Elmshauser, Béla; Friedman, Evan; Jo, Yoon Joo

Working Paper

Deception Aversion

CESifo Working Paper, No. 12154

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Elmshauser, Béla; Friedman, Evan; Jo, Yoon Joo (2025) : Deception Aversion, CESifo Working Paper, No. 12154, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/331620

CES ifo Working Papers

12154
2025

September 2025

Deception Aversion

Béla Elmshauser, Evan Friedman, Yoon Joo Jo

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Deception Aversion*

Béla Elmshauser† Evan Friedman‡ Yoon Joo Jo§

First version: 2nd June 2025
Current version: 22nd September 2025

Abstract

Conveying private information to interested parties is central to almost every economic and social activity. In such interactions, the sender may *lie* by misreporting the truth, but may also *deceive* by inducing inaccurate beliefs about the payoff-relevant state. While a huge experimental literature documents aversion to lying, there is little evidence regarding aversion to deceiving others. Deception aversion is conceptually difficult to document because it depends on unobserved second-order beliefs: the sender's belief over the receiver's belief (over the payoff-relevant state). In this paper, we introduce a novel game and show theoretically how to identify deception aversion from choice data alone, with minimal assumptions on second-order beliefs. We run a laboratory experiment and find strong support for deception aversion that is robust to several natural variations of the game. Many subjects lie in order to avoid deception, and structural estimates imply that 30% of subjects are deception-averse.

Keywords: lying; deception; lying aversion; deception aversion; image concerns; strategic communication; psychological game theory

JEL Classification: C44, C72, C92

*We thank Kai Barron, Juan Francisco Blazquiz-Pulido, Valeria Burdea, Syngjoo Choi, Tilman Fries, Alessandro Ispano, Nicolas Jacquemet, Philippe Jehiel, Agne Kajackaite, Kiryl Khalmetski, Ernest Lai, Wooyoung Lim, Georgia Michailidou, Suanna Oh, Luca Polonio, Collin Raymond, Claire Rimbaud, Michael Thaler, and seminar participants at Paris School of Economics, Indiana University, Purdue University, NYU-Abu Dhabi, National University of Singapore, University of British Columbia, ESWC-2025, SAET-2025, and the Workshop on Narratives, Memory, and Beliefs-2025 for helpful comments. This work was supported by the Agence Nationale de la Recherche (grant ANR-24-CE28-0028-01).

†Email: bela.elmshauser@psemail.eu
‡Email: evan.friedman@psemail.eu
§Email: yoonjo@tamu.edu

Conveying private information to interested parties is central to many economic and social activities. For example, an expert gives advice (e.g., Crawford and Sobel [1982]), a prosecutor persuades a judge (e.g., Kamenica and Gentzkow [2011]), or a seller communicates the value of an object to a potential buyer (e.g., Kim [2012]).

In standard economic theory, the messages available to the sender have no intrinsic meaning, and she will send whatever message maximizes her expected payoff. In contrast, a large experimental literature documents a *preference for truth-telling* whereby subjects tend to tell the truth at the expense of their own material payoffs. One leading explanation is an *aversion to lying*—that people are simply averse to sending messages that are not literally true.[1] As our examples illustrate, however, what matters for the receiver is not the literal meaning of messages, but the information conveyed by those messages. Hence, there is a distinction between lying and *deception*—defined as the sending of messages that induce inaccurate beliefs (Sobel [2020]). A natural question then is: are people also *averse to deceiving* others?

A sender deceives if she sends a message that *she believes* induces more inaccurate beliefs than another message she could have sent. Hence, deception depends on the sender's *second-order beliefs*, the mapping from messages to her belief over the receiver's belief (over the payoff-relevant state). The fundamental issue with studying deception aversion is that second-order beliefs are unobservable.

In this paper, we introduce a novel game and show theoretically how to identify deception aversion from choice data alone, with minimal assumptions on the sender's second-order beliefs. In particular, we do not assume that beliefs are determined in equilibrium, nor do we elicit them directly. We run a laboratory experiment and find strong support for deception aversion. Many subjects *lie in order to avoid deception*, and structural estimates imply that 30% of subjects are deception-averse.

The game we introduce is a variation of the standard paradigm from the lying aversion literature in which people privately observe the outcome of a *die roll* and are paid in proportion to the outcome they report (Fischbacher and Follmi-Heusi [2013]).

Our version differs in two key ways. First, it is a sender-receiver (cheap-talk) game in which an outcome of the die is communicated by the sender to the receiver. Second, in addition to the die, the sender also privately observes the outcome of a *coin flip*, which

[1] In "Related literature" below, we discuss other motives for truth-telling that have been considered in the literature.

determines the conditional distribution of die outcomes, with higher die outcomes being more likely following heads and less likely following tails. Hence, the message (which reports a die outcome only) may give information about the coin. After observing the message, the receiver reports a belief over the probability that the coin is heads and is incentivized for more accurate beliefs.[2] Hence, the coin is payoff-relevant for the receiver; and for the sender to induce inaccurate beliefs about the coin constitutes deception.

We consider various specifications for the sender's payoffs,[3] but in all cases, the sender communicates about the die, which influences beliefs about the coin. The games capture many common economic interactions, e.g., the seller of an object who manipulates the buyer's beliefs over quality (the coin) by reporting signals of quality (the die).

We first ask: what are the behavioral signatures of deception aversion? To answer this, we consider a theoretical framework that allows for very general lying and deception costs. We assume only that (1) lying costs depend on both the die outcome and the sender's message and that (2) deception costs depend on the coin outcome and the second-order beliefs induced by the sender's message. Importantly, we model second-order beliefs—the entire mapping from messages to the sender's belief over the receiver's belief about the coin—as an unobserved object upon which we impose no a priori restrictions.[4] Within this framework, a sender who exhibits *coin effects*—an effect of the coin, conditional on the die—must be averse to deception.[5]

The intuition for why coin effects indicate deception aversion is straightforward. Conditioning on die outcome, any given message will induce the same belief in the receiver and result in the same material payoff and lying cost for the sender, independent of the coin. Hence, without deception costs, the sender's message cannot be affected by the outcome of the coin. On the other hand, deception costs can lead to coin effects because the same message will be more or less deceptive, depending on the coin's realization.

While our general results do not impose any restriction on beliefs, it is natural for

[2] In the lab, we use the binarized scoring rule (Hossain and Okui [2013]).

[3] In the *message-payoff* game, the sender is paid in proportion to the message itself (as in the standard paradigm), independent of the receiver's reported belief. In the *belief-payoff* game, the sender is paid in proportion to the receiver's reported belief so that it is in her material interest to convince the receiver the coin is heads. In both cases, the coin is payoff-relevant for the receiver, but *not* the sender.

[4] Indeed, our only assumption on second-order beliefs is that they are measurable with respect to the receiver's information set, which only contains the message observed.

[5] Similarly, a sender who exhibits *die effects*—an effect of the die, conditional on the coin—must be averse to lying.

subjects to anchor beliefs on the (perhaps) naive assumption that senders have a tendency to tell the truth.[6] Due to the positive correlation between the coin and die, such naïveté pushes first- and second-order beliefs to be monotonically increasing in the message. With increasing second-order beliefs, deception-averse senders are predicted to exhibit *positive coin effects* whereby they send higher messages following heads as opposed to tails. Moreover, the size of the coin effect is predicted to increase in the degree of deception aversion. Hence, an individual subject's *average coin effect* is a natural reduced-form statistic that allows us to summarize the heterogeneity of deception aversion in the population.

We find that many subjects exhibit positive coin effects (and very few have negative coin effects), consistent with deception aversion and increasing second-order beliefs. In particular, many subjects *lie in order to avoid deception*. Among these, perhaps the most interesting are those who exhibit both coin and die effects, implying a mix of both lying and deception aversion. This suggests that these subjects do find lying costly, and yet they still choose to lie in order to avoid deception.

We formulate a simple structural model in which lying and deception costs are both linear. We jointly estimate the cost parameters and second-order beliefs by fitting the model to the data of individual sender-subjects.[7] We find that the model provides a very good fit for most subjects, capturing much of the heterogeneity in the data. Structural estimates imply that 36% of subjects are averse to lying only, 15% are averse to both lying and deception, and 15% are averse to deception only. Hence, while lying aversion is more prevalent than deception aversion, we still find that a large fraction (30%) of subjects are averse to deception.

