Experimental Economics 16.3 (2012): 233-247

DOI: http://dx.doi.org/10.1007/s10683-012-9324-x

Why do people tell the truth? Experimental evidence for pure lie aversion

Raúl López-Pérez · Eli Spiegelman

Abstract A recent experimental literature shows that truth-telling is not always motivated by pecuniary motives, and several alternative motivations have been proposed. However, their relative importance in any given context is still not totally clear. This paper investigates the relevance of pure lie aversion, that is, a dislike for lies independent of their consequences. We propose a very simple design where other motives considered in the literature predict zero truth-telling, whereas pure lie aversion predicts a non-zero rate. Thus we interpret the finding that more than a third of the subjects tell the truth as evidence for pure lie aversion. Our design also prevents confounds with another motivation (a desire to act as others expect us to act) not frequently considered but consistent with much existing evidence. We also observe that subjects who tell the truth are more likely to believe that others will tell the truth as well.

We are indebted to Matthieu Chemin, Joan Crespo, Bruno Deffains, Claude Fluet, Sean Horan, Hubert Kiss, Pierre Laserre, and Marc Vorsatz for helpful comments. We also gratefully acknowledge financial support from the Spanish Ministry of Education through the research project ECO2008-00510, and helpful research assistance by Sayoko Aketa and David Sánchez. Eli Spiegelman thanks Claude Fluet for guidance, encouragement and support.

R. López-Pérez
Department of Economic Analysis, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain
e-mail: raul.lopez@uam.es

E. Spiegelman (✉)
Departement des Sciences Économiques, Université de Québec a Montréal, Montréal, Canada
e-mail: spiegele@vaniercollege.qc.ca

E. Spiegelman
Department of Economics and Political Science, Vanier College, St-Laurent, Canada

Keywords Communication · Honesty · Guilt aversion · Lie aversion · Norms

JEL Classification C70 · C91 · D03 · D64

1 Introduction

In many important economic settings, people may increase their expected material gain by providing information that they believe to be false—in short, by lying. Immediate examples include accounting, auditing, insurance claims, job interviews, labor negotiations, regulatory hearings, and tax compliance. Based on the standard *Homo economicus* assumption that all agents are self-interested money maximizers, economic theory predicts that people will always respond to these situations mendaciously. However, a recent experimental literature shows that people often tell the truth in cases where the theory would not predict it. The broad research question this paper addresses is: Why do they do that? One potential motivation for honesty that has attracted attention in the literature is *pure lie aversion*, the idea that people suffer a utility cost when they tell a lie (e.g., Ellingsen and Johannesson 2004; Kartik 2009). Intuitively, people care about social norms or ethical principles that forbid lying, such as those based on religions like Christianity and Islam, and feel bad if they utter a lie.[1] In this paper, we experimentally investigate the extent to which people are motivated by pure lie aversion.

Experimental methods are necessary for exploring the relevance of lie aversion. While the observation of truthful behavior in everyday life seems in line with the idea that (some) people are lie-averse, more controlled decision contexts are required to avoid confounds with other potential motivators of honesty. For instance, pure lie aversion indicates that the lie itself is the source of motivation. This must be distinguished from motivations linked to the consequences of the lie. In some occasions, for example, a person might believe that truth-telling will lead to a higher material payoff (this might be particularly frequent in repeated interactions, where reputational issues are key). Altruism might be also a motivator of honesty, if the communicator believes that deception will result in a harmful choice. Related to this, note that communication might reduce social distance (Bohnet and Frey 1999) or create some sort of social identity (Orbell et al. 1990; Buchan et al. 2006), and that could in turn reinforce altruistic feelings.

Another layer of complexity in the natural setting concerns the interaction of lying with beliefs. Beliefs can interact with consequences, as in the idea, first tested in Dufwenberg and Gneezy (2000), that people feel bad if they believe that their choice will result in others' not receiving the payoff that they expected. This so-called *guilt aversion* results in a psychological game (as in Battigalli and Dufwenberg 2007, 2009), as preferences depend on beliefs about the payoff actually expected by the other person—the so-called 'second-order expectations'. In this setting, people might avoid false promises: lies that incorrectly inflate the other's payoff expectations. Such

[1] Ellingsen and Johannesson (2004) and Gneezy (2005) review some psychological literature on lie aversion. Gneezy (2005) also considers the views of some classical philosophers on the morality of deception.

a hypothesis has been object of experiments including Charness and Dufwenberg (2006), and has received some verification.[2]

People might not only care about payoff expectations, but also about choice expectations. As an example, people might be conformist in the sense that they are more likely to comply with a certain norm if they believe others expect them to. Applied to honesty, this implies that a message sender will be more likely to follow a norm of honesty when she believes that the message receiver expects the truth. We call this *belief-dependent lie aversion*. While this theory has not received much attention in the literature (see Peeters et al. 2007 for an exception), it is potentially important because it is consistent with much experimental evidence and with behavioral patterns studied in psychology. For instance, Rosenthal (2003, p. 151) asserts the existence of 'hundreds of studies [demonstrating] that one person's expectations for the behavior of another person can actually affect that other person's behavior', even when those expectations are not made explicit.

