the (significant) difference between CTSR treatments 1 and 2 (36% vs. 17%) suggests that assuming a fixed cost of lying will not by itself do the job.

3. Simple guilt

B&D (2007) introduce a theory of guilt aversion, which applies to extensive games with monetary payoffs. The basic idea is that player i suffers from guilt to the extent that he believes that player $j \neq i$ gets a lower (monetary) payoff than i believes j believes she will get.[3] For a two-player game, a psychological utility function of player 1, u_1, can be defined thus:

$$u_1(z, \alpha_2) = \pi_1(z) - \theta_1 \max\{0, \mathbb{E}_{\alpha_2}[\pi_2] - \pi_2(z)\}, \tag{1}$$

where z is the outcome of the game (terminal node reached), $\pi_i(z)$ is the dollar payoff of player i at z, α_2 is player 2's pre-play belief on how the game will be played, $\mathbb{E}_{\alpha_2}[\pi_2]$ is 2's subjective expected payoff calculated using α_2, and θ_1 is an exogenously given positive constant.

[1] Hellmuth (2005) is not a unique case. Leading poker texts actively encourage lies, or at least very deceptive use of language and demeanor. For some colorful testimony, we refer to several examples in Brunson (1978/2002); see *e.g.* pp. 80–81, 88–89, 105–106, 427–428 (the first three of these examples are crafted by "Crazy Mike" Caro).

[2] From here on, Battigalli and Dufwenberg will be abbreviated with B&D.

[3] This conforms well with findings in social psychology, *e.g.* by Baumeister et al. (1994, 1995).

We refer to B&D (2007) for more discussions about mathematical details, and here concentrate on the interpretation of (1) and its application to Gneezy's games. Eq. (1) says that in a situation of conflict of material interests, like Gneezy's games, the increase in player 1's payoff $[\pi_1(B) - \pi_1(A)] > 0$ may be offset by the guilt cost due to the increase in the disappointment of player 2 caused by his lower payoff

$$\max\{0, \mathbb{E}_{\alpha_2}[\pi_2] - \pi_2(B)\} - \max\{0, \mathbb{E}_{\alpha_2}[\pi_2] - \pi_2(A)\} \geq 0, \tag{2}$$

where inequality (2) is strict if 2 initially expects to get more than $\pi_2(B)$ ($\mathbb{E}_{\alpha_2}[\pi_2] - \pi_2(B) > 0$). The extent of this psychological cost is given by θ_1, which measures 1's sensitivity to guilt. Note that player 1's utility depends on a variable he does not know, the first-order belief α_2 of the co-player. To compute the expected utility of his different courses of action he has to use his second-order beliefs about the first-order beliefs of player 2, $\beta_{1,2}$.

Guilt aversion induces in players a tendency to live up to what they perceive others to expect. Moreover, communication may then move beliefs, motivation, and behavior. For example, if player 1 makes a promise to player 2 this may be credible because if 1 believes 2 believes him he will (being guilt-averse) wish to deliver.

4. Taking simple guilt to data

B&D's (2007) guilt aversion theory can be applied to Gneezy's cheap talk game by introducing incomplete information in their framework. Because player 2 has no knowledge of the monetary payoffs, he does not even know whether material interests are common or in conflict; he only knows that player 1 knows them (cf. B&D (2009), Section 6.2). But the CTSR game situation is sufficiently simple that it can be formally described by introducing a few compelling assumptions and belief-dependent variables.