An important question is when to expect deception aversion. A natural conjecture is that there will be less deception aversion in situations that feel more "adversarial." To test this, we ran two versions of the deception game, based on different specifications of the sender's payoffs. In the *message-payoff* game, the sender is paid in proportion to the message itself, independent of the receiver's reported belief. In the *belief-payoff* game, the sender is paid in proportion to the receiver's reported belief, and hence there is a direct incentive for the sender to deceive the receiver. While sender strategies

[6] We formalize this simple idea with a level k-type model in the spirit of spirit of Wang et al. [2010].

[7] Using the receiver-subjects' elicited first-order beliefs as a proxy for the sender's second-order beliefs, we posit a one-parameter family of second-order belief functions for the estimation.

differ across the two treatments, we find that the structural model implies a similar prevalence of deception-averse subjects. Hence, a direct incentive to deceive does not drive out deception aversion, suggesting robustness across a variety of strategic settings.

Finally, because deception does hurt the receiver, an important question is whether the effects we observe are due to an aversion to deception per se, or if they are fully driven by *social* or *other-regarding* preferences. To test this, we also run a treatment based on the *pure lying* game in which the sender can lie in order to directly implement her desired sender-receiver allocation, but there is no opportunity to deceive. Using this game as a control, we argue that the deception aversion documented in other treatments cannot be fully driven by preferences over final outcomes.

In the remainder of this section, we review related literature. Section 1 introduces the theoretical framework. Section 2 gives the experimental design. Section 3 analyzes the beliefs data and introduces a family of models consistent with beliefs. Section 4 gives reduced-form results based on coin and die effects. Section 5 presents the structural model and estimates. Section 6 establishes that deception aversion cannot be fully driven by social preferences, and Section 7 concludes.

Related literature. Our most novel findings concern deception aversion, on which there is little existing work. However, our game features opportunities to lie as well as to deceive, so we speak to the large and rapidly growing literature on lying aversion. Recent work by Gneezy et al. [2018] and Abeler et al. [2019] conclude that the typical findings from "die rolling" experiments can be explained by a model with (1) a fixed cost of lying and (2) image concerns—i.e., a "reputation for honesty" cost such that it is costly to send messages that are typically sent by liars.[8] Our theoretical framework is general enough to incorporate both types of lying costs, as well as others found in the literature.[9] This allows us to conclude that our finding of "coin effects," while explicable by deception aversion, cannot be explained by lying aversion.

Our definitions of lying and deception follow Sobel [2020], who only recently formalized these as distinct notions. Previous work has largely focused on studying the

[8] Khalmetski and Sliwka [2019] propose a cost of sending a message that is proportional to the equilibrium fraction of liars that send that message. Dufwenberg and Dufwenberg [2018] instead posit a cost that is proportional to the expected size of the lie, also determined in equilibrium.

[9] Our general notion of lying costs also allows for the third family of costs considered in the meta-study of Abeler et al. [2019]—costs based on social norms or comparisons. For instance, it may be that people feel less bad about lying if they believe others are lying too, as proposed by Gibson et al. [2013].

relationship between lying aversion and payoff consequences. Gneezy [2005] considers experimental sender-receiver games in which the sender has the opportunity to lie about which of two actions yields the receiver a higher payoff.[10] He finds that the fraction of subjects willing to lie depends on payoff consequences, both their personal benefit and the harm caused to others.[11] The earlier literature variously uses the terms "lying" and "deception," but does not make a distinction between them. While Gneezy [2005] refers to "deception," it is much closer to what Sobel [2020] (and we) refer to as lying. In a setup based on Gneezy [2005], Sutter [2009] pre-empts the modern distinction between lying and deception: by eliciting the sender's beliefs over the receiver's action following messages, he concludes that some senders tell the truth in order to deceive.[12]

Only recently, in a purely theoretical paper, Eilat and Neeman [2023] were the first to formalize a notion of deception costs. Taking an equilibrium approach, they assume that only deviations from equilibrium messages incur a deception cost—a cost that increases in the distance between the receiver's equilibrium belief and that induced by the deviation.[13] In contrast, in our non-equilibrium approach, deception is an entirely subjective phenomenon based on the sender's second-order beliefs. We make the natural assumption that deception costs are increasing in the distance between the induced second-order belief and the state.

The paper most closely related to ours is Choi et al. [2025], which also experimentally disentangles lying and deception aversion.[14] They do so in the context of a two-stage reputation building game in which the senders, whose "preference type" is either aligned or misaligned with the receiver, sends one of two messages about the payoff-relevant

[10] Gneezy et al. [2013] study games in which "the decision to lie increases own payment independent of the counterpart's decision, but potentially at a cost for the counterpart."

[11] One interpretation of Gneezy's findings is that an individual's decision of whether or not to lie depends directly on payoff consequences. Qualifying this conclusion, Hurkens and Kartik [2009] show that his results are consistent with the hypothesis that people are either completely unwilling to lie or will lie whenever it leads to a more favorable allocation.

[12] Blazquiz-Pulido et al. [2024] combine a strategic sender-receiver game with a non-strategic version and more detailed belief elicitation to improve "the identification of the senders' intentions to deceive." Alempaki et al. [2025] show that some subjects deceive by being evasive rather than lying directly.

[13] Earlier work by Kartik [2009] characterizes equilibria in the presence of lying costs.

[14] Another related paper is Farina et al. [2024], who experimentally study verifiable disclosure games in which it is possible to deceive, but not to lie. They argue that some subjects display patterns indicative of deception aversion. An important feature of our cheap-talk games, as well as the games of Choi et al. [2025], is that they permit subjects to be classified as lying-averse, deception-averse, both, or neither.

state.[15] By making one type of sender a computer with a pre-programmed strategy, one variation of their game admits (non-deceptive) lying in equilibrium, and another admits (truthful) deception. First, they document deviations from equilibrium that are consistent with lying and deception aversion. However, to classify individual subjects (e.g. as deception-averse rather than making an "inference error") requires identifying second-order beliefs, which they pursue through direct elicitation in a second experiment. We, on the other hand, introduce a novel game and develop an approach that does not require belief elicitation or an assumption of equilibrium.

By modeling a direct dependence of utility on second-order beliefs, we contribute to the field of psychological game theory (Geanakoplos et al. [1989]). In particular, deception aversion is related to the concept of guilt aversion (Battigali and Dufwenberg [2007] and Charness and Dufwenberg [2006]), which is an aversion to upsetting others' expectations of payoffs, as opposed to inducing inaccurate beliefs. The experiments in this literature are typically based on trust, dictator, and ultimatum games and involve belief elicitation or directly giving information about the receiver's expectations (e.g. Ellingsen et al. [2010]). More closely related to our work, Battigalli et al. [2013] show that the results of Gneezy [2005], previously explained in terms of lying aversion, are also consistent with guilt aversion under some mild assumptions on the sender's second-order beliefs. On the other hand, the games and identification strategy we introduce allow us to disentangle lying and deception aversion.

1 Theory

1.1 A game with lying and deception opportunities

A sender has 2-dimensional private information, which can be thought of as the outcome of a *coin* Θ and a *die* D. We use $\theta \in \{0, 1\}$ for realizations of Θ and $d \in \{1, 2, ..., T\}$ for realizations of D. We let $p_{\Theta,D}(\theta, d)$, $p_{\Theta}(\theta)$, and $p_D(d)$ denote the joint and marginal probability mass functions (PMFs), respectively, which we assume have full support. Letting $p_{D|\theta}(d)$ be the PMF of D conditional on $\Theta = \theta$, we assume that the the likelihood

[15] Their setup recalls Ettinger and Jehiel [2010] and Ettinger and Jehiel [2021], who theoretically predict and experimentally validate the use of a "deceptive tactic" in similar games whereby players initially tell the truth and then switch to lying over the course of play.

ratio $\frac{p_{D|1}(d)}{p_{D|0}(d)}$ is strictly increasing in d. Hence, the coin and die are *positively correlated*.

After observing the coin and die, the sender sends a message $m \in \{1, 2, ..., T\}$ to the receiver, which we think of as declaring an outcome of the die (though this need not be truthful). Hence, her strategy $\sigma_S : \{0, 1\} \times \{1, 2, ..., T\} \to \Delta\{1, 2, ..., T\}$ gives her probability of sending each message as a function of her private information. We use $\sigma_S(m|\theta, d)$ for the probability message m is sent after realizing (θ, d). After observing the message, the receiver reports a belief $b \in [0, 1]$ over the probability $\Theta = 1$, so that her pure strategy is $\sigma_R : \{1, 2, ..., T\} \to [0, 1]$. The restriction of the receiver to pure strategies is immaterial for the reason described in the next paragraph.

