Jawer, Julia; Nielsen, Hedda; Weizsäcker, Georg

Attempting to detect a lie: do we think it through?

International Journal of Game Theory

Provided in Cooperation with:
Springer Nature

Suggested Citation: Jawer, Julia; Nielsen, Hedda; Weizsäcker, Georg (2025) : Attempting to detect a lie: do we think it through?, International Journal of Game Theory, ISSN 1432-1270, Springer, Berlin, Heidelberg, Vol. 54, Iss. 1, https://doi.org/10.1007/s00182-025-00930-w

International Journal of Game Theory (2025) 54:19
https://doi.org/10.1007/s00182-025-00930-w

ORIGINAL PAPER

Attempting to detect a lie: do we think it through?

Julia Jawer[1] · **Hedda Nielsen**[2] · **Georg Weizsäcker**[2]

Accepted: 9 November 2024 / Published online: 26 May 2025

Abstract
Game-theoretic analyses of communication rely on beliefs—especially, the receiver's belief about the truth status of an utterance and the sender's belief about the reaction to the utterance—but research that provides measurements of such beliefs is still in its infancy. Our experiment examines the use of second-order beliefs, measuring belief hierarchies regarding a message that may be a lie. In a two-player communication game between a sender and a receiver, the sender knows the state of the world and has a transparent incentive to deceive the receiver. The receiver chooses a binary reaction. For a wide set of non-equilibrium beliefs, the reaction and the receiver's second-order belief should dissonate: she should follow the sender's statement if and only if she believes that the sender believes that she does not follow the statement. The opposite is true empirically, constituting a new pattern of inconsistency between actions and beliefs.

Keywords Strategic information transmission · Lying · Higher-order beliefs

JEL classification D01 · D83

1 Introduction

A robust finding in social psychology is that most people are bad at lie detection. Success rates of identifying a truth as a truth, and a deception as a deception, are close to the level that one can achieve by flipping a coin (Bond and DePaulo 2006; Vrij 2008). In this paper, we focus on the second-order beliefs of those attempting to

✉ Georg Weizsäcker
weizsaecker@hu-berlin.de

Julia Jawer
jjawer@econ.au.dk

Hedda Nielsen
nielsenh@hu-berlin.de

[1] Aarhus University, Aarhus, Denmark

[2] Humboldt-Universität zu Berlin, Berlin, Germany

detect lies, the receivers of a message, as a possible factor in this weak performance. Judging a statement's veracity requires judging the sender's incentives, i.e., the extent to which he would benefit from each of his possible statements. Notably, this judgement is about the sender's *subjective* incentives—how much does he believe to benefit?—and it thus depends on his belief about his audience's reaction. That is, in a two-person setup the receiver should aim to predict what he, the sender, believes her, the receiver, to do in response to each possible statement.[1]

For deception games, where the sender wins if and only if he deceives the receiver, the logic of best response generates an interesting prediction about such beliefs: the receiver chooses her reaction to the statement so that it does not concur with her second-order belief. If she believes that the sender believes that she trusts his statement (hence, that he can deceive her with a lie), then she should not trust it. Conversely, if she believes that he believes that she does not trust the statement, then she should trust it. We call this prediction *belief rationality* and test it in a two-player communication game, finding it to be violated for the receivers in the game: their empirical correlation between second-order beliefs and choices is significantly positive (Spearman coefficient of 0.2), not negative.

This pattern not only contradicts belief rationality but is also suboptimal in the sense of missing that the participants acting as senders in our game, in fact, aim to exploit credulity: senders who believe that the receiver is likely to trust show a far higher rate of lying, with a difference of two thirds of the lying rate of the remaining group of senders. This difference is lost on the receivers—a player-role specific inability to take another person's perspective, and to think it through.[2]

This result describes a new kind of inconsistency, between second-order beliefs on the one hand, and actions and first-order beliefs on the other hand. For the wider research on communication and language, this makes for a bit of a can of worms: are such belief measurements informative, and should researchers measure second-order beliefs? Asking perhaps more provocatively, should they base their analyses on the concept of beliefs at all? If one observes such strong inconsistencies in respondents' belief statements, why rely on them?[3]

One cannot resolve such a question for all cases, instead it is better discussed for specific communications, together with an assessment of the generality of the discussion, including its limits. We make two observations with respect to the game that we analyze. First, it is a one-shot game with inexperienced players. That is, the participants in our experiment have no possibility to learn or to base their beliefs

[1] For an extensive classification of belief hierarchies in communication, see Weizsäcker (2023).

[2] The pattern is 'player-role specific' in that the senders in our game have a more accurate view of the correlations between the receivers' actions and beliefs.

[3] A related methodological discussion between empirical linguists and psychologists/economists is about the use of laboratory studies, versus naturally occurring speech. For brevity, we do not delve on this discussion here, but merely note that the question of whether or not to use belief-based analyses evolves around a similar trade-off: that between the use of controlled, "stylized" environments versus the use of real-world, "rich" language.

on past data. Our results are about initial behaviors and initial beliefs.[4] Second, we point out that our game is an unusually well-defined communication game in the sense that it has simple, binary actions spaces—e.g., only two possible statements for the sender—and that the propositions which the beliefs refer to are highly transparent: the possible statements have a truth status, meaning that they can be true or false, and the players of the game have no doubt about this fact or about its relevance. In particular, all messages in our game have a clear conventional meaning. The receiver's first-order belief report, assessing the statement, refers to the statement's truthfulness and has a transparent connection to the receiver's action—to follow the statement or not. Likewise, the sender's relevant first-order belief is clearly about this action of the receiver. Like in comparable experiments with belief elicitation (e.g., salesman-pitch games of Sheremeta and Shields (2013) or promise games of Charness and Dufwenberg (2006)), the imposition of belief measurements and the corresponding analysis does not, arguably, interfere much with the understanding of the game itself. That is, we use a game where asking for belief reports is innocuous and relatively unlikely to change the participants' thinking.

One may question, however, the usefulness of examining *second*-order beliefs. Although they are in heavy use in theories of language, including the basic versions of relevance theory (Sperber and Wilson 1995) or Gricean meaning, one may ask whether the interlocutors have any mental representation of second-order beliefs, and how they relate to first-order beliefs and to choices. Here, again, we argue for the focus on specific games. The question is clearly an empirical one and it is therefore useful to provide evidence on it, data set by data set.

The evidence in this paper gives reason to be cautious, but it also holds some good news for proponents of belief-based analyses: the receiver's choice tends to be highly consistent with her first-order belief about the truth. This consistency is *larger* than in normal-form games where the first-order belief is about an opponent's action (Costa-Gomes and Weizsäcker 2008; Rey Biel 2009; Polonio and Coricelli 2019). Our paper's novel type of belief rationality violation arises between beliefs of lower versus higher order, and it arises only for one of the two players: the receiver's belief about the truth tends to be inconsistent with her belief about the sender's first-order belief. The connection between beliefs of first and second order was not a focus of previous studies on players' beliefs in games. The more specialized literature on beliefs in communication experiments has, with only one exception that we are aware of, not even made any numeric measurements of the receiver's second-order beliefs—perhaps due to methodological concerns.[5] This highlights, again, the need for an appropriate data context with a sufficiently simple and transparent game.