We thus have at least five potential motivators of honesty in natural settings (selfishness, altruism, guilt aversion, lie aversion and its belief-dependent variant). Our experimental design allows us to investigate the relevance of pure lie aversion, controlling for the impact of the other motivations. We consider two treatments of a very simple decision problem where one agent must send a (truthful or false) message to another. Belief-dependent lie aversion predicts a positive but different rate of truth-telling across treatments, whereas pure lie aversion predicts the same positive rate in both treatments. Altruism, guilt aversion, and selfishness predict zero truth-telling. In both treatments, furthermore, we measure first- and second-order beliefs about truth-telling, which should be correlated with behavior according to belief-dependent lie aversion.

Overall, 38.76 percent of the subjects chose to tell the truth in our study, but there was not a significant difference in the rate of truth-telling across treatments. Our results therefore suggest that pure lie aversion is a significant force behind honesty, whereas its belief-dependent variant is not, at least in this context. Yet we do find that first- and second-order beliefs co-vary significantly with behavior. For instance, subjects who lied in our study were significantly more likely to believe that others would lie as well. This suggests a need to enrich the assumptions surrounding the theory of lie aversion (as suggested in Bicchieri 2005; López-Pérez 2012, or Erat and Gneezy 2012).

Our study contributes to a burgeoning experimental literature examining deception and honesty—e.g., Gneezy (2005), Charness and Dufwenberg (2006), Vanberg (2008), Hurkens and Kartik (2009), Sutter (2009). It is most closely related to Peeters et al. (2007), Fischbacher and Heusi (2008), Erat and Gneezy (2012), and Sánchez-Pagés and Vorsatz (2009), which provide evidence consistent with the existence of lie-averse agents, controlling for several potential confounds. We offer additional evidence in this line and complement this literature with a design that can discriminate

[2]We stress that the crucial difference between pure lie aversion and guilt aversion is that the latter posits a relation between beliefs and the activation of the bad feelings, whereas pure lie aversion assumes that the bad feelings are activated simply by uttering a lie. The name given to such feelings, in contrast, is largely immaterial for the distinction. Pure lie aversion does not rule out that the bad feelings are what psychologists call guilt.

between two forms of lie aversion,[3] controlling moreover for the effect of altruism, social distance, social identity, and guilt aversion on truth-telling.

The rest of the paper is structured as follows. The next section formalizes and discusses the idea of belief-dependent lie aversion (BDLA). Section 3 describes our experimental design and procedures, and explains the predictions by BDLA and other theories. Section 4 reports the results from our experiment, and Sect. 5 concludes.

2 Belief-dependent lie aversion

For illustrative purposes, we focus our description on a very simple decision problem: A *sender* privately observes a random signal θ with n possible realizations and must then choose a message for another agent (the *receiver*), who makes no choice. A message must announce one of the possible values of θ. Thus, whatever the actual value of θ, the sender always has n messages available. Clearly, a message can be true or false depending on whether it announces the actual value of θ. For simplicity, we assume a literal meaning to these messages, so that there can be no coordination problem between messages and values of θ.

Let $x_S(z)$ denote the sender's monetary payoff at terminal node z, and $I(z)$ denote an indicator taking value 1 at terminal node z if the sender lied in the history of z, and value 0 otherwise. Suppose then that at any z, the receiver assigns some probability to $I(z) = 1$, and that the sender has second-order beliefs about this probability. These beliefs can take the form of a probability distribution; let $\mu(z) \in [0, 1]$ denote its mean. That is, the sender at node z thinks that 'on average' the receiver will think that the sender has lied with probability $\mu(z)$. Furthermore, let $D_S(z) = \max[0, I(z) - \mu(z)]$ denote the sender's perception of the degree of deviance that the receiver expects at z. Note that $D_S(z) > 0$ only if the sender lied; our aim is a simple formalization of the intuition that people are more likely to respect a norm (of honesty in this case) if others expect them to: in their normative decisions, people tend to conform to others' expectations. Thus, the sender's preferences $\succ$ are defined over the set of vectors $[x_S(z), D_S(z)]$, and satisfy rationality, plus three axioms. The first is a monotonicity axiom: other things equal, people prefer more money to less:

Axiom 1 (monotonicity) *Given* $D_S(z) = D_S(z')$, $[x_S(z), D_S(z)] \succ [x_S(z'), D_S(z')]$ *if* $x_S(z) > x_S(z')$.

The second axiom is a continuity hypothesis. Intuitively, this states that any deviance can be compensated by a sum Δ of money:

[3] Peeters et al. (2007) run an experiment with a sender-receiver game played over 100 rounds with rematching, and analyze the performance of a model of 'consequentialistic preferences' with characteristics similar to what we denominate here belief-dependent lie aversion, and another of 'deontological preferences' similar to pure lie aversion. Although some results tend to lend weight to the idea of belief-dependent lie aversion, a test of such model in the repeated sender-receiver game is hindered by the fact that it predicts multiple equilibria. In the conclusion, we discuss how some of our results could help to understand dynamic play in repeated games like theirs.