From the point of view of player 2 (henceforth *receiver*) the pair of dollar payoff functions is an unknown $\pi^t = (\pi_1^t, \pi_2^t) \in \mathbb{R}_+^{\{A,B\}} \times \mathbb{R}_+^{\{A,B\}}$ determined by a treatment parameter $t \in T$ observed only by player 1 (*sender*).[4] The sender chooses the message $m \in \{m^A, m^B\}$ as a function of the observed value of t. The CTSR game has four terminal nodes, or paths, $Z = \{m^A, m^B\} \times \{A, B\}$; but messages do not affect material payoffs, therefore we write, for example, $\pi_i^1(B)$ instead of $\pi_i^1(m^A, B)$, as we did in Eq. (2). The size of set $\{(\pi_1^t, \pi_2^t) : t \in T\} \subset \mathbb{R}_+^{\{A,B\}} \times \mathbb{R}_+^{\{A,B\}}$ reflects the ignorance of player 2. We assume this set is large; in particular, it contains the three treatments of Gneezy's experiment; see Assumptions 1 and 2. We also assume that the sender knows the set $\{(\pi_1^t, \pi_2^t) : t \in T\}$ of payoff functions contemplated by the receiver. According to B&D's (2007) theory, the receiver has a first-order belief $\alpha_{2,1} \in \Delta(T \times S_1)$, where $S_1 = \{m^A, m^B\}^T$ is the set of cheap talk strategies of the sender, as conceived by the receiver.[5] The plan of the receiver on how to play the game can be represented as a belief about his own strategy $\alpha_{2,2} \in \Delta(S_2)$, where $S_2 = \{A, B\}^{\{m^A, m^B\}}$. Without loss of generality we assume that $\alpha_{2,2}$ assigns probability one to a pure strategy. In particular, we focus on two pure strategies of the receiver: the "Yes-man" or trusting strategy Y = (A if m^A, B if m^B), and the "contrarian" strategy N = (B if m^A, A if m^B). The first-order belief $\alpha_2 = \alpha_{2,1} \times \alpha_{2,2} \in \Delta(T \times S_1 \times S_2)$ determines a probability distribution on $T \times \{A, B\}$ and hence a subjective expected payoff $\mathbb{E}_{\alpha_2}[\pi_2]$. We let $\Pi_2^Y = \mathbb{E}_{\alpha_{2,1} \times Y}[\pi_2]$ denote the receiver's expected payoff if he plans to trust the sender; similarly, $\Pi_2^N = \mathbb{E}_{\alpha_{2,1} \times N}[\pi_2]$ denotes his expected payoff if he plans to do the opposite of what the sender suggests. Symmetry considerations and a principle of insufficient reason yield the following assumption about first-order beliefs of the receiver:

Assumption 1. The first-order beliefs of the receiver about payoffs and the sender, $\alpha_{2,1}$, are such that the expected payoff from strategy Y (resp. strategy N) conditional on the received message $m \in \{m^A, m^B\}$ is well defined and independent of m, hence equal to Π_2^Y (resp. Π_2^N). Therefore strategy Y (resp. N) is the unique best response if and only if $\Pi_2^Y > \Pi_2^N$ (resp. $\Pi_2^N > \Pi_2^Y$).

Our analysis focuses on the behavior of the sender, therefore our key assumptions concern his second-order beliefs.

Assumption 2. The second-order beliefs of the sender about the receiver, $\beta_{1,2}$,[6] are independent of t and such that the sender believes that

(i) Assumption 1 holds,
(ii) the receiver is subjectively rational, i.e. he best responds to his beliefs $\alpha_{2,1}$,
(iii) (Π_2^Y, Π_2^N) (a feature of the receiver's belief $\alpha_{2,1}$) is continuously distributed with support $[0, \overline{\Pi}]^2$ where $\overline{\Pi} > 15$,
(iv) the probability that $\Pi_2^Y \geq \Pi_2^N$ is more than 50%: $\mathbb{P}_{\beta_{1,2}}[\Pi_2^Y \geq \Pi_2^N] > 0.5$.

[4] The set of functions with domain X and codomain Y is denoted by Y^X. For example, $\pi_i^t \in \mathbb{R}_+^{\{A,B\}}$ is a pair of numbers: $\pi_i^t = (\pi_i^t(A), \pi_i^t(B))$. It is common knowledge that monetary payoffs (gross of show up fee) in experiments cannot be negative. Hence it is common knowledge that in the CTSR game $\pi_i^t \in \mathbb{R}_+^{\{A,B\}}$, $i = 1, 2$.

[5] We can think of $\alpha_{2,1}$ as the marginal of an extended belief $\overline{\alpha}_{2,1} \in \Delta(T \times \Theta_1 \times S_1)$ that also encompasses the guilt type of player 1. But this is not necessary here.

[6] By Assumption 1, the initial beliefs of player 2 determine his beliefs conditional on each message. Therefore here we can model α_2 as a point in $\Delta(T \times S_1 \times S_2)$ and $\beta_{2,1}$ as a point in $\Delta(S_2 \times \Delta(T \times S_1 \times S_2))$.