[4] In the experiment, the participants also switch their roles between sender and receiver, after their first interaction. The analysis that we refer to in this Introduction was specified in our pre-plan and uses only the data generated in the participants' first roles. Appendix A describes the data that are generated in the participants' second roles, showing only a weak inconsistency between the receivers' actions and second-order beliefs (Spearman coefficient of 0.05).

[5] The exception is Agranov et al. (2024) who study differences in second-order beliefs about truth telling in market exchange with and without competition. In experiments on games without communication, second-order beliefs are measured more widely. See Manski and Neri (2013) and the discussion in Schotter and Trevino (2014). In more applied fields of economics, measurements of second-order beliefs have recently become influential as measures of social norms (e.g., Bursztyn et al. 2020).

Table 1 Payoff tables

	Table A			Table B	
	Option A	Option B		Option A	Option B
Sender	4	12	Sender	12	4
Receiver	12	4	Receiver	4	12

We build our measurement on a game that has proven to be suitable, by Peeters et al. (2015).[6] Their game is constant sum and makes the sender's motive to deceive highly transparent to the participants. It uses a particular feature to facilitate belief measurements: the game is mirror-symmetric in the sense that a lie about one state of the world is the mirror image of a lie about the other state of the world. Therefore, and under the mild assumption of label independence of the strategies, asking very few questions per participant is sufficient for a full elicitation of a belief hierarchy, up to second-order beliefs. This is further explained in the next section, with details on our experimental design. Subsequent sections contain, respectively, our (pre-registered) hypotheses, the experimental results and a concluding discussion that examines possible explanations of our findings.

2 Experiment

Two players, sender and receiver, interact anonymously in a one-shot fashion. The sender knows which of two equally probable states of the world, A or B, occurs. Each state corresponds to a payoff table, reproduced in Table 1. The sender sends a message indicating the state of the world: either he announces "Table A has been selected", or "Table B has been selected". The receiver reads the message and chooses a column in the payoff table (which is unknown to her), either Option A or Option B. Payoffs are such that the receiver wants to learn the truth and the sender wants her to miss it: one and only one player wins the game—i.e., obtains the high payment—and the receiver wins if and only if she matches her choice to the identity of the table. To aid the transparency of deception incentives, the instructions are explicit in raising the possibility that the sender can send a non-truthful message at his own will. All of these procedures are equivalent to those in the original experiment by Peeters et al. (2015).[7]

[6] The authors of the original study do elicit second-order beliefs, but only for the sender (owing to the different nature of their research question). Their existing measurements provide us with benchmarks for our results regarding the variables that we duplicate, while allowing for the inclusion of a new variable—the receiver's second-order belief.

[7] The only substantive differences between the design of Peeters et al. (2015) and ours, apart from our elicitation of the receiver's second-order belief, is that we have a slightly larger number of observations and that we use different payments for the belief variables.

In addition to playing the game, participants report their first-order beliefs and their second-order beliefs. As they act in different player roles, these beliefs are role specific. For first-order beliefs, the sender indicates his subjective probability of the event that the receiver follows the message—choosing the option with the label that is indicated in the sender's message—and the receiver indicates her subjective probability of the event that the message corresponds to the truth. For second-order beliefs, each player reports his or her subjective expectation of the opponent's answer to her or his first-order question.[8] The payment for the belief tasks rewards accuracy via the Binarized Scoring Rule (Hossain and Okui 2013)—paying 12 euro or 4 euro with probabilities that depend on belief accuracy—and the instructions indicate the rule's incentive compatibility in a way that is transparent while giving the participants the option to skip over the details.[9]

As mentioned earlier, the fact that the game is symmetric in labels A and B is key. Due to this feature, it suffices to elicit all responses under one out of two scenarios and impute responses for the other scenario by using the mirrored label for all states, messages and choices.[10] Like in the experiment of Peeters et al. (2015), the instructions explain this property in simple words. The procedure is equivalent to a full elicitation of all possible scenario-contingent actions and the beliefs about them, under the assumption that strategies and beliefs are label independent.

Across 12 sessions, we recruited a total of 251 participants and matched them in pairs.[11] Like in Peeters et al. (2015), each participant plays the game once in each role, for a total of two games, with fixed partners but without any feedback after the first game. The chronological order of playing the two games is randomized across participants. All payoffs in the tables are in euro amounts. For each participant, one of the two games is randomly selected as payoff relevant at the conclusion of the experiment, and one of three tasks is paid: the actual game, the first-order belief

[8] Full instructions are available in Appendix B. The choice of the precise belief hierarchies that are elicited corresponds to established practice in the literature, see e.g. the survey in Weizsäcker (2023). Note that the second-order belief is elicited as a point belief. For full generality of the belief hierarchy, this belief would be a distribution over the possible distributions that the first-order belief can assume. However, with only two payoffs in the game, the maintained assumptions of the next section guarantee that the mean of the distribution suffices to predict behavior.

[9] 19% of participants open the detailed scoring rule information and look-up does not correspond to significant differences in first- or second-order beliefs in our main sample.

[10] We ask for the sender's message in the scenario that A is the state of the world, and impute his message for the case that the state is B. For the receiver's action we ask what option she chooses if she receives the message that Table A was selected, and impute her choice for the scenario that the message indicates B. For all belief statements, we ask for the beliefs about the other player's behavior under the scenario that is used in the instructions, and impute the beliefs for the other scenario as the correspondingly mirrored beliefs.

[11] In sessions with odd numbers of participants, one person's choices were payoff irrelevant for the other participants, although he/she obtained payments as if matched with one of the others. In the pre-registration, we indicated a maximum of 245 participants. Turnout was slightly higher than expected and we decided to work with all data before looking at them. The experiment was programmed in z-Tree (Fischbacher 2007) and conducted at the WZB-TU experimental laboratory in November 2022. Participants were recruited through an online database using ORSEE (Greiner 2015). The data set is available at https://www.ifo.de/en/ebdc-dataset/experimental-data-attempting-detect-lie-do-we-think-it-through and via DOI https://doi.org/10.7805/jnw-sob-2025.

task, or the second-order belief task. The payment occurs in addition to paying a participation fee of 6 euro per participant.

3 Hypotheses

Notating the two players' indexes by $\{s, r\}$, let the scenario of the sender be the state of the world, $\theta_s \in \{A, B\}$ and the scenario of the receiver be the message $\theta_r \in \{A, B\}$, with the interpretation that the sender knows the true state θ_s and the receiver hears the message "Table θ_r has been selected." The two players' families of actions are denoted as $a_i(\theta_i) \in \{0, 1\}$, for $i \in \{s, r\}$, where the action with value 1 is, in each case, the action that corresponds to the players' scenario: the sender tells the truth and the receiver follows the message. The players' first-order beliefs are $b_i^1(\theta_i) \in [0, 1]$ (a belief about a_{-i}) and their second-order beliefs are $b_i^2(\theta_i) \in [0, 1]$ (a belief about the mean of b_{-i}^1), for $i \in \{s, r\}$. The assumption of label independence, which we henceforth make, is that $a_i(A) = a_i(B) =: a_i$, $b_i^1(A) = b_i^1(B) =: b_i^1$ and $b_i^2(A) = b_i^2(B) =: b_i^2$, for $i \in \{s, r\}$.[12] To denote distributions of actions, $\sigma_i(a_i)$ describes the probability of player i choosing action a_i.