The receiver's payoffs are such that, if she believes $\Theta = 1$ with probability b', it is strictly optimal to report $b = b'$. We abstract from the details here, but in our laboratory implementation, we use the binarized scoring rule (Hossain and Okui [2013]).

We consider two different versions of the game, defined by differences in the sender's payoffs. In the *message-payoff* game, the sender's payoff $u^M(m)$ depends *only* on the message m she sends and is strictly increasing in m. In the *belief-payoff* game, the sender's payoff $u^B(b)$ is strictly increasing in the receiver's reported belief b. In both cases, the coin is payoff-relevant for the receiver, but *not* the sender.

Economic applications. These games are two natural variations that capture many common economic interactions. In the belief-payoff game, the sender has a material incentive to convince the receiver that $\theta = 1$. Hence, there is a direct incentive to deceive. This captures the seller of an object who manipulates the buyer's beliefs over quality (the coin) by reporting signals of quality (the die). The message-payoff game is a simplified variation in which the sender's payoff increases in the message sent, and therefore does not depend on the receiver's action (as in the standard "die rolling" paradigm). Hence, unlike in the belief-payoff game, deception comes only as an externality.

Experimentally, the comparison of the two games is interesting. A natural conjecture is that the direct incentive to deceive in the belief-payoff game will decrease deception aversion, but it may also increase it if it makes deception opportunities more salient. Hence, the comparison provides insight into the role of strategic forces in modulating deception aversion. Importantly, however, our general strategy for identifying lying and deception aversion, outlined in Section 1.3, is the same for both games (and, in fact, many other variations; see Remark 3).

1.2 The equilibrium benchmark

We first establish the Perfect Bayesian equilibrium (PBE) benchmark. In both message-payoff and belief-payoff games, all PBE are pooling in the sense that both $(\theta = 0)$-senders and $(\theta = 1)$-senders use the same (or equivalent) strategies, which is a consequence of the fact they have the same incentives and message space. As a result, no information is transmitted and the receiver does not update her beliefs after any (on-path) message.

Proposition 1. *(i) In all PBE of the message-payoff game, the sender always sends the highest message and the receiver does not update beliefs (on path): $\sigma_S(T|\theta, d) = 1$ for all (θ, d) and $\sigma_R(T) = p_\Theta(1)$. (ii) In all PBE of the belief-payoff game, $(\theta = 0)$-senders and $(\theta = 1)$-senders use strategies that induce the same distribution over messages and the receiver does not update beliefs (on path): $\sum_d \sigma_S(m'|0, d)p_{D|0}(d) = \sum_d \sigma_S(m'|1, d)p_{D|1}(d)$ for all m', and $\sigma_R(m') = p_\Theta(1)$ for any m' in the support of $\sigma_S(\cdot)$.*[16]

Proof. See Appendix 8.1. □

With lying and/or deception costs, there may be equilibria in which information is transmitted. One approach taken in the literature is to characterize equilibria in the presence of lying costs (e.g. Gneezy et al. [2018]) or, more recently, deception costs (Eilat and Neeman [2023]). In this paper, however, we favor a non-equilibrium framework that allows us to identify deception aversion on the individual subject level under minimal assumptions on the sender's second-order beliefs.

1.3 A non-equilibrium framework with lying and deception costs

We let $\beta : \{1, ..., T\} \to [0, 1]$ be the sender's *second-order beliefs*, so that $\beta(m) \in [0, 1]$ is the sender's belief over the receiver's belief of the probability that $\Theta = 1$ after observing message m. We take this as exogenous and we initially impose *no restriction* whatsoever on this function.[17] There is of course the implicit assumption that second-order beliefs are measurable with respect to the receiver's information set. In particular, we exclude

[16] In both the message-payoff and belief-payoff games, these are PBE if and only if, following any off-path message, the receiver forms some belief $b' \leq p_\Theta(1)$ and best responds by reporting $b = b'$.

[17] We do not even impose that second-order beliefs are consistent with any sender strategy. For example, we allow for $\beta(m') = 1$ for all m', which is not consistent with any strategy as it implies that the $(\theta = 0)$-sender sends every message m' with zero probability.

any "magical thinking" in which the sender believes that the receiver has access to an independent signal about the realization of the coin Θ. Importantly, we as experimenters observe coin-die realizations (θ, d), but *not* second-order beliefs β.

Following Sobel [2020], the sender *lies* if message m does not coincide with die outcome d. The sender *deceives* by sending message m if (she believes that) the message does not lead to the most accurate belief about the payoff-relevant state Θ. Unlike lying, the definition of deception depends on the unobserved second-order beliefs.

Definition 1. Suppose the sender observes (θ, d). The sender *lies* at (θ, d) if $m \neq d$. The sender *deceives* at (θ, d) if $m \notin \underset{m' \in \{1,...,T\}}{argmin} |\beta(m') - \theta|$.

Lying and deception are conceptually distinct. In particular, telling the truth may be deceptive, and lying need not be.

Remark 1. Telling the truth at (θ, d) is deceptive if $d \notin \underset{m' \in \{1,...,T\}}{argmin} |\beta(m') - \theta|$. Lying at (θ, d) is not deceptive if $m \in \underset{m' \in \{1,...,T\}}{argmin} |\beta(m') - \theta|$.

We now posit a very general utility function for the sender, incorporating psychological costs to lying and deception.

Lying costs depend on die realization d and message m through the function $c^{lie}(m, d)$. This is extremely general, allowing for a fixed cost of lying, costs that increase in the distance between d and m, and costs that cannot be captured by conventional notions of distance. This also allows for image concerns—i.e. a "reputation for honesty" cost from being *perceived as a liar* (e.g. Gneezy et al. [2018] and Abeler et al. [2019]).[18] Referring to all of these costs collectively as "lying costs," our goal is not to distinguish between the different components of lying costs, but to separate all such costs from deception costs.

Deception costs depend on the coin realization θ and the sender's second-order belief $\beta(m)$ induced by message m through the function $c^{dec}(\beta(m), \theta)$. This is also very general,

[18] Let $\Lambda : \{1, ..., T\} \to [0, 1]$ be the sender's exogenously specified *second-order perception* of being a liar such that $\Lambda(m)$ is the sender's belief of the receiver's belief that $m \neq d$ after observing message m. Given Λ, we may suppose that $c^{lie}(m, d) = g(m, \Lambda(m), d)$ for some function g. Similarly, we can also incorporate costs based on social norms or comparisons (e.g. Gibson et al. [2013])—i.e. feeling less bad about lying if others are also lying—by defining a subjective belief over others' lying behavior.

though for some results, we will specialize to the natural case that costs increase in the distance $|\beta(m) - \theta|$.

Letting $u^g(m;\beta)$ be the expected material payoff in game $g \in \{M, B\}$, i.e. $u^g(m;\beta) = u^M(m)$ in the message-payoff game or $u^g(m;\beta) = u^B(\beta(m))$ in the belief-payoff game,[19] the sender's overall *subjective* utility is the expected material payoff minus the two types of psychological costs:

$$\phi^g(m, \theta, d; \beta) = u^g(m;\beta) - c^{lie}(m, d) - c^{dec}(\beta(m), \theta). \tag{1}$$

We assume that senders maximize utility, and so we let $\sigma_S^*(\theta, d) := \underset{m' \in \{1,...,T\}}{argmax}\ \phi^g(m', \theta, d; \beta) \subset \{1, ..., T\}$ be the set of optimal (pure) messages following coin-die realization (θ, d). Note that this depends on second-order beliefs β and game g, which we suppress in the notation. For simplicity, we assume a unique best response so that σ_S^* is *single-valued*, an assumption which holds generically.[20]

Our first result shows how to identify the existence of deception aversion from choice data alone, without any assumption on second-order beliefs.