Note that part (iv) of this second assumption is in line with the evidence reported in Section 2. With this, we can express the expected utility of type θ_1 from sending message m^z ($z \in \{A, B\}$) given treatment t in a relatively simple form:

$$U_z^t(\theta_1) = [\pi_1^t(z) - \theta_1 D^Y(\pi_2^t(z))]P^Y + [\pi_1^t(z') - \theta_1 D^N(\pi_2^t(z'))](1 - P^Y), \tag{3}$$

where $z,\ z' \in \{A,\ B\},\ z \neq z',\ P^Y = \mathbb{P}_{\beta_{1,2}}[\Pi_2^Y \geq \Pi_2^N] > 0.5$ is the probability of the trusting strategy Y,[7] $D^Y(x) = \mathbb{E}_{\beta_{1,2}}[\max\{0, \Pi_2^Y - x\} | \Pi_2^Y \geq \Pi_2^N]$ is the expected disappointment of a trusting receiver if he gets x dollars, and $D^N(x) = \mathbb{E}_{\beta_{1,2}}[\max\{0, \Pi_2^N - x\} | \Pi_2^Y < \Pi_2^N]$ is the expected disappointment of a contrarian receiver if he gets x dollars. Of course, all these probabilities and expectations depend on the second-order beliefs of the sender, $\beta_{1,2}$, but we do not make it explicit in Eq. (3) to simplify the notation. The following assumption simplifies the analysis[8]:

Assumption 3. The sender expects that, on average, trusting and contrarian receivers are equally disappointed by any payoff in the relevant range, that is, $D^Y(x) = D^N(x)$ for each $x \in [0, \overline{\Pi}]$.

Letting $D(x)$ denote the common expectation of the sender of the disappointment of trusters and contrarian receivers, the expected utility gain from lying can be expressed as follows:

$$U_B^t(\theta_1) - U_A^t(\theta_1) = [\pi_1^t(B) - \pi_1^t(A) - \theta_1(D(\pi_2^t(B)) - D(\pi_2^t(A)))](2P^Y - 1). \tag{4}$$

The disappointment of the receiver when he gets x dollars, $\max\{0, \Pi_2 - x\}$, is decreasing and convex in x. The sender's expectation of this disappointment, $D(x)$, is the integral of $\max\{0, \Pi_2 - x\}$ with respect to the unknown expectation Π_2, given second-order beliefs $\beta_{1,2}$. Therefore, also $D(x)$ must be decreasing and convex. Our assumptions about the second-order beliefs of the sender imply that these properties hold strictly (see Appendix A):

Lemma 1. *The expected disappointment $D(x)$ is strictly decreasing and strictly convex on $[0, \overline{\Pi}]$.*

Corollary 1. *For each $x \in [0, \overline{\Pi})$, the incremental ratio $(D(x) - D(x+h))/h$ is strictly decreasing in h on $(0, \overline{\Pi} - x)$.*

Proof. Let $\Delta(h) = D(x) - D(x+h)$. By definition $\Delta(0) = 0$. Lemma 1 implies that $\Delta(h)$ is strictly concave. Therefore the incremental ratio $\Delta(h)/h$ is strictly decreasing. □

Now recall that, by Assumption 2, $2P^Y > 1$. Furthermore, $\pi_1^t(A) < \pi_1^t(B)$ and $\pi_2^t(A) > \pi_2^t(B)$ in each treatment $t = 1, 2, 3$. By Lemma 1, $D(\pi_2^t(B)) - D(\pi_2^t(A)) > 0$ for each $t = 1, 2, 3$. Therefore the difference in Eq. (4) is decreasing in θ_1 and the indifference equation $U_B^t(\theta_1) - U_A^t(\theta_1) = 0$ has a unique and positive solution

$$\hat{\theta}^t = \frac{\pi_1^t(B) - \pi_1^t(A)}{D(\pi_2^t(B)) - D(\pi_2^t(A))}. \tag{5}$$

A sender of type θ_1 lies in treatment t if and only if $\theta_1 < \hat{\theta}^t$.