The experimental observations (including beliefs) that correspond to the game's unique Weak Perfect Bayesian Equilibrium are $\sigma_i(a_i) = b_i^1 = b_i^2 = \frac{1}{2}$, for $i \in \{s, r\}$. The equilibrium is uninformative, prescribing that sender and receiver each randomize with equal probability in each scenario. A more interesting case for the analysis arises for non-equilibrium beliefs, which leads to the milder set of assumptions that we call belief rationality: we continue to maintain the assumption that the players maximize their subjective expected utility (SEU) and that this is twice-mutually known by the players, but we relax the assumption that beliefs are in equilibrium. Since these assumptions imply that players maximize their subjectively perceived probabilities of receiving the high payment in the game, we can straightforwardly predict the connection between actions and first-order belief:

Hypothesis 1: $\sigma_s(a_s)$ decreases in b_s^1 and $\sigma_r(a_r)$ increases in b_r^1.

A remark is in order about the fact that the formulation of Hypothesis 1 is weaker than the corner solution that SEU predicts: if $b_s^1 > \frac{1}{2}$, then $\sigma_s(a_s) = 0$, and if $b_r^1 > \frac{1}{2}$, then $\sigma_r(a_r) = 1$. The reasons for the weaker formulation of the hypothesis are (a) that it covers non-degenerate choice frequencies in the experiment, and (b) that the weaker statement allows for possible alternative formulations of the players' objectives. For instance, adding the assumption of a random utility perturbation of player i's actions (in a way that is uncorrelated with b_i^1, as usually assumed in models of probabilistic choice) would be covered by the hypothesis. Also, many natural formulations of aversion against stating lies or expressing distrust (but retaining SEU) are covered by the hypothesis.

[12] Note that the assumption's two sets of restrictions on beliefs correspond, respectively, to mutual first-order knowledge and mutual second-order knowledge of the property that $a_i(A) = a_i(B)$, for $i \in \{s, r\}$.

Table 2 Data averages

	Actions		Beliefs			
	$\bar{a}_s$	$\bar{a}_r$	$\bar{b}_s^1$	$\bar{b}_s^2$	$\bar{b}_r^1$	$\bar{b}_r^2$
Sender	0.6667		0.5038	0.4446		
	(0.4733)		(0.2462)	(0.1890)		
	[0.0002]		[0.8160]	[0.0008]		
Receiver		0.6563			0.5477	0.5305
		(0.4768)			(0.2469)	(0.2772)
		[0.0004]			[0.0549]	[0.1870]

In (parentheses): standard deviations. In [brackets]: p-values of Wilcoxon signed-rank tests against a value of 0.5, two-sided

Under the assumption of belief rationality, the first-order belief is consistent with the second-order belief in the sense that each player expects the other player to act in congruence with Hypothesis 1:

Hypothesis 2: b_s^1 increases in b_s^2 and b_r^1 decreases in b_r^2.

Part (b) of the above-made remark on the "weaker" hypothesis formulation applies to Hypothesis 2, too. For many possible formulations of the players' objectives, a change in a player's second-order belief may not induce a jump to the corner solution for the first-order belief, but the direction of the change applies nevertheless and is described by Hypothesis 2.

Combining the two previous hypotheses yields the connection between actions and second-order beliefs that we aim to test in this paper, chiefly for the receiver:

Hypothesis 3: $\sigma_s(a_s)$ decreases in b_s^2 and $\sigma_r(a_r)$ decreases in b_r^2.

4 Results

Table 2 shows the data averages of actions and beliefs in both player roles.[13] For each participant, we use only the observations from one of the two player roles: the one that he or she holds in his or her first game.[14] Average responses are denoted as $\bar{a}_s$, $\bar{b}_s^1$, etc., for the empirical means of the sender's action, her belief b_s^1, etc.

The table shows that about two thirds of the senders tell the truth and about two thirds of the receivers follow the message. While the best response of receiver participants to the behavior of sender participants would have been to always follow the

[13] For a visualization of the first- and second-order belief distributions, see Figs. 2 and 3 in Appendix A.

[14] This data restriction rules out order effects. Some order effects are detectable in our data set and our pre-plan specified, for this case, to include only the data from each participant's first game in the main analysis. All results retain their direction and significance when including all data, but the results when using the second game alone show only an insignificant inconsistency between the receivers' actions and second-order belief—see Appendix A.

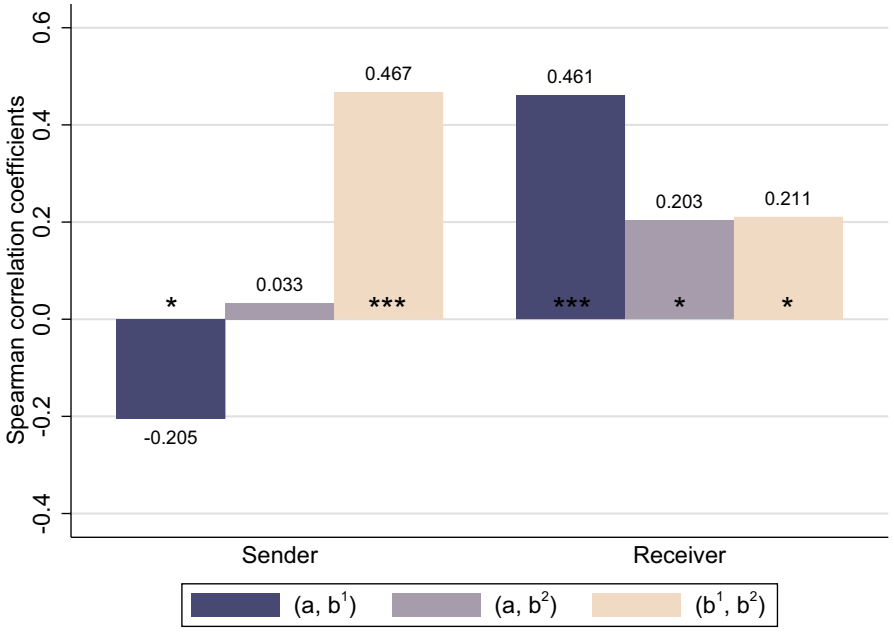

Fig. 1 Correlation coefficients. Spearman correlation coefficients between actions (a), first-order beliefs (b^1) and second-order beliefs (b^2), for both senders (first three bars) and receivers (last three bars). $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ in two-sided tests

message,[15] the average beliefs reported in the table appear to rationalize the actions fairly well; the receivers' average first-order beliefs are that senders tell the truth in about half of the cases, which makes the average receiver roughly indifferent between following and not following. These first-order beliefs, in turn, can be justified by second-order beliefs; on average, receivers predict roughly that the senders expect the receivers to trust the message with probability one half. Moreover, these second-order beliefs are quite accurate, on average.