Proposition 2. *If $\sigma_S^*(0, d) \neq \sigma_S^*(1, d)$ for some d, then the sender is strictly averse to deception: either $c^{dec}(\beta(\sigma_S^*(0, d)), 1) > c^{dec}(\beta(\sigma_S^*(0, d)), 0)$, $c^{dec}(\beta(\sigma_S^*(1, d)), 0) > c^{dec}(\beta(\sigma_S^*(1, d)), 1)$, or both.*

Proof. Suppose for purposes of contradiction that $c^{dec}(\beta(\sigma_S^*(0, d)), 1) \leq c^{dec}(\beta(\sigma_S^*(0, d)), 0)$ and $c^{dec}(\beta(\sigma_S^*(1, d)), 0) \leq c^{dec}(\beta(\sigma_S^*(1, d)), 1)$. By definition of $\phi^g(\cdot)$, we have that $c^{dec}(\beta(\sigma_S^*(0, d)), 1) \leq c^{dec}(\beta(\sigma_S^*(0, d)), 0) \implies \phi^g(\sigma_S^*(0, d), 1, d) \geq \phi^g(\sigma_S^*(0, d), 0, d)$. By definition of $\sigma_S^*(\cdot)$, we have that $\phi^g(\sigma_S^*(1, d), 1, d) > \phi^g(\sigma_S^*(0, d), 1, d)$ and thus that $\phi^g(\sigma_S^*(1, d), 1, d) > \phi^g(\sigma_S^*(0, d), 0, d)$. A symmetric argument yields $\phi^g(\sigma_S^*(0, d), 0, d) > \phi^g(\sigma_S^*(1, d), 1, d)$, a contradiction. □

Hence, any *coin effect*—an effect of the coin on the message, conditional on die outcome—indicates deception aversion.[21] The intuition is as follows. Conditioning on

[19] This implicitly assumes that the sender believes the receiver best responds to her belief by reporting it accurately.

[20] For any (θ, d), for almost all (c^{lie}, c^{dec}), (1) has a unique maximizer for almost all β.

[21] Technically, at this level of generality, coin effects may indicate that the sender is deception *loving*, so it is more accurate to say "non-neutral deception preferences." For later results, we will refine the deception cost function to be able to identify deception aversion specifically.

die outcome, any given message will induce the same belief in the receiver and result in the same material payoff and lying cost for the sender, independent of the coin. Hence, without deception costs, the sender's message cannot be affected by the outcome of the coin. On the other hand, deception costs can lead to coin effects because the same message will be more or less deceptive, depending on the coin's realization.

Remark 2. Proposition 2 shows that coin effects are sufficient for identifying deception aversion. They are not, however, necessary. For instance, a sender with perfectly "flat" second-order beliefs ($\beta(m) = \beta(m')$ for all m, m') is incapable of deception, and so will not exhibit coin effects even if she is extremely deception-averse. Hence, our experiment identifies a *lower bound* on the fraction of subjects who are deception-averse.

In a similar fashion, we may identify lying aversion with minimal assumptions. Any *die effect*—an effect of the die on the message, conditional on the coin outcome—indicates lying aversion.

Proposition 3. *If $\sigma_S^*(\theta, d) \neq \sigma_S^*(\theta, d')$ for some θ and $d \neq d'$, then the sender is strictly averse to lying: either $c^{lie}(\sigma_S^*(\theta, d), d') > c^{lie}(\sigma_S^*(\theta, d), d)$, $c^{lie}(\sigma_S^*(\theta, d'), d) > c^{lie}(\sigma_S^*(\theta, d'), d')$, or both.*

Proof. See Appendix 8.1. The proof closely parallels the proof of Proposition 2. □

There is thus a separability between deception and lying aversion that allows us to conclude that coin effects indicate deception aversion and die effects indicate lying aversion. A useful corollary states that purely deception-averse senders will ignore the die, and purely lying-averse senders will ignore the coin.

Corollary 1. *(i) A purely deception-averse sender (i.e. with no lying costs, $c^{lie}(m, d) = c^{lie}(m', d')$ for all m, m', d, d') will ignore the die: $\sigma_S^*(\theta, d) = \sigma_S^*(\theta, d')$ for all θ, d, and d'. (ii) A purely lying-averse sender (i.e. with no deception costs, $c^{dec}(\beta(m), \theta) = c^{dec}(\beta(m'), \theta')$ for all m, m', θ, θ') will ignore the coin: $\sigma_S^*(0, d) = \sigma_S^*(1, d)$ for all d.*

Senders who are averse to both deception and lying may exhibit both coin and die effects. Indeed, we find many such subjects in our experiment.

With additional structure on second-order beliefs and deception costs, we may make more precise predictions. For instance, consider the following monotonicity assumption.

Assumption 1. *(1) Second-order beliefs $\beta(m)$ are strictly increasing in m, and (2) deception costs $c^{dec}(\beta(m), \theta)$ are an increasing function of the (absolute) distance between $\beta(m)$ and θ, i.e. $c^{dec}(\beta(m), \theta) = f(|\beta(m) - \theta|)$ for (weakly) increasing f.*

Part (1) of the assumption states that second-order beliefs are strictly increasing. In Section 3, we show theoretically that this will be the case for naive senders who believe that receivers are sufficiently credulous. While we do not elicit second-order beliefs in our experiment, we find that receivers' *first-order* beliefs tend to be increasing, consistent with increasing second-order beliefs. Part (2) is an extremely weak assumption, stating that the cost of deception increases in the size of the deception, as measured by the distance between θ and the induced belief $\beta(m)$. Because we only assume that $f(\cdot)$ is weakly increasing, this allows for *no deception costs* whatsoever, a fixed cost of deceiving *at all*, or a fixed cost of deceiving above some threshold.

Under Assumption 1, the sender will tend to send higher messages when $\theta = 1$. That is, we should expect *positive coin effects*: $\sigma_S^*(1, d) - \sigma_S^*(0, d) \geq 0$. Moreover, if two senders have the same lying cost function and second-order beliefs, if one finds deception more costly, she will exhibit an even larger coin effect.

Proposition 4. *Suppose Assumption 1 holds. (i) $\sigma_S^*(1, d) - \sigma_S^*(0, d) \geq 0$ for all d. (ii) Let 1 and 2 be two senders with the same second-order beliefs β and lying cost function c^{lie}. If* deception is more costly for 1 than 2,[22] *then $\sigma_S^{*1}(1, d) \geq \sigma_S^{*2}(1, d)$, $\sigma_S^{*1}(0, d) \leq \sigma_S^{*2}(0, d)$, and $\sigma_S^{*1}(1, d) - \sigma_S^{*1}(0, d) \geq \sigma_S^{*2}(1, d) - \sigma_S^{*2}(0, d)$ for all d.*

Proof. (i): For purposes of contradiction, suppose $\sigma_S^*(1, d) < \sigma_S^*(0, d)$, which implies by Assumption 1 that $0 \leq \beta(\sigma_S^*(1, d)) \leq \beta(\sigma_S^*(0, d)) \leq 1$ and thus that (a) $c^{dec}(\beta(\sigma_S^*(0, d)), 0) \geq c^{dec}(\beta(\sigma_S^*(1, d)), 0)$ and (b) $c^{dec}(\beta(\sigma_S^*(1, d)), 1) \geq c^{dec}(\beta(\sigma_S^*(0, d)), 1)$. If $u^g(\sigma_S^*(1, d)) - c^{lie}(\sigma_S^*(1, d), d) \geq u^g(\sigma_S^*(0, d)) - c^{lie}(\sigma_S^*(0, d), d)$, then (a) implies that $\phi^g(\sigma_S^*(1, d), 0, d) \geq \phi^g(\sigma_S^*(0, d), 0, d)$, contradicting the strict optimality of $\sigma_S^*(0, d)$ at $(0, d)$. If $u^g(\sigma_S^*(1, d)) - c^{lie}(\sigma_S^*(1, d), d) \leq u^g(\sigma_S^*(0, d)) - c^{lie}(\sigma_S^*(0, d), d)$, then (b) implies that $\phi(\sigma_S^*(0, d), 1, d) \geq \phi(\sigma_S^*(1, d), 1, d)$, contradicting the strict optimality of $\sigma_S^*(1, d)$ at $(1, d)$.

(ii): We begin by showing that $\sigma_S^{*1}(1, d) \geq \sigma_S^{*2}(1, d)$. First, to simplify notation, let $\sigma^i := \sigma_S^{*i}(1, d)$, $u^g(i) := u^g(\sigma_S^{*i}(1, d))$, $k(i) := c^{lie}(\sigma_S^{*i}(1, d), d)$, and $c_j(i) :=$

[22] *Deception is more costly for 1 than 2* if $c_1^{dec}(\beta(m), \theta) = f(|\beta(m) - \theta|)$ and $c_2^{dec}(\beta(m), \theta) = g(|\beta(m) - \theta|)$ where, for every $\delta \geq 0$, $f(\delta + \epsilon) - f(\delta) \geq g(\delta + \epsilon) - g(\delta)$ for every $\epsilon \geq 0$.