Axiom 2 (continuity) *Consider two terminal nodes, one with associated vector* $[x_S, D_S]$ *and another with* $D_S^* \neq D_S$. *For any* x_S, D_S, *and* D_S^*, *there exists some nonzero amount of money* Δ *which depends on* x_S *and* $(D_S^* - D_S)$, *such that the sender is indifferent between the terminal nodes if she gets* $x_S + \Delta$ *at the node with deviance* D_S^*.

We remark that people may be heterogeneous regarding Δ, a point that we clarify further in Sect. 3. The final axiom makes explicit the negative connotation of the term 'deviance'. Intuitively, a lie evokes stronger (negative) feelings as the deviation from what one thinks was expected increases:

Axiom 3 (deviance aversion) *Given two terminal nodes* z *and* z' *with identical monetary payoff* x_S *and associated deviances* $D_S(z)$ *and* $D_S(z')$, $z \succ z'$ *if* $D_S(z) < D_S(z')$.

To highlight the differences in predictions exploited by our experimental design, it is worth comparing (a) belief-dependent lie aversion with (b) guilt aversion (as in Charness and Dufwenberg 2006). Both correspond to psychological games, in that preferences depend on beliefs. Yet the beliefs in theory (b) are about *the receiver's payoffs*, while in theory (a) they are about *the sender's actions*. In (b), furthermore, the lie itself is not normative: a lie can increase, decrease, or leave guilt unchanged, depending on how it affects payoff expectations. In (a), by contrast, a player feels badly for lying, albeit less badly when she believes that the lie is expected.

Finally, we can also compare theory (a) and pure lie aversion. Applying our notation, lie aversion assumes that preferences depend on the vector $[x_S(z), I(z)]$, and posits $[x_S(z), 0] \succ [x_S(z), 1]$ for any $x_S(z)$. Thus, the main difference between theory (a) and pure lie aversion is that the latter does not depend on the agent's beliefs, but only on whether he/she lies.

3 Experimental design and procedures

Our design uses a one-shot decision problem where pure lie aversion makes a prediction different than such theories as altruism, belief-dependent lie aversion, guilt aversion and selfishness, thus permitting us to estimate the relevance of pure lie aversion without these confounds. In the experiment, a message *sender* privately observes a random signal on the computer screen (more precisely, a green or a blue circle) and then must choose a message for the *receiver*. Two messages are always possible: 'The green circle has appeared' or 'The blue circle has appeared'. Monetary payoffs are as follows: the sender gets 15 Euros if he announces the green circle and 14 if he announces the blue one, whereas the receiver always gets 10 Euros. Note (i) the sender's payoff depends on the message sent, not on the realization of the signal, (ii) the sender faces a dilemma between honesty and material interest if the signal happens to be blue, as telling the truth is then costly, and (iii) the receiver does not observe the realization of the random signal and hence cannot verify whether the mes-

sage received is false or true, but does know the payoff set.[4] The experiment has two treatments (High and Low); they differ only in the (common knowledge) probability of the blue circle, which is 0.25 in Low, and 0.75 in High.

We ran 20 computerized sessions (10 High and 10 Low) at the Universidad Autónoma de Madrid, with a total of 258 participants. The sessions were conducted in two waves, the first (106 subjects) in November 2010, the second between September and October 2011. The software used was z-Tree (Fischbacher 2007). Participants were students from different disciplines, and the distribution of disciplines was similar in both treatments (Chi-square $(7) = 7.583$; $p = 0.371$).[5] Participants were not students of the experimenters. After being seated at a visually isolated computer terminal, each participant received written instructions that described the decision problem (see the appendix). Subjects could read the instructions at their own pace and we answered their questions in private. Understanding of the rules was checked with a control questionnaire that all subjects had to answer correctly before they could start making choices.

The instructions attempted to diminish potential demand effects or other confounds. For instance, we stated that this was an experiment on decision-making and that 'there are no tricky questions, you must simply choose as you prefer'. A potential motivation by any subject to behave so as to 'please' the experimenter, therefore, arguably put no constraints on her choice. Additionally, the instructions did not use language pertaining to truth or lies, or contain any indication to be truthful. Finally, subjects could ascertain that their choices would have a small effect on the aggregate, as they were informed that the total number of participants would exceed 40. Any potential motivation to increase the overall rate of truth-telling, therefore, should not play a major role in our setting.

Participants were anonymously matched in pairs. Before their roles (sender/receiver) were randomly determined, all chose as if they were senders. Since the receivers are totally passive, this cannot affect their choices afterwards.[6] We used the strategy method to elicit the decision; that is, before knowing the actual realization of the random signal, subjects indicated what message they would send for each contingency

[4]This distinguishes our study from 'deception games' (e.g. Gneezy 2005), in which the receiver does not know the payoff set. In those studies, this ignorance eliminates the concern that the sender's decision is influenced by his knowledge that the receiver knows whether the action was harmful, leaving the harm intact. In our study, this concern does not arise because the receiver is not harmed by the sender's choice. Furthermore, the receiver's knowledge of the sender's incentives plays an integral part in the experimental treatments, as these incentives induce the expectations we attempt to manipulate with our (High and Low) treatments, described below.