Our main interest lies, however, in the correlation between the variables, in the sense of the previous section's hypotheses. Figure 1 reports Spearman correlation coefficients for each player role, also indicating their statistical significance in two-sided tests.[16]

[15] The observation that both senders and receivers tend to follow the "truth dafault" is consistent with a large set of observations in the lying literature (Bond and DePaulo 2006; Levine 2014). All of our measurements are consistent with, and indeed close to, those in Peeters et al. (2015). A noteworthy consequence of adhering to the truth default is that receivers earn more money in the experiment than senders: the expected payments that result from actions in the game are 7.17 euro for senders and 9.03 euro for receivers.

[16] Having pre-formulated and derived our hypotheses as directed hypotheses, we may consider one-sided tests as appropriate. However, the fact that some test statistics deviate from zero in the opposite direction makes it easier to interpret two-sided tests.

The correlations show that Hypothesis 1 is supported for both player roles, that Hypothesis 2 is supported for the sender but rejected for the receiver, and that Hypothesis 3 is neither rejected nor supported for the sender but rejected for the receiver. In particular, the results highlight a striking asymmetry between the belief hierarchies of the sender and those of the receiver. While the senders correctly anticipate a strong positive relation between the receivers' first-order beliefs and their actions (correlations of about 0.47 in each case), the receivers' views are off target; the correlation between b_r^1 and b_r^2 is significantly positive, whereas the correlation that it tries to predict, that of a_s and b_s^1, is significantly negative (both around 0.2, in absolute terms). The receivers fail to understand that senders tend to exploit the credulity of receivers.[17]

To assess the magnitude of the effect, we consider only those receivers whose second-order belief expresses a non-zero tendency—that is, we drop the 12% of receivers who report a second-order belief of $b_r^2 = 0.5$—and to ask how differently these receivers act depending on the sign of their tendency. For participants with $b_r^2 > 0.5$, the frequency of following the message is 72%, and for participants with $b_r^2 < 0.5$ (who should have a higher propensity to follow the message, according to Hypothesis 3) the average frequency of following is merely 60%. A corresponding separation of senders shows that, in fact, senders lie much more often if they expect the receiver to have a positive tendency to follow; senders with $b_s^1 > 0.5$ show a lying rate of 42%, versus only 25% for the senders with $b_s^1 < 0.5$.[18]

One may attempt to explain the violation of Hypothesis 3—the positive correlation between the receivers' actions and their second-order belief—by an illusion of transparency; perhaps the receivers believe (too much) that the sender can detect their inclination to follow the message or not. We point out, however, that this explanation fails to explain a key pattern: that the positive correlation between b_r^1 and b_r^2 violates Hypothesis 2. Any second-order belief, even an illusion-of-transparency-laden one, should generate a first-order belief that reflects the incentive to exploit the belief. Rather, the receivers' first-order beliefs tell the opposite story. The inconsistency of b_r^1 and b_r^2 is, thus, a separate phenomenon from the illusion of transparency.

5 Concluding discussion

A striking pattern of our results is that the sender appears to be "smarter" than the receiver. The inconsistency between second-order beliefs on the one hand, and actions and first-order beliefs on the other, appears for the receiver but not for the sender. The receiver does not appear to think the situation through and fails to predict that a sender who believes in the receiver's gullibility will exploit this. In contrast, the sender's beliefs do not show such a violation of belief rationality.

[17] In Appendix A, we show that these results are robust to using a nonlinear and multivariate choice model instead of correlations.

[18] For robustness, Table 5 in Appendix A shows the frequencies of actions corresponding to different cutoff values for b_s^1 and b_r^2, instead of 0.5 as in the main text. The result is generally robust to changes in the cutoff value.

The behavioral and experimental literature on biased beliefs in communication cannot, to our knowledge, explain this asymmetry. To the extent that this literature focuses on the role of second-order beliefs, it has an emphasis on lying cost and guilt aversion, as in Charness and Dufwenberg (2006), Kartik (2009) and Gneezy et al. (2018), or level-k reasoning as in Crawford (2003).[19] The hypotheses that we develop and test in this paper would, at least in spirit, also apply to many of these more general models. Recent models of biased beliefs in games include Cohen and Li (2023) and Fong et al. (2023) who extend the notion of cursed equilibrium (Eyster and Rabin 2005) to extensive-form games, including communication games. The theory of Cohen and Li (2023) would, if anything, tend to predict that the receiver's beliefs are more accurate than the sender's, not vice versa. It may be useful to further examine these and other theories with respect to the conditions under which the receiver may show an inconsistency of higher-order beliefs that the sender does not show. In the following, we aim to make this discussion more precise and discuss a potential reason for the sender-receiver asymmetry in our game.

The hypotheses that we test in the paper's previous sections invoke only the assumptions of subjective expected utility theory (SEU) and twice-mutual knowledge of it. Several of the above-listed models are therefore rejected in our data only by virtue of the fact that they, too, rely on these SEU assumptions.[20] A notable exception are theories of guilt aversion that use second-order beliefs differently: if the receiver expects that the sender expects her to follow him, but she does in fact not follow him, then her utility is reduced because she dislikes disappointing his expectation. This creates an incentive for the receiver to act in accordance with her second-order belief. However, we do not regard this as a plausible explanation for our results, for two reasons. First, we make the same observation as we made earlier when discussing the illusion of transparency: while the guilt-aversion argument can explain the violation of Hypothesis 3 (about actions and second-order beliefs), it would not explain the violation of Hypothesis 2 (about first-order beliefs and second-order beliefs). Second, we regard it as implausible that the guilt-aversion argument should apply more to the receiver than to the sender. The sender's action is salient in the game and not telling the truth is therefore a salient attempt to create false expectations: it is obviously a lie if one states that "Table A was chosen" if, in fact, Table A was not chosen. The receiver, in contrast, may perceive greater wiggle room to morally justify the decision not to follow the statement. Guilt aversion should therefore affect the sender more than the receiver, in contrast to our findings.

The salience of not telling the truth is related to a relevant observation that was made in level-k analyses of deception by Crawford (2003): if the thought process of a player starts with the instinctive level-0 behavior that the sender is

[19] Level-k analyses presume that player can engage only in a limited, and typically rather small, number of iterations of best response behavior. An initial, "instinctive" level-0 behavior is assumed and a level-k player makes up to k cognitive steps of best responding to best responding.. to this level-0 behavior. For more references on level-k analysis in communication games, see Crawford et al. (2013).

[20] A full level-k analysis cannot, however, be applied to our game, as it does not make a prediction about how beliefs are stated for orders of beliefs that are higher than one's own k.

truthful and the receiver is credulous, then a pair of plausible level-1 behaviors would be for the sender to lie and for the receiver to follow the sender's message. This suggests that deception games, by their nature, create an asymmetry in "truth default" orientation between the two player roles, in level-k analyses and potentially also in more general approaches.