$c_j^{dec}(\beta(\sigma_S^{*i}(1,d)),1)$ (the deception cost j would incur by sending i's message). For purposes of contradiction, suppose $\sigma^1 < \sigma^2$, which by Assumption 1 implies that 1's message is more deceptive than 2's. By strict optimality of 1's action, we have that $u^g(1)-k(1)-c_1(1) > u^g(2)-k(2)-c_1(2) \iff u^g(1)-k(1)-u^g(2)+k(2) > c_1(1)-c_1(2)$. By strict optimality of 2's action, we have that $u^g(2)-k(2)-c_2(2) > u^g(1)-k(1)-c_2(1) \iff u^g(1)-k(1)-u^g(2)+k(2) < c_2(1)-c_2(2)$. But then we have that $c_1(1)-c_1(2) < c_2(1)-c_2(2)$, which violates the assumption that 1 finds deception more costly than 2. Hence, it must be that $\sigma_S^{*1}(1,d) \geq \sigma_S^{*2}(1,d)$. A symmetric argument yields $\sigma_S^{*1}(0,d) \leq \sigma_S^{*2}(0,d)$, and therefore that $\sigma_S^{*1}(1,d)-\sigma_S^{*1}(0,d) \geq \sigma_S^{*2}(1,d)-\sigma_S^{*2}(0,d)$. □

To gain some intuition, consider the special case of a purely deception-averse sender (i.e. with no lying costs). Under Assumption 1, she will always send the highest message T when $\theta = 1$ because this both maximizes expected material payoffs and minimizes deception costs. On the other hand, when $\theta = 0$, there is a tradeoff: higher messages result in greater material payoffs but also incur greater deception costs. Thus, following $\theta = 0$, the costlier is deception, the lower is the message sent. In the extreme case, if deception is sufficiently costly, the sender will always message 1 following $\theta = 0$. We summarize this in the following corollary.

Corollary 2. *Suppose Assumption 1 holds. (i) A purely deception-averse sender (i.e. with no lying costs) will use the strategy $\sigma_S^*(1,d) = T$ and $\sigma_S^*(0,d) = x$ for all d and some x. (ii) Let 1 and 2 be two purely deception-averse senders with the same second-order beliefs β. If deception is more costly for 1 than 2, then $\sigma_S^{*1}(0,d) \leq \sigma_S^{*2}(0,d)$ for all d.*

To summarize, coin effects indicate deception aversion and die effects indicate lying aversion. With additional monotonicity assumptions on β and c^{dec}, any coin effects will have a positive sign whose magnitude increases in the degree of deception aversion.

The results of this section are quite general. Proposition 2, Proposition 3, and Corollary 1 depend only on the separability of the lying and deception cost terms, while Proposition 4 and Corollary 2 require the additional Assumption 1. No results, however, require any assumption on the shape of lying costs c^{lie}. We offer four remarks.

Remark 3. All results in this section go through under various generalizations of the game itself. First, the sender's expected material utility $u(m;\beta)$ can be left completely general (i.e. any function of m, parameterized by β). Second, because the sender's

optimal strategy does not depend on the joint distribution of Θ and D *conditional on* second-order beliefs, the results are also robust to arbitrary correlations between Θ and D. We are interested in the positive correlation case because, as we argue in Section 3, this gives rise to very regular second-order beliefs that typically satisfy Assumption 1.

Remark 4. An alternative formulation of deception costs would be to replace θ in the definition of $c^{dec}(\beta(m),\theta)$ with $\beta^*(\theta) := \beta(m'') : m'' \in \underset{m' \in \{1,...,T\}}{argmin} |\beta(m') - \theta|$, i.e. the *most accurate* belief the sender believes she can induce. However, the two formulations are behaviorally equivalent.[23] We favor ours only because it simplifies notation and makes some results more transparent.

Remark 5. While we have only defined deception and deception costs with respect to the payoff-relevant state Θ, our notion of lying costs is general enough to also encompass deception costs with respect to the non-payoff-relevant state D.[24] Allowing for such costs, all results go through as stated, except we should replace "deception aversion" with "Θ-deception aversion" (and "lying aversion" with "lying aversion and/or D-deception aversion"). In such case, our results allow us to separately identify Θ-deception aversion.

Remark 6. We have imposed that, after sending message m, the sender believes that the receiver will form belief $\beta(m) \in [0,1]$ with probability one. This can be relaxed by allowing $\beta(m)$ to be a random variable supported on $[0,1]$ with associated Borel measure $\mu(\cdot|m) \in \Delta[0,1]$. Letting $\beta' \in [0,1]$ be an arbitrary *realization* of $\beta(m)$, define *expected* deception costs $\bar{c}^{dec}(\beta(m),\theta) := \int c^{dec}(\beta',\theta) d\mu(\beta'|m)$. After generalizing Assumption 1,[25] all results go through as stated, except with $\bar{c}^{dec}(\beta(m),\theta)$ replacing $c^{dec}(\beta(m),\theta)$.

[23] Fix β, which induces $\beta^*(\theta)$. To go from $\tilde{c}^{dec}(\beta(m),\beta^*(\theta))$ to $c^{dec}(\beta(m),\theta)$, set $c^{dec}(\beta(m),\theta) = \tilde{c}^{dec}(\beta(m),\beta^*(\theta))$ for all m and θ. To go from $c^{dec}(\beta(m),\theta)$ to $\tilde{c}^{dec}(\beta(m),\beta^*(\theta))$, consider two cases. First, if $\beta^*(0) \neq \beta^*(1)$, set $\tilde{c}^{dec}(\beta(m),\beta^*(\theta)) = c^{dec}(\beta(m),\theta)$ for all m and θ. Second, if $\beta^*(0) = \beta^*(1) := \beta^*$, then $\beta(m) = \beta^*$ for all m. In such case, both $c^{dec}(\beta(m),\theta)$ and $c^{dec}(\beta(m),\beta^*(\theta))$ are constant in m for any given θ, and therefore behaviorally equivalent.

[24] Let $\gamma : \{1,...,T\} \to \Delta\{1,...,T\}$ be the sender's exogenously specified *second-order belief* of the receiver's belief about D after sending message m. Given γ, we may suppose that $c^{lie}(m,d) = h(m,\gamma(m),d)$ for some function h.

[25] **Assumption 1*.** (1) $m' > m \implies \beta(m') \succ_{fosd} \beta(m)$, and (2) $\bar{c}^{dec}(\beta(m),\theta) = \int c^{dec}(\beta',\theta) d\mu(\beta'|m)$ where $c^{dec}(\beta',\theta) = f(|\beta' - \theta|)$ for (weakly) increasing f.

2 Experimental Design

The experiment consisted of three treatments, referred to as *message-payoff*, *belief-payoff*, and *pure lying*. Each corresponds to a different version of a sender-receiver game, and each subject participated in exactly one treatment, i.e. a between-subject design. All sessions followed the same basic structure, with one section of 41 rounds, followed by a questionnaire. We first describe the structure common to all treatments.

Common structure of all treatments.

Private information. The sender's private information is identical in all treatments. The sender privately observes (1) the outcome of a coin-flip—heads or tails—both equally likely and (2) the outcome of an 8-sided die roll. The coin and die are positively correlated, with higher die outcomes more likely after observing heads and less likely after observing tails. The conditional distributions of die outcomes following heads and tails are upper- and lower-triangular, respectively, as depicted in Figure 1. This information was communicated to subjects via this exact image as well as through tables. In terms of the theoretical framework of Section 1, $T = 8$, $\theta = 1$ corresponds to heads, $\theta = 0$ corresponds to tails, $p_\Theta(1) = \frac{1}{2}$, $p_{D|1}(d) = (8+d)/100$, and $p_{D|0}(d) = (17-d)/100$. By using an 8-sided die, as opposed to the more standard 6-sided, the triangular distributions can be expressed using whole number percentage-probabilities.