[5]This is relevant because, as we show elsewhere (López-Pérez and Spiegelman 2012), there is a correlation between honest behavior and the subject's discipline. Since the distribution of disciplines is similar in both treatments, we can be sure that any potential difference in behavior across treatments is not due to differences in the subjects' studies. Note also that there were not significant differences across treatments in the average values for political position ($p = 0.683$; Mann-Whitney test), gender ($p = 0.452$), or religiosity ($p = 0.165$).

[6]One could think of an alternative design in which the senders send messages to the experimenter, and hence there is no need for the receivers. In this case, however, belief-dependent lie aversion predicts that the subjects' second order beliefs about the experimenter's expectations should affect their decision. Controlling for such beliefs could be difficult. In addition, the degree (and relevance) of lie aversion could depend on the status of the recipient.

(blue/green).[7] This method maximizes the amount of data gathered, provides information on the chosen message profiles which facilitates the test of the theories, and permits the elicitation of subjects' beliefs in a manner that facilitates comparisons across treatments.

We elicited two beliefs from all subjects immediately after they had indicated the messages they would send. First, we asked each subject to estimate the percentage of all subjects who sent the message 'green' when the signal was blue—in other words, their expectations about deception when the signal was blue. We will refer to this number in what follows as a subject's first-order belief. Second, we asked each subject to estimate the average percentage estimated by all subjects in the previous question. We call this estimation a subject's second-order belief, and use it to approximate the belief $\mu(z)$ upon which belief-dependent lie aversion is taken to depend. Both first- and second-order beliefs were elicited in an incentive-compatible manner, as we paid 3 Euros when the absolute error was less than or equal to 5 percentage points.[8] Belief elicitation could not affect the choices, since all choices were made before beliefs were mentioned. Only after beliefs were elicited, one subject in each pair was randomly selected as the real sender, the other as receiver. The color of the circle was generated based on the relevant probabilities (High or Low), the actual sender informed of the color, and the message previously selected by the sender sent to the receiver. Afterwards, subjects answered a brief questionnaire which included some socio-demographic information and a question about their reasons for their message choice when the circle was blue. This ended the experiment. Subjects were paid in private by an assistant who was not informed about the details of the experiment. Each session lasted approximately 40 minutes, and on average subjects earned 12.70 Euros.

4 Discussion

This design isolates the effect of pure lie aversion, as this theory makes a different prediction than other potential motivations. To appreciate this, consider the predictions by each motivation. First, it is clear that a selfish sender will always announce 'green'. Since the receiver's payoff is not affected by the sender's choice, second, there is no altruistic reason ever to announce 'blue'. Third, guilt aversion cannot explain any 'blue' messages either because it is common knowledge that the receiver's payoff is constant across messages, as are, therefore, the sender's second-order expectations. By contrast, a (pure) lie-averse sender will tell the truth in any treatment—i.e., announce the observed color of the signal—if the utility cost of lying is large enough.

[7] In principle, the strategy method might induce different behavior than the specific-response method, where participants know the realization of the signal. Yet a control treatment to check for possible effects of the strategy method showed no significant effect (the data is available in a web appendix at http://www.uam.es/raul.lopez). We also note that Brandts and Charness (2011) review the experimental studies that use both methods and find no treatment differences in most of them.

[8] First-order (second-order) beliefs were paid only if the subject was later selected as a receiver (sender). We did this in order to avoid payoff asymmetries. Our belief-elicitation protocol is simple and rather easy to describe in instructions, and is not marred by any hedging problem.

Given the payoff constellation in our study, it seems safe to assume that this will be the case for some types.[9] Further, the probability of the appearance of the blue signal is irrelevant for a lie-averse player, so that this theory makes the following prediction:

Prediction LA (lie aversion) The rate of truth-telling will be the same across treatments. Deriving predictions for belief-dependent lie aversion requires us to specify beliefs. These can depend only on the message, of course, not on the color of the circle. Denote respectively by μ^{B} and μ^{G} the second-order belief of a lie upon 'blue' and 'green' messages and assume that both senders and receivers have correct expectations. Since sending the 'blue' message on the green signal is clearly suboptimal, it follows that $\mu^{\mathrm{B}} = 0$ and $\mu^{\mathrm{G}} < 1$, so that a lie after the blue signal will imply a nonzero deviance. To precisely derive μ^{G} we take into account Axioms 2 and 3 which imply that senders will respond honestly if the monetary gain from lying is less than some sum of money Δ, which falls with μ^{G} and is allowed to vary across individuals given μ^{G}. Given a monetary gain of 1 Euro from lying, we denote by $f(\mu^{\mathrm{G}})$ the strictly positive, non-decreasing, and continuous function representing the *fraction* of people who lie as μ^{G} changes.[10] Assuming that there is some measure for whom $\Delta = 0$ (i.e., who experience no lie aversion) is enough to imply $f(0) > 0$. Taking $\mu^{\mathrm{B}} = 0$ into account, and further denoting with p_{B} the probability of the blue signal, it follows by Bayes' theorem that the chance that message 'green' is a lie is defined by

$$\Pr\big[blue|\text{“green”}\big] = \frac{f(\mu^{\mathrm{G}}) \cdot p_B}{1 - p_B + f(\mu^{\mathrm{G}}) \cdot p_B}. \tag{1}$$