Following a similar logic, we offer an admittedly speculative reasoning for why senders and receivers show differences in their consistency of beliefs of first and second order. First, recall that in our game, the two players' pairs of beliefs (first- and second-order, for each player) are concerned with three unknowns: the state of the world θ_s, the sender's statement a_s, and the receiver's reaction a_r. The roles of θ_s and a_r are fully symmetric between the first-order beliefs of players—the receiver's first-order belief assesses $Pr(\theta_s = A | a_s = \text{“}A\text{”})$, while the sender's first-order belief assesses $Pr(a_r = A | a_s = \text{“}A\text{”})$—and the second-order beliefs show an analogous symmetry of roles for θ_s and a_r. In contrast, the statement a_s plays a distinct role, as all beliefs are conditional on it.

Now, observe that the state of the world θ_s is connected to a_s by the latter's conventional meaning: the statement literally expresses the state of the world's value. The conditionality of θ_s on a_s is therefore straightforward and the players are likely to internalize the correlation between a_s and θ_s with lower cognitive effort than for the correlation between a_s and a_r. We take this to be a possible driver of differences in consistency when reporting a pair of first-order and second-order beliefs. In particular, we may speculate that a player is more likely to report a pair of beliefs that satisfies belief rationality if the conditionality that appears in the relatively more complex belief—the second-order belief—is easier to grasp. If so, then a receiver who evaluates the correlation between a_s and a_r in her second-order belief encounters greater difficulty than a sender who evaluates the correlation between a_s and θ_s in his second-order belief. The statement's conventional meaning thus creates an asymmetry in the ease of expressing a "smart" pair of beliefs: the receiver cannot "think it through" as easily as the sender. This mirrors what we find in our data.

Finally, we note that this speculation about the ease of showing belief rationality does not contain a reason for the correlation between the receiver's beliefs of first and second order to have the *opposite* sign. For this effect, however, one may invoke the "truth default" again: if belief rationality is relatively difficult to grasp for the receiver, and it therefore disciplines her belief statements with little stringency, then it may take only a small degree of leaning towards truth/trust (in all beliefs) in order to flip the correlation's sign.

Appendix A: further results

Main sample: first game

We here provide additional results using data from the experiment's first game (i.e., using the main sample that was collected before role switching occurred in the experiment).

In Fig. 2 we visualize the combinations of first- and second-order beliefs reported by senders and receivers. As shown by the slope of the regression line, the correlation between the first- and second-order beliefs is positive for both senders and receivers, but stronger for senders—in line with Fig. 1.

Comparing the distributions of beliefs for senders who tell the truth to those who do not (see the different markers), we observe differences that are in line with the statements about means appearing in Sect. 4. Moreover, in Table 3, we provide the average first- and second-order beliefs conditional on the action chosen by the participant.

To examine the correlations of our experimental variables in a nonlinear choice model, Table 4 reports Probit regressions for a_s and a_r. It shows again that the correlation between the receiver's action and her second-order belief is positive, not negative. Moreover, the regressions show that if one controls for first-order beliefs, the second-order beliefs have

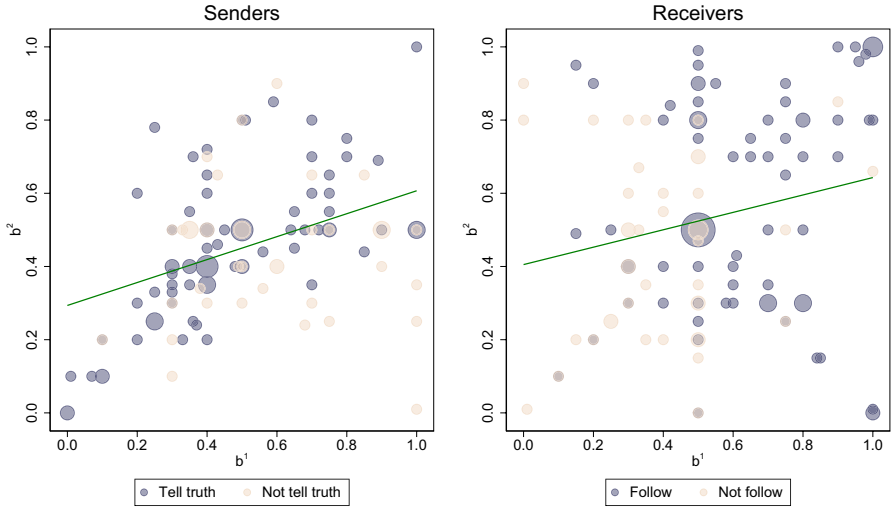

Fig. 2 First- and second-order beliefs. Combinations of first- (b^1) and second-order beliefs (b^2) for senders (left) and receivers (right). The size of the bubble corresponds to the number of participants who reported the particular combination of first- and second-order beliefs. The line represents a linear regression of second-order beliefs on first-order beliefs. Data include each participant's first game

Table 3 Data averages: split by action

	Proportion	Beliefs $\bar{b}^1$	$\bar{b}^2$
Sender			
Tell truth	0.6667	0.4671 (0.2423)	0.4495 (0.1959)
Not tell truth	0.3333	0.5773 (0.2400)	0.4349 (0.1761)
Receiver			
Follow	0.6563	0.5677 (0.2468)	0.5157 (0.2731)
Not follow	0.3438	0.5035 (0.2443)	0.5633 (0.2870)

Standard deviations in parentheses. Data include each participant's first game

Table 4 Probit regressions

	a_s			a_r		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	1.0282***	0.3170	0.7230*	−1.0048**	−0.1140	−1.3203**
	(0.2730)	(0.2954)	(0.3186)	(0.3334)	(0.2423)	(0.4312)
b_i^1	−1.1524*		−1.4688**	2.7333***		2.6142***
	(0.4746)		(0.5278)	(0.6329)		(0.6503)
b_i^2		0.2569	1.0414		1.0056*	0.7414
		(0.6122)	(0.7243)		(0.4085)	(0.4585)
Observations	123	123	123	128	128	128

Standard errors in parentheses. Data include each participant's first game. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 5 Action frequencies of senders and receivers

Cutoff	a_s		a_r	
c	$b_s^1 < c$	$b_s^1 > c$	$b_r^2 < c$	$b_r^2 > c$
0.1	1.00	0.66	0.67	0.66
0.2	0.88	0.65	0.64	0.66
0.3	0.94	0.63	0.46	0.70
0.4	0.75	0.63	0.57	0.70
0.5	0.75	0.58	0.60	0.72
0.6	0.72	0.56	0.61	0.73
0.7	0.69	0.58	0.58	0.82
0.8	0.70	0.50	0.60	0.81
0.9	0.68	0.50	0.62	1.00

Average values of a_s (truth telling) and a_r (following the statement), separated by beliefs strictly above versus below c

little predictive power. This is consistent with our maintained assumptions of SEU maximization, where the only way in which second-order beliefs correlate with actions is via their role as the basis for first-order beliefs.

Additional data: second game

The previous analyses focused on the data from each participant's first game. In this appendix we reproduce the main analyses, as well as the scatterplots from the previous section, using only data from each participant's second game. The senders featured here had previously taken on the role of receivers, while the receivers had previously taken on the role of senders. Therefore, both had the opportunity to think through the game from a different role's perspective.