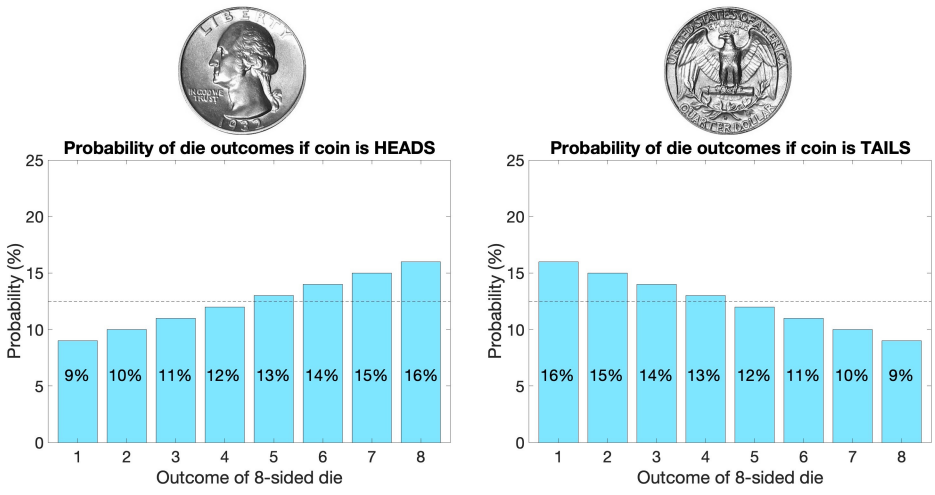


Figure 1: *Distribution of coin and die outcomes.* This figure depicts the conditional distributions (PMFs) of die outcomes following heads and tails, respectively.

Rounds 1-40. In each of the first 40 rounds, all subjects played in the role of the sender. In each round, subjects privately observed new coin and die outcomes (independent of previous rounds and of other subjects' outcomes). After observing this information, subjects sent a message stating "*The outcome of my die roll is _.*" where they were allowed to send any number 1 through 8. Importantly, messages were *not* immediately transmitted. It was only in round 41 that *one* of each subject's 40 messages was randomly selected and shown to another subject in the role of the receiver. Hence, there is no feedback and the games are strategically one-shot.

As the private information and action space of the sender are the same in all treatments, the experimental interface in rounds 1-40 was identical across treatments.

Round 41. In round 41, all subjects played in the role of the receiver. Each subject observed *one* message that was sent by a randomly selected (other) subject in a randomly selected previous round. Subjects were told "*A participant (other than you) and a round are randomly selected. In this round, the participant sent the message: "The outcome of my die roll is _."*" After observing this message, subjects were prompted to take a single action, which was to either (1) report a belief over the probability the coin was heads (in the message-payoff and belief-payoff treatments) or (2) simply click a button to acknowledge having observed the message (in the pure lying treatment).

Questionnaire. After the main experiment, there was a 24-question questionnaire.[26] Most importantly, immediately after round 41, there were eight (unincentivized) questions that asked for what the subjects' beliefs *would have been* had they observed each of the eight possible messages.[27] We did this for all three treatments, even though beliefs were not elicited in round 41 of the pure lying treatment. We also include some free response questions, some questions to measure strategic sophistication, demographics, and a 10-item Raven's matrix test of cognitive ability (Raven [1936]).

Message-payoff treatment. In the message-payoff game, the sender's payoff depends only on the message that is randomly selected and shown to the receiver. The payment, which is in terms of a probability of earning \$10,[28] is proportional to the ran-

[26] One message-payoff session and one belief-payoff session involved slightly different comprehension questions and a longer questionnaire of 30 questions that were streamlined for the later sessions. We describe here the questions that are common to all sessions.

[27] These were presented as eight separate questions, one for each message (including the same message that appeared in round 41). Each question appeared in random order on a separate screen that looked identical to the screen in round 41 of the message-payoff and belief-payoff treatments.

[28] All payments are in "probability points." Our main motivation for this is that, in order to use the

domly selected message, as shown in Table 1. In terms of the theoretical framework of Section 1, $u^M(m) = 0.32 + 0.04m$.

The receiver's task and payment were as follows. After being shown a message in round 41, subjects were asked "*What is your guess for the probability this participant's coin was HEADS in this round?*". Subjects then reported their beliefs, expressed as probabilities in percentage terms. Beliefs were incentivized via the binarized scoring rule (BSR) (Hossain and Okui [2013]), resulting in a probability of earning \$10. This ensures that a subject's subjective probability of earning \$10 was uniquely maximized by reporting her exact belief. The BSR is incentive-compatible for any risk attitude.

	Message							
	1	2	3	4	5	6	7	8
Sender's payment:	36%	40%	44%	48%	52%	56%	60%	64%

Table 1: *Message-payoff game—sender's payment.* This table gives the probability the sender earns \$10 as a function of the randomly selected message from rounds 1-40.

Belief-payoff treatment. In the belief-payoff game, the sender earns a higher payoff if the randomly selected message convinces the receiver that her coin was heads. Specifically, the sender earns \$10 with a probability that equals the receiver's reported belief. In the language of the theoretical framework, $u^B(b) = b$.

We note that for senders who believe the receiver is Bayesian and credulous—i.e. believing every message is sent truthfully—her second-order beliefs become $\beta(m) = 0.32 + 0.04m$ and her subjective expected payoffs coincide with those in the message-payoff game. We refer to this as the *truthful/Bayesian* benchmark.

The task and payment of the receiver are exactly as in the message-payoff game, based on reported beliefs and the BSR. Hence, the screens shown to subjects in round 41 are identical across message-payoff and belief-payoff treatments.

Pure lying treatment. The pure lying game is a *control game* that offers the opportunity to lie, but not to deceive: the sender has the same private information, but the coin is no longer payoff-relevant. Instead, the payoffs to both sender and receiver are a function of the randomly selected message. Hence, the sender directly determines her desired sender-receiver allocation, but she may have to lie in order to implement it.

binarized scoring rule, the receiver's payoffs are necessarily in probability points, so this ensures that sender and receiver payoffs are in the same units. This also has the potential to mitigate the effects of risk aversion as expected utility is linear in probability points.

The exact payoffs are in Table 2. The sender's payoffs are exactly as in the message-payoff game, such that a higher message yields a higher probability of earning \$10: $u^P(m) = 0.32 + 0.04m$. For the receiver, a higher message yields a *lower* probability of earning \$10, according to the payoff function $u_R^P(m) = 0.68 - 0.04m$. Note that $u^P(m) + u_R^P(m) = 1$ for all m, but the realization of payments is independent across sender and receiver. Instead of reporting a belief, receivers are asked to simply click a button to acknowledge they observed the message. In this way, messages are transmitted to the other party in a similar way across treatments, an important consideration for making the treatments more comparable in terms of lying costs.

	Message							
	1	2	3	4	5	6	7	8
Sender's payment:	36%	40%	44%	48%	52%	56%	60%	64%
Receiver's payment:	64%	60%	56%	52%	48%	44%	40%	36%

Table 2: *Pure lying game—sender and receiver payments*. This table gives the probabilities with which both sender and receiver earn \$10 as a function of the randomly selected message from rounds 1-40.

The reason for running the pure lying treatment is twofold. First, because the coin is not payoff-relevant, there is no opportunity to deceive and so even a deception-averse sender should not exhibit coin effects. However, since higher die outcomes are more likely following heads, some subjects may feel "entitled" to higher payment and therefore send higher messages following heads. Hence, the pure lying game serves as a control that allows us to determine if coin effects in the other treatments are the result of genuine deception aversion as opposed to such *entitlement effects* (Hoffman et al. [1994]). Second, because deception does hurt the receiver, an important question is whether the effects we observe are due to an aversion to deception per se, or if they are fully driven by *social* or *other-regarding* preferences. In Section 6, we show how the pure lying game can be used as a control for testing this alternative hypothesis as well.

Logistical details. Our sessions were run at the Economic Research Laboratory (ERL) of Texas A&M University. Subjects were undergraduate students, recruited via the ORSEE software (Greiner [2015]). We had 70, 82, and 50 subjects in each of message-payoff, belief-payoff, and pure lying treatments respectively.[29] The average payment was

[29]We aimed for 80, 80, and 50 subjects, respectively (fewer in the pure lying treatment as it is the

$22 for an experiment lasting approximately 50 minutes. Experimental instructions and screenshots are given in Online Appendices 9.5 and 9.6.

Discussion of the experimental design. A few features of our design merit discussion. Firstly, for the large majority of the experiment—the first 40 of 41 rounds—subjects play as the sender. We do this because we are primarily interested in sender strategies, due to our focus on documenting lying and deception aversion. With 40 rounds, simulations (unreported) suggest that 98% of subjects will realize at least 5 distinct die outcomes for which both heads and tails are realized (6.7 on average). Hence, we are able to have good estimates of individual-level coin effects, which require both heads and tails realizations for a given die outcome. Alternatively, we could have used the strategy method, but this may have led to demand effects.