Given the assumption of correct expectations, the predicted belief μ^{G}—and corresponding level of dishonesty—is defined implicitly by

$$\mu^* = \frac{f(\mu^*) p_B}{1 - p_B + f(\mu^*) p_B} \tag{2}$$

One can show that there exists at least one non-zero solution to (2), and that the solution increases with p_B.[11] This means that the perceived deviation associated with lying should fall with p_B, therefore increasing the lie rate. The following prediction summarizes the discussion above:

Prediction BDLA (belief-dependent lie aversion) The rate of truth-telling will be positive in both treatments, but lower in treatment High; in other words, the High treatment will display the highest rate of lying.

---

[9] If the cost of telling the truth was higher, many lie-averse types could decide not to tell the truth. We would be unable, therefore, to provide an accurate estimation of the percentage of subjects who dislike lies—i.e., of the relevance of pure lie aversion.

[10] See the web appendix for a more detailed example.

[11] Consider the function $h(\mu) = \mu - \mu \cdot p_B + (\mu - 1) \cdot f(\mu) p_B$. Given our assumptions, this function is continuous and such that $h(0) < 0$; $h(1) > 0$. The intermediate value theorem therefore implies the existence of some μ^* such that $h(\mu^*) = 0$. Note also that uniqueness of μ^* is ensured for instance if $h' > 0$, which imposes some restrictions on the distribution of Δ (we clarify this point further in the web appendix). If the distribution is such that uniqueness does not hold, we assume that individuals coordinate their beliefs on the highest μ^*, so that prediction BDLA below is still satisfied.

Table 1 Percentage of choice of each strategy in each treatment

Treatment	Strategies				Total
	(G, G)	(G, B)	(B, G)	(B, B)	
High	46.97 %	39.39 %	2.27 %	11.36 %	100 %
Low	54.76 %	38.10 %	2.38 %	4.76 %	100 %
Aggregate	50.78 %	38.76 %	2.33 %	8.14 %	100 %

Note: $N = 132$ and 126 in treatment High and Low, respectively

Note that there are two main intuitions behind this result: (i) lying by the sender is more likely if she believes that the receiver expects a lie with high probability, and (ii) the receiver expects a lie with higher probability in High because the blue signal is more likely in that treatment, and therefore there are more occasions to get a larger payoff by lying after the blue signal (recall that lies only make sense when the signal is blue; otherwise they are costly).

5 Results

A sender has four possible pure strategies in the experiment. Denoting them as the message sent upon seeing a green (G) and blue (B) circle, respectively, they are: 'payoff maximizing' (G, G); 'honest' (G, B); 'mythomaniac' (B, G); and 'payoff minimizing' (B, B). Table 1 indicates the percentage of subjects who played each strategy in each treatment (High/Low), and in aggregate.[12] The most frequent choices in both treatments correspond to strategies (G, G) and (G, B), while other strategies are much less frequently chosen.[13]

Our design includes two controls to discriminate between pure and belief-dependent lie aversion. For the first control, let $f(s)_T$ denote the frequency of choice of strategy s in treatment T ($T = \mathrm{H}, \mathrm{L}$). According to belief-dependent lie aversion only strategies (G, G) or (G, B) should be chosen, and moreover (Prediction BDLA) the null hypothesis $f(\mathrm{G}, \mathrm{G})_\mathrm{H} \leq f(\mathrm{G}, \mathrm{G})_\mathrm{L}$ should be rejected in favor of the alternative $f(\mathrm{G}, \mathrm{G})_\mathrm{H} > f(\mathrm{G}, \mathrm{G})_\mathrm{L}$. As Table 1 indicates, this will not be possible. Pure lie aversion, in contrast, predicts no difference in the rate of choice of the strategies (G, G) or (G, B) across treatments. None is found: a Mann-Whitney test fails to reject hypotheses $f(\mathrm{G}, \mathrm{G})_\mathrm{H} = f(\mathrm{G}, \mathrm{G})_\mathrm{L}$ ($p > 0.2$), and $f(\mathrm{G}, \mathrm{B})_\mathrm{H} = f(\mathrm{G}, \mathrm{B})_\mathrm{L}$ ($p > 0.8$).

As a second control, we can use the subjects' elicited first and second-order beliefs about deception. We start by assuming that beliefs correctly anticipate behavior. Since belief-dependent lie aversion predicts a higher rate of choice of (G, G) in the High treatment, it correspondingly predicts higher expectations of deception in that treatment. Lie aversion, in contrast, predicts no difference in behavior or—therefore—in beliefs across treatments. Table 2 reports data about subjects' average

[12] We pool the data from the two waves of subjects (November 2010 and September–October 2011), as they are statistically identical in terms of their strategy choices. A Chi-square analysis of the joint distribution fails to reject independence (d.f. = 3; stat = 2.637; p-value = 0.451).

[13] None of the theories so far considered in this paper can explain why some small fractions of the subjects chose the payoff minimizing strategy (B, B) or the mythomaniac one (B, G) in both treatments. We discuss this issue later.