Figure 3 provides a visualization of senders' and receivers' first- and second-order beliefs.

With respect to the data averages of actions and beliefs, the most significant shift occurs among the receivers. As seen in Table 6, only half of the receivers follow the

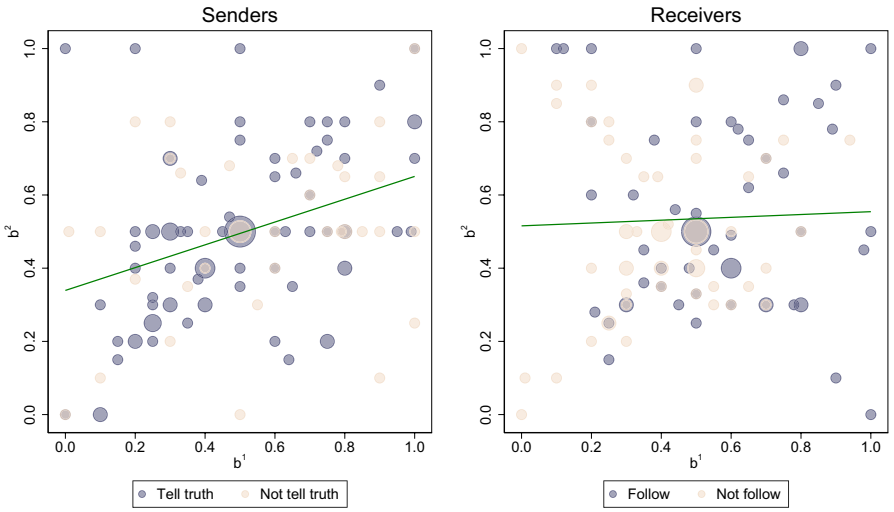

Fig. 3 First- and second-order beliefs. Combinations of first- (b^1) and second-order beliefs (b^2) for senders (left) and receivers (right). The size of the bubble corresponds to the number of participants who reported the particular combination of first- and second-order beliefs. The line represents a linear regression of second-order beliefs on first-order beliefs. Data include each participant's second game

Table 6 Data averages

	Actions		Beliefs			
	$\overline{a}_s$	$\overline{a}_r$	$\overline{b}_s^1$	$\overline{b}_s^2$	$\overline{b}_r^1$	$\overline{b}_r^2$
Sender	0.6875		0.5081	0.4863		
	(0.4653)		(0.2634)	(0.2227)		
	[0.0000]		[0.7674]	[0.4396]		
Receiver		0.5122			0.4859	0.5207
		(0.5019)			(0.2266)	(0.2370)
		[0.7868]			[0.4464]	[0.7347]

In (parentheses): standard deviations. In [brackets]: p-values of Wilcoxon signed-rank tests against a value of 0.5, two-sided. Data include each participant's second game

message, as opposed to two-thirds who did so without having previously considered the game from the sender's perspective.

In terms of correlations, we again see the most notable shift appearing among receivers. The correlation between a_r and b_r^2 is 0.046 and insignificantly different from zero, indicating that Hypothesis 3 is neither rejected nor supported for the receiver. Table 7 shows that for receivers who previously took on the role of senders, the correlation between their actions and second-order beliefs is on the wrong side of zero (like in the paper's main analysis) but insignificantly so. The correlation is insignificantly different from the corresponding correlation reported in the main text, using the data from each participant's first game. Table 8 provides the results of probit regression using the second game only.

Table 7 Correlation coefficients

	Sender				Receiver		
	a_s	b_s^1	b_s^2		a_r	b_r^1	b_r^2
a_s	1.000			a_r		1.000	
b_s^1	−0.144	1.000		b_r^1	0.324***	1.000	
b_s^2	−0.073	0.396***	1.000	b_r^2	0.046	0.063	1.000

Data include each participant's second game. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8 Probit regressions

	a_s			a_r		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.8333	0.5850*	0.7945*	−0.9614**	−0.1734	−1.1489**
	(0.2708)	(0.2811)	(0.3206)	(0.2914)	(0.2744)	(0.3847)
b_i^1	−0.6646**		−0.7022	2.0595***		2.0551***
	(0.4622)		(0.5070)	(0.5464)		(0.5460)
b_i^2		−0.1968	0.1191		0.3923	0.3641
		(0.5229)	(0.5840)		(0.4784)	(0.5088)
Observations	128	128	128	123	123	123

Standard errors in parentheses. Data include each participant's second game. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9 Data averages

	Actions		Beliefs			
	$\bar{a}_s$	$\bar{a}_r$	$\bar{b}_s^1$	$\bar{b}_s^2$	$\bar{b}_r^1$	$\bar{b}_r^2$
Sender	0.6773		0.5060	0.4659		
	(0.4684)		(0.2546)	(0.2075)		
	[0.0000]		[0.9362]	[0.0046]		
Receiver		0.5857			0.5174	0.5257
		(0.4936)			(0.2387)	(0.2579)
		[0.0066]			[0.3870]	[0.2136]

In (parentheses): standard deviations. In [brackets]: p-values of Wilcoxon signed-rank tests against a value of 0.5, two-sided. Data include both first and second games

Full sample: both games

Finally, in Tables 9, 10, 11, we reproduce our main analysis pooling the data from the first game and the second game. The results are qualitatively robust relative to the analyses that include the first games only.

Table 10 Correlation coefficients

	Sender				Receiver		
	a_s	b_s^1	b_s^2		a_r	b_r^1	b_r^2
a_s	1.000			a_r	1.000		
b_s^1	−0.174**	1.000		b_r^1	0.406***	1.000	
b_s^2	−0.023	0.421***	1.000	b_r^2	0.128*	0.139*	1.000

Data include both first and second games. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 11 Probit regressions

	a_s			a_r		
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.9174***	0.4580*	0.7627**	−1.0082***	−0.1621	−1.2725***
	(0.1935)	(0.2006)	(0.2247)	(0.2161)	(0.1806)	(0.2810)
b_i^1	−0.8819**		−1.0371**	2.4479***		2.3961***
	(0.3321)		(0.3705)	(0.4028)		(0.4096)
b_i^2		0.0046	0.5006		0.7290*	0.5616
		(0.3926)	(0.4535)		(0.3083)	(0.3350)
Observations	251	251	251	251	251	251

Standard errors in parentheses. Data include both first and second games. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix B: instructions

General instructions

Thank you for participating in this experiment. For completing the experiment, you will receive 6 euro. Additionally, you will have the possibility of earning up to 12 euro depending on your decisions and answers throughout the experiment.

Participation will take about 45 min. The experiment consists of four parts: instructions, the experiment itself, a questionnaire, and payment.

Please read through the instructions carefully to make sure that you have fully understood the task and the questions. At the end of the instructions you will be asked to answer a few short questions to check your understanding. If you answer a question incorrectly, the relevant parts of the instructions will appear in a separate window. Please read these parts again carefully and try to answer the question again.