It is important that exactly *one* message is transmitted to the receiver. If multiple messages are transmitted, then the sender might have incentive to create a "portfolio" of messages to appear more credible. Consider, for example, a subject who always messages 8—the payoff-maximizing choice in the message-payoff game. This would reveal to the receiver that the message is not very informative. Hence, if multiple messages are transmitted, it would likely lead to mixing and data that is hard to interpret: indeed, second-order beliefs would be a mapping from the entire set of transmitted messages. However, by including the eight belief questions in the questionnaire, we observe the beliefs the receiver would have reported if they had seen each of the eight messages.

Because of the important conceptual role of second-order beliefs, we could have elicited second-order beliefs directly as part of our questionnaire. However, this would have been difficult to incentivize and explain to subjects. Moreover, it is not necessary for any of our main analyses, which our theoretical results in Section 1 make clear.

Finally, our theoretical analysis depends on an "accepted meaning of messages," in particular that messages are commonly understood as declaring a die outcome. It is for this reason that we force messages to take the form *"The outcome of my die roll is _."* as opposed to simply reporting a number. In our view, this, and the fact that we do not provide feedback, makes adopting other languages implausible.

control). We ran 5 sessions for each of the message-payoff and belief-payoff treatments and 3 sessions of the pure lying treatment. The exact number of subjects differed from our targets due to randomness in subject recruiting.

3 Beliefs

Before examining the empirical sender strategies, we analyze the receivers' first-order beliefs, which we think of as a rough proxy for the senders' second-order beliefs. While our main analysis does not depend on identifying beliefs, establishing their general pattern will help us to interpret later results.

Figure 2 plots, for all individual receiver-subjects, their beliefs following every message, as reported in the eight questions of the questionnaire. Each of the first three panels corresponds to a different treatment. The bottom-right panel shows that the average beliefs are very similar across treatments.

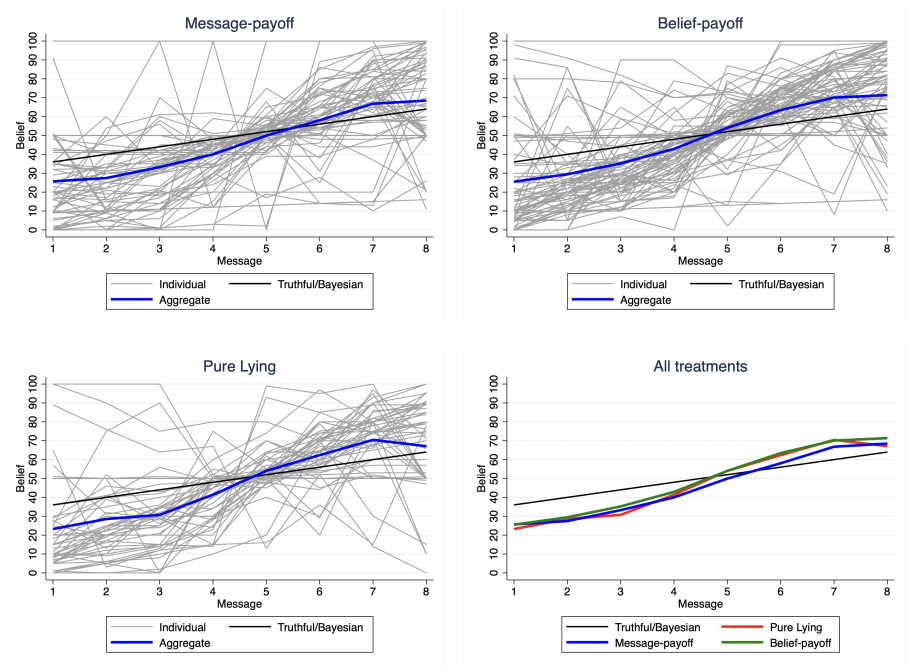


Figure 2: *Elicited first-order beliefs.* The top-left, top-right, and bottom-left panels plot, for each treatment, all individual subjects' elicited first-order beliefs in the receiver role following each of the eight messages, based on the eight questions from the questionnaire. Each panel gives the average belief (blue) and the truthful/Bayesian benchmark—the beliefs formed by a perfectly Bayesian subject who believes the sender always tells the truth. The bottom right panel compares the average beliefs across all treatments.

We note that the eight belief questions, one for each message, appeared in random order and on separate screens. Compared to the strategy method, this likely increases

noise, but reduces demand effects pushing subjects toward any specific pattern of beliefs (e.g. monotonically increasing). Nevertheless, we see significant regularities.

While there is substantial heterogeneity across subjects, a large majority of subjects report beliefs that are perfectly or approximately monotonically increasing in the message. The average beliefs are also perfectly monotonic in the message-payoff and belief-payoff treatments, though average beliefs are very similar following messages 7 and 8. In the pure lying treatment, average beliefs are monotonic, except for a slight dip in beliefs in going from message 7 to message 8.

Inspecting individual subjects' beliefs more carefully, we find that while most subjects have approximately monotonic beliefs, about 17% of subjects have beliefs that are approximately hump-shaped. For most of these subjects, beliefs peak at 7, but for some the peak is at 6. We plot these subjects' beliefs in Figure 3 and find that their average beliefs peak at 7, with a very pronounced dip in going from 7 to 8. Indeed, for these subjects, the average belief following 8 is lower than that following 6 in all treatments.

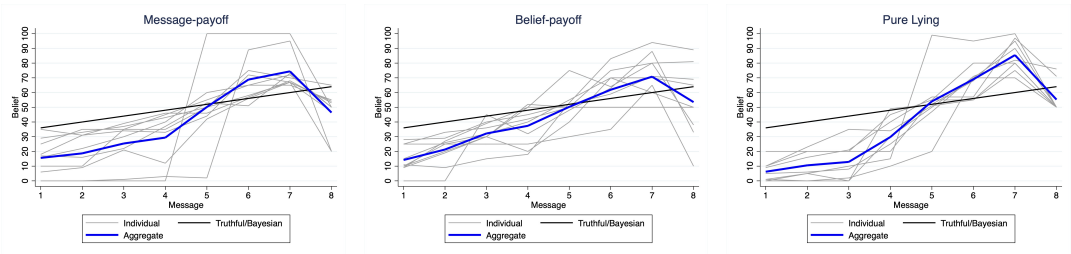


Figure 3: *Elicited first-order beliefs of subjects with hump-shaped beliefs.* We plot here the beliefs for the approximately 17% of subjects who have hump-shaped beliefs. Each treatment has a similar percentage of such subjects.

That beliefs tend to be monotonically increasing means that we expect most deception-averse subjects to send higher messages following heads (Part (i) of Proposition 4), i.e. a positive coin effect. Moreover, monotonic beliefs are not necessary for positive coin effects. Consider the typical pattern for subjects with hump-shaped beliefs: beliefs are minimized at 1 and maximized at 7. A subject with such second-order beliefs who is infinitely deception-averse will message 1 when tails and 7 when heads—a large positive coin effect. Hence, while such subjects may exhibit negative coin effects, they may only do so if they have partial deception aversion. Moreover, under part (2) of

Assumption 1, the only possible negative coin effects for such subjects involve messaging 8 when tails and 7 when heads. Hence, scope for negative coin effects is limited.

3.1 A simple model of boundedly rational beliefs

Our main empirical analysis will not make use of any particular model of beliefs. Indeed, a virtue of the theory developed in Section 1 is that it is agnostic about where the beliefs come from. However, the regularities observed in the beliefs data are consistent with a certain family of models, which we offer as a means to better understand the data.

In Appendix 9.1, we show that the beliefs data are well-described by models in which beliefs are anchored on the (perhaps) naive assumption that senders have a tendency to tell the truth. We formalize this using a level k-type model in the spirit of Wang et al. [2010] with level 0-senders that are *truthful* and level 0-receivers that are *credulous*. We show that, due to the positive correlation between coin and die, this model generates strictly increasing second-order beliefs if there is sufficient naïveté. Moreover, it also predicts that sufficiently sophisticated senders will have hump-shaped second-order beliefs that peak at 7. Hence, the model captures the main qualitative patterns in the data.

We also note that, because the sender's subjective expected payoffs coincide in all three games under the *truthful/Bayesian* benchmark (as discussed in Section 2), the model also helps to rationalize why beliefs are so similar across treatments.