Table 2 Average beliefs about deception in each treatment

	Treatment		Total ($N = 258$)
	High ($N = 132$)	Low ($N = 126$)	
	Mean (S.E.)	Mean (S.E.)	Mean (S.E.)
Average first-order beliefs *'What percentage of subjects will send the green message on seeing the blue light?'*	69.8 (2.50)	69 (2.61)	69.4 (1.80)
Average second-order beliefs *'What will be the average answer to the question above?'*	70.4 (2.28)	70.6 (2.30)	70.5 (1.62)

Note: S.E. = Standard error of the mean

Table 3 Summary statistics on second-order beliefs, according to strategy chosen

Behavior	*N*	Mean	SE
Honest (G,B)	100	59.20	2.69
Minimizer (B,B)	21	52.81	6.32
Blue message overall	*121*	*58.09*	*2.48*
Maximizer (G,G)	131	82.48	1.58
Mythomaniac (B,G)	6	59.33	10.78
Green message overall	*137*	*81.47*	*1.62*
Total	258	70.50	1.62

beliefs in each treatment and in aggregate; these are remarkably constant at around 70 %. A Mann-Whitney test indicates that neither first-order ($p = 0.81$) nor second-order ($p = 0.97$) beliefs are significantly different.[14] Hence, the evidence seems inconsistent with belief-dependent lie aversion and more in line with pure lie aversion.

What if we drop the standard assumption that priors (i.e., beliefs) are common and correct, and assume instead that subjects have heterogeneous priors and play optimally given their own priors? In this case, belief-dependent lie aversion predicts that the sender's decision to lie after the blue signal should depend positively on his belief that the co-player expects him to lie. According to this theory, therefore, a participant in any treatment who reports a high second-order belief about deception is more likely to send message 'green' after the blue signal. Our data in Table 3 are emphatically in line with this: Pooled across treatments, subjects who planned to send the message 'green' when the circle was blue had on average a second-order belief of

[14]Similar tests also reveal that the two waves of subjects are identical on first-order expectations ($p = 0.838$) and second-order expectations ($p = 0.990$).

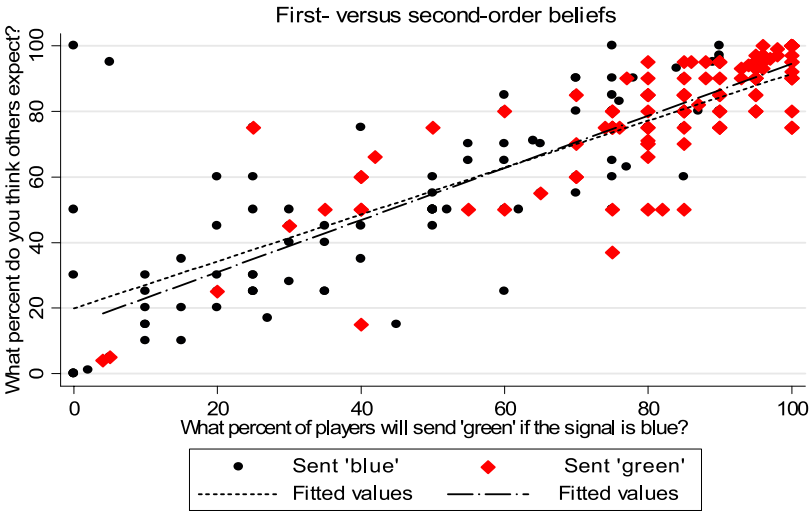

Fig. 1 Beliefs of each subject choosing message 'blue' or 'green' after the blue signal

deception of 81.5 percent; the value for subjects who sent message 'blue' was 58.1 percent.[15] This difference is significant (Mann-Whitney $p < 0.0001$).

Figure 1 provides further illustration of the correlation between beliefs and behavior. Each circle (diamond) in this figure represents a subject who sent message 'blue' ('green') after the blue signal in our treatments, placed according to her second- and first-order beliefs. We can see that the subjects choosing message 'green' are highly concentrated in the upper end of the scale, that is, second-order beliefs significantly predict the decision to lie after the blue signal. Figure 1 also shows by means of two regressions lines that first and second-order beliefs are correlated for each group of agents. Since we saw before that second-order beliefs are correlated with the decision to lie, it follows that first-order beliefs are also correlated with the decision to lie. That is, people who expect many subjects to lie are also more likely to lie. For instance, pooled across treatments, subjects who chose the strategies (G, B) and (G, G) respectively expected that 54 and 84 percent of the participants would lie. This is again a significant difference (Mann-Whitney test, $p < 0.0001$). In summary, therefore, our data indicates a clear correlation between beliefs and behavior (see also Lundquist et al. 2009 on this), but first-order beliefs already capture this relation, and second-order beliefs do not appear to provide much more insight.