Proposition 1. *Under Assumptions 2–3, the thresholds $\hat{\theta}^1, \hat{\theta}^2, \hat{\theta}^3$ are ordered as follows: $0 < \hat{\theta}^2 < \hat{\theta}^1 < \hat{\theta}^3$.*

Proof. Plugging in Eq. (5) the treatments values we have

$$\hat{\theta}^1 = \frac{1}{D(5) - D(6)},\ \hat{\theta}^2 = \frac{1}{D(5) - D(15)},\ \hat{\theta}^3 = \frac{10}{D(5) - D(15)}.$$

Lemma 1 yields

$$0 < \frac{1}{D(5) - D(15)} < \frac{1}{D(5) - D(6)}.$$

Corollary 1 yields

$$\frac{1}{D(5) - D(6)} < \frac{10}{D(5) - D(15)}. \qquad \square$$

To obtain predictions about the frequency of lies we have to postulate a distribution of guilt sensitivity and second-order beliefs in the population of (potential) senders.

Assumption 4. Guilt sensitivity θ_1 and second-order beliefs $\beta_{1,2}$ are independently distributed, and the cumulative distribution function of θ_1, $G : \mathbb{R}_+ \to [0, 1]$, is continuous and strictly increasing.

[7] By Assumption 2, the probability that the receiver is indifferent is zero, therefore $\mathbb{P}_{\beta_{1,2}}[\Pi_2^Y \geq \Pi_2^N] = \mathbb{P}_{\beta_{1,2}}[\Pi_2^Y > \Pi_2^N]$.

[8] Assumption 3 says that the difference function $\Delta^{YN}(x) = D^Y(x) - D^N(x)$ is identically zero. Our results still hold if $\Delta^{YN}(x)$ is assumed to be non-negative, weakly decreasing and weakly convex. We argue that this is plausible. If the sender believes that trusting receivers are on average more optimistic, hence more disappointed than contrarians, then $\Delta^{YN}(x) \geq 0$. Since disappointment must be zero when monetary payoff x is high, the same holds for $\Delta^{YN}(x)$. With this, it is a small step to further assume that $\Delta^{YN}(x)$ is weakly decreasing and convex.

The independence assumption is not necessary for our results, but it simplifies the analysis. Let F denote distribution of second-order beliefs $\beta_{1,2}$. The observed frequency of lies in treatment t within a large random sample of senders is approximately

$$F^t(\text{lies}) = \int G(\hat{\theta}^t(\beta_{1,2}))F(\mathrm{d}\beta_{1,2}),$$

where we made explicit the dependence of threshold $\hat{\theta}^t$ on second-order beliefs. Given Assumption 4 about the distribution of guilt sensitivity and second-order beliefs, Proposition 1 yields the qualitative result observed in Gneezy's experiment:

Proposition 2. *Under Assumptions 2–4, the frequencies of lies in treatments 1–3 are as follows:*

$$0 < F^2(\text{lies}) < F^1(\text{lies}) < F^3(\text{lies}) < 1.$$

5. Conclusion

Simple guilt provides a psycho-foundation for honesty, in some situations. It presumes that motivation is belief-dependent, in a particular way, and therefore words may move beliefs, motivation, and behavior. For example, Charness and Dufwenberg (2006) report experimental evidence that for these reasons promises may foster trust and cooperation in situations characterized by hidden action (moral hazard).[9] Our take on Gneezy's design is that the sender is similarly forced to move the receiver's beliefs, and through anticipation this shapes the sender's behavior in line with the observed treatment effects.[10]

While the belief-dependence of guilt allows that communication moves beliefs, it does not have to be that way in all settings. Poker regulars do not take between-deals chit-chat at face value. In our earlier example, Johnny probably expects Phil to lie. Even if the long-run effect (say relative to silence) is to increase Phil's payoff by $\$x$ at Johnny's expense, this is just what Johnny expects. Therefore Phil suffers no remorse. We propose that the other examples from the introduction (car sales, tax returns, etc.) where people lie routinely can be partly understood similarly.