Afterwards, you will complete the experiment - just once. After you have answered all the questions in the experiment, we will ask you to complete a short questionnaire. Once all participants have completed the experiment, your payment will be calculated and displayed to you. The payout will take place directly after the experiment. We ask you to remain seated until you are called for the payout.

Your answers and decisions are anonymous and will be treated confidentially and used exclusively for scientific purposes.

In order for the data from the experiment to be reliable, certain rules must be followed during the experiment. We would therefore ask you to switch off your cell phones and put them out of reach. We also ask you to only use the functions of your computer that are necessary for the experiment and otherwise refrain from using electronic devices.

Please restrict your communications during the experiment to questions about the experiment and address these questions only to the experimenters. If you have any questions after you have read the instructions, please speak quietly. Your question will then be answered personally and quietly. If a questions is relevant for all participants, the experimenter will repeat the question aloud and answer it.

If you disrupt the experiment or do not adhere to the above rules, we will unfortunately have to exclude you from the experiment and you will not be paid.

Experiment instructions

Procedure

In this experiment, we simulate an interaction between two individuals where Person *S* (*Sender*) sends a message to Person *R* (*Receiver*). Person *S* will have to choose one of two possible messages to send, and Person *R* will have to choose one of the two possible reactions to the message, which will later be called *Option A* and *Option B*.

At the start of the experiment, you will be randomly matched with another participant. Neither you nor the participant you are matched with will learn the identity of his/her match. Moreover, you will not learn your role until the end of the experiment. This means that you will have to make decisions as both Person *S* and Person *R*.

Table *A* and Table *B* represent two possible scenarios of the experiment. Which of the two is the relevant scenario is uncertain at the beginning of the experiment—it is selected at random by the computer. The numbers in the tables represent payoffs of Person *S* and Person *R*, in euro amounts.

	Table A			Table B	
	Option A	Option B		Option A	Option B
Person *S*	4	12	Person *S*	12	4
Person *R*	12	4	Person *R*	4	12

The payoffs depend on the table selected by the computer (*Table A* or *Table B*) and the choice made by Person *R* (*Option A* or *Option B*). Person *R* earns more money if and only if option and table coincide.

For example, if Person *R* chooses *Option A* and *Table A* is randomly chosen, Person *R* gets 12 euro and Person *S* gets 4 euro as a payoff. In case that Person *R* chooses *Option A* and *Table B* is randomly chosen, Person *R* gets 4 euro and Person *S* gets 12 euro as a payoff. And so on, for the two cases where Person *R* chooses *Option B*.

Note also: Following this payoff rule, Person *S*, who sends the message, is always better off if Person *R* chooses an option that does not coincide with the table selected by the computer.

The experiment proceeds as follows. The computer randomly selects one of the two tables with equal probability. Only Person *S* is informed about the table that has been selected by the computer. Person *S* chooses which of the following two messages to send to Person *R*:

- "Table A has been selected",
- "Table B has been selected".

Note, Person *S* is completely free in the choice of her message.

Person *R* observes the message from Person *S* and chooses one of the two options:

- Option A,
- Option B.

Note, also Person *R* is completely free to choose any of the two options.

After making choices about messages and options, both Person *S* and Person *R* are asked to answer two further questions. First, you are asked to estimate the other person's action. Second, you are asked to predict what the other person estimates your action to be. For both questions, the closer your answer is to the true value, the larger your payoff will be.

Altogether, in the course of the experiment, you will first have to answer three questions as Person *S* and then three questions as Person *R*.

Payment

Your payment from the experiment consists of two parts: a fixed payment (participation fee) of 6 euro and a variable payment that depends on your answers.

To determine the amount of the variable payment, the computer will randomly select your role (Person *S* or Person *R*) and one of the three questions corresponding to this role.

If the first question—the question about your choice (message for Person *S*, option for Person *R*)—is selected, the variable payment is calculated according to the payoff table. Depending on the random choice of the table (*A* or *B*) made by the computer and the choice of the option (*A* or *B*) made by Person *R* (either you or a participant matched with you), you will earn either 4 euro or 12 euro.

If the second question or the third question—the questions about expectations—is selected, the variable payment is calculated using a rule called the Binarized Scoring Rule. According to this rule, you can also earn 12 or 4 euro and the probability of earning a high payoff (12 euro) increases if your answer is closer to the true value; conversely, the probability of earning a low payoff (4 euro) increases if your answer is further from the true value. While you need not understand the exact algorithm of the Binarized Scoring Rule, a full description of it can be found below, in the section labelled "Supplement:...". All you have to know is that the rule makes it optimal for all participants to state their true beliefs in response to the second question and to the third question.

Supplement: Binarized Scoring Rule

This section describes the payment rule for the second and third questions. In these two questions you are not asked to make a choice as in the first question. Instead, you need to estimate the answer of another person on a scale from 0 to 100. These estimates show how likely you consider the other person to provide a particular answer. Here, their answer is either a choice or an estimate of the other person.

The expected payment for either of two questions increases with the accuracy of your estimate. The measure of the realized error in your estimate (subsequently referred to by the letter l) is calculated as follows: $l = ((x - \theta)/100)^2$, where x is your guess about the answer of another person matched with you, and θ is the actual answer of this person. Thus, l measures the distance between your estimate and the actual answer of the other person.

To explain θ: if you are asked to estimate a *choice* of the other person (which you are in the second question), then θ is either 0 or 100, depending on the actual choice of this person. Your decision screen will clarify which of the possible choices of the other person corresponds to 0 and which to 100. If you are asked to guess an *estimate* made by the other person (in the third question), θ is the actual estimate provided by the other person.

Your payment is calculated as follows. The computer draws at random an additional integer number z between 0 and 100, with equal probability for each integer. If the error measure l is strictly less than $z/100$ ($l < z/100$), you receive 12 euro. If l is greater or equal than $z/100$ ($l \geq z/100$), you receive 4 euro.

It follows from this rule that it is optimal for you to enter a relatively high number x (close to 100) if you think that the answer θ of the other person is large. Conversely, it it optimal for you to enter a relatively small number if you think that the answer of the other person matched with you is small. In this way, you maximize the probability of receiving 12 euro for every possible realization of the integer z. Please note: how large or small your optimal x is, depends on your precise assessment of the how likely the other person chooses each of their possible answers.

With this in mind, the Binarized Scoring Rule implies that it is always optimal for the participants to truthfully state their own estimate. This was proven by Hossain and Okui (2013).[21]

Understanding checks

1. What is the outcome for Person *S* if *Table B* is randomly selected by the computer, Person *S* sends message *"Table A has been selected"*, and Person *R* chooses *Option B*?

[21] Tanjim Hossain, Ryo Okui; The Binarized Scoring Rule, The Review of Economic Studies, Volume 80, Issue 3, 1 July 2013, Pages 984–1001.