It is worthwhile to note that, while a relation between beliefs and behavior is predicted by belief-dependent lie aversion, it is not incompatible with pure lie aversion, simply because the latter theory makes no definite prediction in this respect once one allows for heterogeneous beliefs between players. However, our results suggest that a theory of lie aversion might be complemented with some assumptions in this respect. For instance, in a context of heterogeneous priors, the data are consistent with the idea

[15] Note that the first percentage refers to the subjects choosing either (G, G) or (B, G), whereas the second one refers to the subjects playing either (G, B) or (B, B).

that people are averse to breaking norms of honesty, particularly if they do not expect others to break those norms, i.e., to lie (c.f. Cialdini et al. 1990; Bicchieri 2005; López-Pérez 2012).[16]

In suggesting that lie aversion and honest behavior interact with first-order beliefs, we are well aware of two caveats. First, even if the conjecture is correct, the measured association may not be valid. Participants in our experiment stated their beliefs after choosing their actions, and if those who lied themselves would 'prefer to believe' that others were lying, too, then they might come to believe such thing in order to (unconsciously) avoid cognitive dissonance. Note however that this effect should be at least attenuated by the payment for accurate beliefs. Second, the association between honesty and stated beliefs is vulnerable to the now relatively well-known argument about the (false) consensus effect (e.g. Ross et al. 1977). According to this hypothesis, people project their own behavior, attitudes and beliefs onto others, tending to overestimate the extent to which others act and think the same way they do. Thus subjects who do not lie tend to think others don't lie either, and by extension that others expect them not to lie. Hence the beliefs do not drive the action; rather, some personal characteristic drives both beliefs and actions, generating a spurious relationship. In this respect, the strong correlation between first and second-order beliefs (Spearman's $\rho = 0.827$, p-value < 0.0001) shown in Fig. 1 is consistent with the consensus effect.[17] Further research should clarify whether the relation between honest behavior and beliefs is just a result of this effect, or due to an interaction between lie aversion and beliefs.

We finish with a brief discussion of the behavior of the 8.14 % of subjects who chose the 'minimizing' strategy (B, B). One potential explanation of this behavior is that those subjects were trying to avoid the receiver's suspicion that they might be lying, perhaps because they expect that such a suspicion would bring disapproval (on disapproval-aversion see López-Pérez and Vorsatz 2010). Some evidence points in this line. To start, note that a disapproval-averse subject should choose (B, B) in High and (G, G) in Low, as the blue (green) message is more likely to be trusted in the High (Low) treatment. Hence, choice (B, B) should be more frequent in High. Indeed, the only significant effect that the different treatments seem to have induced is a (marginal) difference in the minimization behavior (Mann-Whitney test; p-value $= 0.053$). Second, we asked the subjects at the end of the experiment to write an open-form reason for their message choice when the signal was blue. Frequently, the justifications of those subjects choosing strategy (B, B) made reference either to the probability of the blue signal or to the other party. For instance, one subject choosing (B, B) in treatment High justified sending the blue message after the blue signal in the following manner: 'The color was actually blue and moreover the blue circle had a likelihood of appearance of 75 % so that participant B would consider me sincere with 75 % of probability'.[18] Other examples included 'the blue circle was

[16] If this were true, our design could underestimate the relevance of pure lie aversion. In effect, some people could be lie-averse but lie in our experiment because they expect most others to lie as well.

[17] There are two high-leverage outliers who report very low first- and very high second-order expectations. If these are removed, then the regression line is statistically indistinguishable from the 45-degree line.

[18] Note: The instructions referred to the sender/receivers as type-A/type-B participants.

the most likely to appear' [from treatment High], and 'since the green circle is most likely to appear, B would very likely think that the green circle would appear' [from someone choosing strategy (G, G) in treatment Low].

6 Conclusion

This paper reports the results from an experiment investigating whether pure lie aversion affects truth-telling. Our design allows us to prevent confounds with other potential motivations for truth-telling that have been considered in the literature. Participants in our design know that uttering a lie will increase their own monetary payoff and at the same time will inflict no harm on anybody, or affect anyone's payoff expectations. Further, they are isolated from each other: They do not know anything about their co-player and their decisions are anonymous. Nobody should tell the truth in our setting for altruism or to shape the receiver's payoff expectations. In contrast, people could tell the truth if they dislike lies. We consider two variants of this idea. Pure lie aversion predicts truth-telling if the cost is low, irrespective of other variables. Belief-dependent lie aversion, in turn, implies that people will tell the truth if it is not very costly and moreover they believe that others expect truth-telling from them. Our two treatments permit us to discriminate between these motivations.

Our main results are the following: (1) Overall, nearly 40 % of the subjects choose the strategy consistent with pure lie aversion; (2) we find no significant evidence for belief-dependent lie aversion; (3) there is a correlation between beliefs and honest behavior, so that people telling the truth expect a higher fraction of others to tell the truth as well. These results suggest that pure lie aversion is a widespread motive, possibly influenced by beliefs (as suggested by Bicchieri 2005), and have implications for understanding behavior. For instance, surveys are often used to explore societal trends, and questionnaires are also employed in experiments (such as the current one, for example). Responders often get no reward for their answers, so that one could expect them to answer in a random manner and hence consider their answers as simply 'hot air'. Yet our study suggests that *some* responders might tell the truth even if they suffer a small cost, so that truth-telling should arguably be more pronounced if it involves no cost (as in most surveys, if not all). Of course, other factors may affect responses. A desire for privacy or an aversion to disapproval from the people running the survey may lead to biased responses to sensitive questions. In addition, respondents may not perfectly recall the information the questions require. For instance, the reasons that subjects give to justify past decisions may be psychologically distinct from their motivations at the moment of action. While these factors should not be ignored, our results suggest at least that a complete disregard for surveys or questionnaires is not warranted.