The theory of simple guilt is capable of picking up the central tendencies of Gneezy's data for the CTSR game, but this must not be misinterpreted as suggesting that other forms of motivation are not important as well. For example, while we do not invoke a fixed cost of lying, it is clear from other experiments that many people may dislike lying even in situations where such deception furnishes material gains for everyone (see e.g. Erat and Gneezy, 2012). Furthermore, this motivation can be combined with simple guilt to explain the comparison between the CTSR game and the Dictator game. First, some senders in the CTSR treatments may tell the truth even if they are not guilt averse because they dislike lying *per se*. Second, if it is commonly believed that some people dislike lying, then receivers should have on average more optimistic expectations than they would have in the passive player role of the corresponding Dictator treatments. Understanding this, even the guilt averse senders with low cost of lying *per se* are less prone to deceive.

Another example of motivation not considered here is B&D's (2007) "guilt from blame". While simple guilt models a conscience which is "internalized" in the sense that player i consults his own beliefs of the degree to which he hurts another player j relative to j's expectations, under guilt from blame i suffers to the extent that he believes j infers (at the end of the game) that i set out (at the beginning of the game) to hurt j relative to j's expectations. See B&D (2007) for formal details. Guilt from blame may be important in many settings where players' impressions of each other are shaped by play – see e.g. Charness and Dufwenberg (2011) for an example – but is largely irrelevant as regards Gneezy (2005) design. Since player 2 has no information about the treatment, his inferences regarding the extent to which player 1 set out to hurt player 2 relative to 2's expectations are the same across treatments. If player 1 understands this, guilt from blame predicts the same behavior for 1 in all treatments.

Acknowledgements

We thank Ernst Fehr for encouraging us to explore if guilt aversion could explain Gneezy's data, and Simone Cerreia-Vioglio, Olof Johansson-Stenman, Peter Martinsson, and the many referees for their helpful comments. Pierpaolo Battigalli gratefully acknowledges financial support from ERC Grant 324219.

[9] See also Dufwenberg and Gneezy (2000). Some exciting controversy and debate surrounds Charness and Dufwenberg (2006); see e.g. Bellemare et al. (2011), Ellingsen et al. (2010), Ismayilov and Potters (2012), Rueben et al. (2009), and Vanberg (2008).

[10] Some of the papers in this special issue document effects that may reflect simple guilt at play. For example, Gino et al. ((2013)) consider moral flexibility, so that one is more likely to view dishonesty as morally acceptable and thus feel less guilty about benefiting from cheating when one's dishonesty benefits others. Jiang ((2013)) considers how a subtle variation in the rules of a game affect cheating. Fosgaard et al. (2013) examine how the information provided to participants affect the likelihood of cheating, implicitly changing the beliefs of the players.

Appendix A.

Proof of Lemma 1. By Assumptions 2 and 3, there is a density function $\beta : [0, \overline{\Pi}] \to \mathbb{R}$ strictly positive on $(0, \overline{\Pi})$ such that

$$D(x) = \int_0^{\overline{\Pi}} \max\{\Pi_2 - x, 0\}\beta(\Pi_2)\mathrm{d}\Pi_2 = \int_x^{\overline{\Pi}} (\Pi_2 - x)\beta(\Pi_2)\mathrm{d}\Pi_2.$$

To ease notation we write $D(x|\Pi_2) = \max\{\Pi_2 - x, 0\}$. Fix two payoffs $x < y$ in $[0, \overline{\Pi}]$. We first show that $D(x) > D(y)$, hence $D(\cdot)$ is strictly decreasing. Observe that

$$D(x|\Pi_2) - D(y|\Pi_2) = \begin{cases} 0, & \text{if } \Pi_2 \le x \\ \Pi_2 - x > 0, & \text{if } \Pi_2 \in (x, y) \\ y - x > 0, & \text{if } \Pi_2 > y \end{cases}$$

Therefore

$$D(x) - D(y) = \int_x^{\overline{\Pi}} [D(x|\Pi_2) - D(x|\Pi_2)]\beta(\Pi_2)d\Pi_2 > 0$$

because is $\beta(\Pi_2)$ strictly positive on $(x, \overline{\Pi})$.