- Options: 4 , 8 , 12

2. What is the outcome for Person *S* if *Table B* is randomly selected by the computer, Person *S* sends message *"Table B has been selected"*, and Person *R* chooses *Option B*?
   - Options: 4 , 8 , 12
3. From the perspective of which role will you need to make decisions?
   - Options: Person S, Person R, both
4. Is the probability of you earning a higher payoff from the questions about expectations decreasing, increasing or independent of you giving an answer that is closer to the true value?
   - Options: decreasing, increasing, independent

Experiment

[Text in box remains on left hand side of screen throughout experiment]

Table A

	Option A	Option B
Person S	4	12
Person R	12	4

Table B

	Option A	Option B
Person S	12	4
Person R	4	12

1. Payoffs from the experiment depend on either *Table A* or *Table B*
2. The computer will randomly select one of these tables (each with equal probability)
3. Only Player S is told which table is selected
4. Player S sends one of the following messages to Player R: *"Tables A has been selected"* or *"Table B has been selected"*
5. Player R is asked to choose either *Option A* or *Option B*
6. If the option chosen coincides with the table selected by computer, Player S gets 4 euro and Player R receives 12 euro. If they do not coincide, the payoffs are reversed.

Person *S*

You have now been matched with another participant of the experiment.

Question 1S

Suppose you are Person *S* and the computer has randomly selected *Table A*.

Which message do you send to Person *R*?

- "Table A has been selected"
- "Table B has been selected"

NOTE: Above you only chose a message to send when Table A is selected. To simplify the experiment, we use symmtry: to determine payoffs, we assume that if you send "Table A has been selected" in the case above, then by symmetry you would send "Table B has been selected" if Table B was actually selected. If instead you send "Table B has been selected" in the case above, then we assume that you would send "Table A has been selected" if Table B was actually selected.

You will now be asked to estimate the answers of the Person *R* with whom you are matched. More precisely, you will be asked to indicate how likely you consider a given event or answer by Person *R* to be. Remember, stating a likelihood is equivalent to stating how many out of 100 Person *R*s you think would choose a certain answer.

Question 2S

Suppose you are Person *S* and that you sent the message "Table A has been selected".

How likely do you think it is that Person *R* chooses Option A?

☐ out of 100

Question 3S

Suppose you are Person *S*.

Person *R* is asked the following question: "Suppose you are Person *R* and that Person *S* sent you the message "Table A has been selected". How likely do you think it is that Table A was randomly selected by the computer?".

You, as Person *S*, are asked to estimate the answer given by Person *R*. Please read the question in the previous paragraph once again and estimate: what answer do you think does Person *R* gives to this question?

☐ out of 100

Person *R*

Remember, you are still matched with the same other participant of the experiment.

Question 1R

Now, suppose instead you are Person *R* and that Person *S* sent you the message "Table A has been selected".

Which option do you choose?

- Option A
- Option B

NOTE: Above you only chose an option for when you receive the message "Table A has been selected". Just as before, we simplify and use symmetry: to determine your payoffs, we assume that if you choose Option A in the case above, then you would choose Option B if you receive the message "Table B has been selected". If instead you choose Option B in the case above, then we assume you would choose Option A if you receive the message"Table B has been selected".

You will now be asked to estimate the answers by Person *S*. More precisely, you will be asked to indicate how likely (out of 100) you consider a given event or answer by Person *S* to be.

Question 2R

Suppose you are Person *R* and that Person *S* sent you the message "Table A has been selected".

How likely do you think it is that Table A was randomly selected by the computer?

☐ out of 100

Question 3R

Suppose you are Person *R*.

Person *S* is asked the following question: "Suppose you are Person *S* and that you sent the message "Table A has been selected". How likely do you think it is that Person *R* chooses Option A?"

You, as Person *R*, are asked to estimate the answer given by Person *S*. Please read the question in the previous paragraph once again and estimate: what answer do you think does Person *S* give to this question?

☐ out of 100

Acknowledgements The authors thank Dorothea Kübler, Ronald Peeters and Joel Sobel for helpful conversations, the German Science Foundation for financial support via CRC TRR 190 (project number 280092119) and the Einstein Foundation Berlin for financial support via the Einstein Visiting Fellowship of Bertil Tungodden. The data analysis was pre-registered as AsPredicted #109270, available at https://aspredicted.org/SML_RN3

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Agranov M, Dasgupta U, Schotter A (2024) Trust me: communication and competition in a psychological game. J Eur Econ Assoc. https://doi.org/10.1093/jeea/jvae012

Bond CFJ, DePaulo BM (2006) Accuracy of deception judgments. Pers Soc Psychol Rev 10:214–243

Bursztyn L, Gonzalez AL, Yanagizawa-Drott D (2020) Misperceived social norms: women working outside the home in Saudi Arabia. Am Econ Rev 110:2997–3029

Charness G, Dufwenberg M (2006) Promises and partnership. Econometrica 74:1570–1601

Cohen S, Li S (2023) Sequential cursed equilibrium. Unpublished manuscript

Costa-Gomes MA, Weizsäcker G (2008) Stated belief and play in normal-form games. Rev Econ Stud 75:729–762

Crawford VP (2003) Lying for strategic advantage: rational and boundedly rational misrepresentation of intentions. Am Econ Rev 93:133–149

Crawford VP, Costa-Gomes MA, Iriberri N (2013) Structural models of nonequilibrium strategic thinking: theory, evidence, and applications. J Econ Lit 51:5–62

Eyster E, Rabin M (2005) Cursed equilibrium. Econometrica 73:1623–1672

Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. Exp Econ 10:171–178

Fong M-J, Lin P-H, Palfrey TR (2023) Cursed sequential equilibrium. Unpublished manuscript

Gneezy U, Kajackaite A, Sobel J (2018) Lying aversion and the size of the lie. Am Econ Rev 108:419–453

Greiner B (2015) Subject pool recruitment procedures: organizing experiments with ORSEE. J Econ Sci Assoc 1:114–125

Hossain T, Okui R (2013) The binarized scoring rule. Rev Econ Stud 80:984–1001

Kartik N (2009) Strategic communication with lying costs. Rev Econ Stud 76:1359–1395

Levine TR (2014) Truth-default theory (TDT): a theory of human deception and deception detection. J Lang Soc Psychol 33:379–392

Manski CF, Neri C (2013) First- and second-order subjective expectations in strategic decision-making: experimental evidence. Games Econ Behav 81:232–254

Peeters R, Vorsatz M, Walzl M (2015) Beliefs and truth-telling: a laboratory experiment. J Econ Behav Organ 113:1–12

Polonio L, Coricelli G (2019) Testing the level of consistency between choices and beliefs in games using eye-tracking. Games Econ Behav 113:566–586

Rey Biel P (2009) Equilibrium play and best response to (stated) beliefs in normal form games. Games Econ Behav 65:572–585

Schotter A, Trevino I (2014) Belief elicitation in the laboratory. Annu Rev Econ 6:103–128

Sheremeta RM, Shields TW (2013) Do liars believe? Beliefs and other-regarding preferences in sender–receiver games. J Econ Behav Organ 94:268–277

Sperber D, Wilson D (1995) Relevance: communication and cognition, 2nd edn. Blackwell Publishing, New Jersey

Vrij A (2008) Detecting lies and deceit: pitfalls and opportunities. John Wiley & Sons, New Jersey

Weizsäcker G (2023) Misunderstandings: false beliefs in communication. Open Book Publishers, UK