Our results might help to understand previous experimental results. For instance, Peeters et al. (2007) note in a repeated game that some subjects tell the truth in most rounds but not always, which seems inconsistent with pure lie aversion. However, if the behavior of the lie-averse types depends on their first-order beliefs and these beliefs change with repetition, we could observe some lie-averse players who change their behavior accordingly. Finally, our results might also provide a benchmark for new experiments. In this respect, we propose three questions with which

we hope to suggest future experiments. First, how strong is lie aversion? Truth-telling in our study was cheap (just 1 Euro), would it decrease radically if its cost increases to, say, 5 Euros? Second, do people dislike telling lies or do they enjoy telling the truth (Sánchez-Pagés and Vorsatz 2009)? One can distinguish between these accounts in a slight variation of our basic decision problem where the sender also has the option to remain silent, in which case she gets 15 Euros (the receiver always gets 10). When the circle is blue, a lie-averse player would maximize her utility choosing silence, whereas a player who (sufficiently) enjoys telling the truth would choose the 'blue' message. Third, neuro-economic research (e.g. Zak 2011; Sommer et al. 2010) hints that moral cognition contains emotional (quick, instinctive and crude), as well as reasoned (slow, deliberative and sophisticated) components. It is conceivable that these components manifest themselves as different kinds of 'other-regarding' preferences. For instance, the pure aversion to lying seems instinctive and emotional, which could explain its prevalence. On the other hand, theories such as belief-dependent lie aversion, which require some interpretation of the co-player's expectations, may be part of the reasoned moral arsenal. One could hence think that such 'reasoned' factors have an effect only if conveniently primed by the context.

References

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.

Bicchieri, C. (2005). *The grammar of society*. London: Oxford University Press.

Bohnet, I., & Frey, B. (1999). Social distance and other regarding behavior in dictator games: comment. *American Economic Review*, *89*, 335–339.

Brandts, J., & Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, *14*, 375–398.

Buchan, N. R., Johnson, E. J., & Croson, R. (2006). Let's get personal: an international examination of the influence of communication, culture and social distance on other regarding preferences. *Journal of Economic Behavior & Organization*, *60*, 373–398.

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

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1990). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology Monograph Supplement*, *58*(6), 1015–1026.

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

Ellingsen, T., & Johannesson, M. (2004). Promises, threats, and fairness. *Economic Journal*, *114*, 397–420.

Erat, S., & Gneezy, U. (2012). White lies. *Management Science*. doi:10.1287/mnsc.1110.1449.

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

Fischbacher, U., & Heusi, F. (2008). Lies in disguise: an experimental study on cheating. Mimeo.

Gneezy, U. (2005). Deception: the role of consequences. *American Economic Review*, *95*(1), 384–394.

Hurkens, S., & Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Experimental Economics*, *12*, 180–192.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies*, *76*, 1359–1395.

López-Pérez, R. (2012). The power of words: a model of honesty and fairness. *Journal of Economic Psychology*, *33*, 642–658.

López-Pérez, R., & Vorsatz, M. (2010). On approval and disapproval: theory and experiments. *Journal of Economic Psychology*, *31*, 527–541.

López-Pérez, R., & Spiegelman, E. (2012). Do economists lie more? Mimeo.

Lundquist, T., Ellingsen, T., Gribbe, E., & Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization*, *70*, 81–92.

Orbell, J., Dawes, R., & van de Kragt, A. (1990). The limits of multilateral promising. *Ethics*, *100*, 616–627.

Peeters, R., Vorsatz, M., & Walzl, M. (2007). Truth, trust, and sanctions: on institutional selection in sender-receiver games. Mimeo.

Rosenthal, R. (2003). Covert communication in laboratories, classrooms, and the truly real world. *Current Directions in Psychological Science*, *12*(5), 151–154.

Ross, L., Greene, D., & House, P. (1977). The false consensus effect: an egocentric bias in social perception and attribution processes. *Journal of Experimental Social Psychology*, *13*, 279–301.

Sánchez-Pagés, S., & Vorsatz, M. (2009). Enjoy the silence: an experiment on truth-telling. *Experimental Economics*, *12*(2), 220–241.

Sommer, M., Rothmayr, C., Döhnel, K., Meinhardt, J., Schwerdtner, J., Sodian, B., & Hajak, G. (2010). How should I decide? The neural correlates of everyday moral reasoning. *Neuropsychologia*, *48*, 2018–2026.

Sutter, M. (2009). Deception through telling the truth!? Experimental evidence from individuals and teams. *Economic Journal*, *119*, 47–60.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica*, *76*(6), 1467–1480.

Zak, P. J. (2011). Moral markets. *Journal of Economic Behavior & Organization*, *77*(2), 212–233.