For each $\lambda \in (0, 1)$, we let $\overline{x}(\lambda)$ denote the corresponding convex combination of x and y: $\overline{x}(\lambda) = \lambda x + (1 - \lambda)y$. We show that $D(\overline{x}(\lambda)) < \lambda D(x) + (1 - \lambda)D(y)$; hence $D(\cdot)$ is strictly convex. First note that $D(x'|\Pi_2) = \max\{\Pi_2 - x', 0\}$ is a convex function of x'. Thus, for each $\Pi_2 \in [0, \overline{\Pi}]$,

$$D(\overline{x}(\lambda)|\Pi_2) \le \lambda D(x|\Pi_2) + (1 - \lambda)D(y|\Pi_2).$$

Next observe that, for each $\Pi_2 \in (x, \overline{x}(\lambda))$, $D(x|\Pi_2) = \Pi_2 - x > 0$ and $D(\overline{x}(\lambda)|\Pi_2) = 0 = D(y|\Pi_2)$, hence

$$D(\overline{x}(\lambda)|\Pi_2) = 0 < \lambda(\Pi_2 - x) = \lambda D(x|\Pi_2) + (1 - \lambda)D(y|\Pi_2).$$

These inequalities and the fact that $\beta(\Pi_2)$ is strictly positive on the (nonempty) open interval $(x, \overline{x}(\lambda))$ imply

$$D(\overline{x}(\lambda)) = \int_{\overline{x}(\lambda)}^{\overline{\Pi}} D(\overline{x}(\lambda)|\Pi_2)\beta(\Pi_2)d\Pi_2 < \int_x^{\overline{x}(\lambda)} [\lambda D(x|\Pi_2) + (1 - \lambda)D(y|\Pi_2)]\beta(\Pi_2)d\Pi_2$$
$$+ \int_{\overline{x}(\lambda)}^{\overline{\Pi}} [\lambda D(x|\Pi_2) + (1 - \lambda)D(y|\Pi_2)]\beta(\Pi_2)d\Pi_2 \le \lambda D(x) + (1 - \lambda)D(y). \qquad \square$$

References

Battigalli, P., Dufwenberg, M., 2007. Guilt in games, American economic review. Papers and Proceedings 97, 170–176.
Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. Journal of Economic Theory 144, 1–35.
Baumeister, R., Stillwell, A., Heatherton, T., 1994. Guilt: an interpersonal approach. Psychological Bulletin 115, 243–267.
Baumeister, R., Stillwell, A., Heatherton, T., 1995. Personal narratives about guilt: role in action control and interpersonal relationships. Basic and Applied Social Psychology 17, 173–198.
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, 437–453.
Brunson, D., 1978/2002. Super System: A Course in Power Poker, 3rd ed. Cardoza Publishing, New York.
Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econome 74, 1579–1601.
Charness, G., Dufwenberg, M., 2011. Participation. American Economic Review 101, 1213–1239.
Dufwenberg, M., Gneezy, U., 2000. Measuring beliefs in an experimental lost wallet game. Games and Economic Behavior 30, 163–182.
Ellingsen, T., Johannesson, M., Tjotta, S., Torsvik, G., 2010. Testing guilt aversion. Games and Economic Behavior 68, 95–107.
Erat, S., Gneezy, U., 2012. White lies. Management Science 58, 723–733.
Fosgaard, T., Hansen, L., Pievosan, M., 2013. Separating will from grace: an experiment on conformity and awareness of cheating. Journal of Economic Behavior and Organization 93, 279–284.
Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games and Economic Behavior 1, 60–79.
Gino, F., Ayal, A., Ariely, D., 2013. Self-serving altruism? The lure of unethical actions that benefit others. Journal of Economic Behavior and Organization 93, 285–292.
Gneezy, U., 2005. Deception: the role of consequences. American Economic Review 95, 384–394.
Hellmuth Jr., P., 2005. Bad Beats and Lucky Draws. HarperCollins, New York.
Ismayilov, H., Potters, J., 2012. Promises as commitments. Unpublished manuscript.
Jiang, T., 2013. Cheating in mind gamesv – the subtlety of rules matters. Journal of Economic Behavior and Organization 93, 328–336.
Rueben, E., Sapienza, P., Zingales, L., 2009. Is mistrust self-fulfilling? Economic Letters 104, 89–91.
Vanberg, C., 2008. Why do people keep their promises? An experimental test of two explanations. Econometrica 76, 1467–1480.