We emphasize that, regardless of the "true" model of belief formation, the games do lead to significant empirical regularities in beliefs. This is not required to identify deception aversion, but it does make it easier to attribute differences across subjects' behavior to differences in deception aversion (e.g., as in Part (ii) of Proposition 4).

4 Reduced-form results

4.1 Aggregate sender strategies

Figure 4 summarizes the aggregate sender strategies for each treatment. The top three panels plot the average messages conditional on coin and die outcome (red for heads; blue for tails), and the bottom three panels plot histograms of messages conditional on coin outcome only.

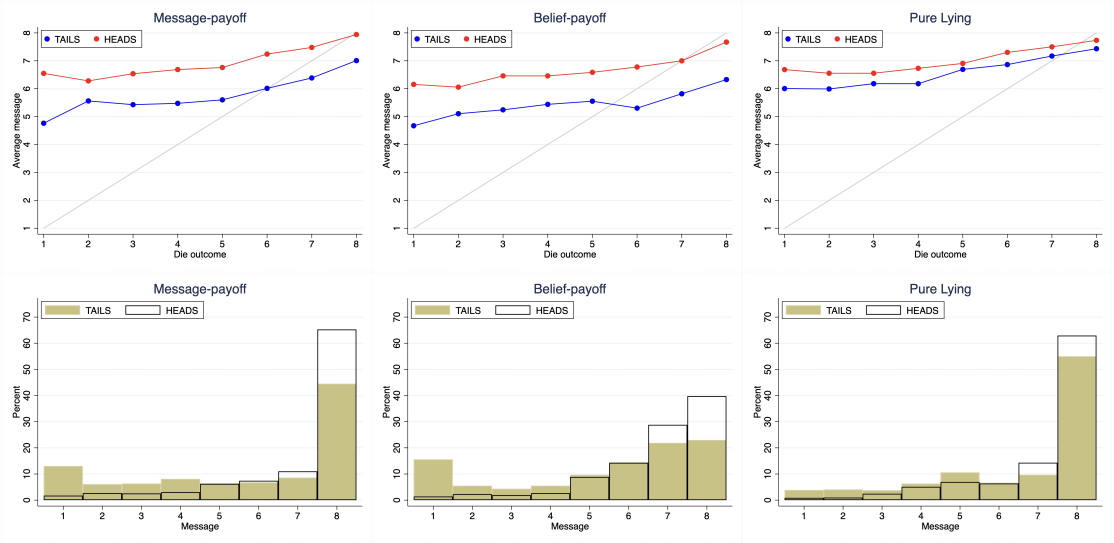


Figure 4: *Average messages conditional on coin and die; distribution of messages conditional on coin.*

From the top panels, there is a clear positive coin effect in both message-payoff and belief-payoff treatments, consistent with deception aversion and increasing second-order beliefs (Proposition 4). There is also a coin effect in the pure lying treatment, but it is much smaller. In columns (1)-(3) of Table 3, we quantify these coin effects by regressing the message sent on an indicator for heads, conditioning on the die outcome (i.e. with 8 dummy variables). We find average coin effects of 1.16, 1.21, and 0.43 in the message-payoff, belief-payoff, and pure lying treatments, respectively, all of which are highly significant. The coin effects in message-payoff and belief-payoff are more than twice the size of those in pure lying, and we find that the difference in coin effects is highly significant. Hence, we find strong evidence for positive coin effects that can be explained by deception aversion, but not entitlement effects (as described in Section 2). These results are reported in columns (4)-(6) of Table 3.

The top panels of Figure 4 show that there is also substantial lying in all treatments. This can be seen by comparing the average messages following each die outcome to the truthful benchmark (upward-sloping gray line). On average, deviations tend to be upward, consistent with lying to increase material payoffs. Despite the significant amount of lying, there are also clear die effects in all treatments, consistent with lying aversion (Proposition 3). In particular, the average messages tend to be approximately

	Coin effect			Difference in coin effects		
	MP	BP	PL	MP vs. PL	BP vs. PL	MP+BP vs. PL
	(1)	(2)	(3)	(4)	(5)	(6)
Heads	1.159***	1.208***	0.430**	0.430**	0.430**	0.430**
	(0.241)	(0.245)	(0.164)	(0.163)	(0.163)	(0.163)
Heads × treatment				0.730**	0.778**	0.761**
				(0.290)	(0.293)	(0.237)
Observations	2800	3280	2000	4800	5280	8080

Table 3: *Coin effects: reduced-form evidence of deception aversion.* In columns (1)-(3), we regress the message sent on indicators for whether the coin was heads, conditioning on the die outcome (the 8 die dummies are omitted); columns (1), (2), and (3) correspond to message-payoff, belief-payoff, and pure lying treatments, respectively. The coefficient on "Heads" is the coin effect, and we observe statistically significant positive effects in all treatments. In columns (4)-(6), we consider the *difference in* coin effects between a given treatment and the pure lying treatment. For this, we regress the message sent on indicators for whether the coin was heads and the interaction between the coin being heads and an indicator for the given treatment, conditioning on the die outcome and the die outcome interacted with the treatment indicator (the 16 dummies are omitted). Column (4) compares message-payoff to pure lying, column (5) compares belief-payoff to pure lying, and column (6) compares message-payoff and belief-payoff pooled together to pure lying. The coefficient on "Heads × treatment" gives the difference in coin effects between the given treatment and pure lying. We observe statistically significant positive effects, indicating that the coin effect is larger in all treatments relative to pure lying. Standard errors are clustered at the subject level.

monotonically increasing in the die outcome.

To better understand which messages drive the observed coin effects, we next study the distributions of messages conditional on coin outcome (bottom panels of Figure 4).

Message-payoff. We first consider message-payoff (bottom-left panel). *Following heads*, higher messages are sent more often: message 8 is sent an overwhelming 69% of the time, messages 5-7 are sent between 6 and 10% of the time, and messages 1-4 are each sent less than 2% of the time. Hence, subjects strongly favor messaging 8 and rarely ever send messages 1-4, presumably those that they believe signal that the coin is tails more likely than not. *Following tails*, we see that 8 is still the most common message, being sent 49% of the time. There is thus a 20 percentage point (p.p.) gap in sending 8 following heads versus tails, which drives much of the average coin effect. Following tails, it is no longer the case that the higher the message, the more often it

is sent: 1 is the second most common message, sent about 12% of the time (as opposed to 1% following heads). The other messages, 2-7, are sent fairly uniformly. We do see, however, that message 4 is sent more often than the nearby messages 3 and 5. Rather than being due to randomness, we argue in Section 5 that this is a purposeful decision for some subjects that is actually intuitive from the perspective of deception aversion.

Belief-payoff. In belief-payoff (bottom-middle panel), we see a similar pattern at the extremes: message 8 is 17 p.p. more likely following heads and message 1 is 13 p.p. more likely following tails. However, while 8 remains the most commonly used message following both heads and tails, much of the mass has been shifted from 8 to 7 and to some extent 6 and 5. What accounts for this downward shift? Noting that senders with hump-shaped second-order beliefs maximize expected material payoffs by sending message 8 in message-payoff and lower messages in belief-payoff (corresponding to the peak of second-order beliefs), this is likely the consequence of rational, self-interested behavior.

Pure Lying. In pure lying (bottom-right panel), we see that there is a much smaller coin effect, and that it is much less pronounced at 8. Notably, we no longer observe any strong tendency to message 1, which is sent only 3% of the time following tails and 1% of the time following heads. Unlike in the other treatments, message 5 is sent more often than neighboring messages 4 and 6, about 10% of the time following tails and 6% of the time following heads. While 5 is still sent less often than implied by the truthful benchmark, its use does suggest there are a small number of inequality-averse subjects as it leads to the sender-preferred inequality-minimizing outcome.

4.2 Heterogeneity in individual sender strategies

We now explore the distribution of deception and lying aversion in the population of subjects. Because coin effects indicate deception aversion and die effects indicate lying aversion, we are able to summarize heterogeneity through two reduced-form statistics.

First, we consider the *average coin effect* for individual i:

$$C_i = \frac{\sum_{d'} \mathbf{1}\{i \text{ observes both } (1,d') \text{ and } (0,d')\}(\bar{\sigma}_S^i(1,d') - \bar{\sigma}_S^i(0,d'))}{|\{d|i \text{ observes both } (1,d) \text{ and } (0,d)\}|},$$

where $\bar{\sigma}_S^i(\theta, d')$ is the average message following coin-die realization (θ, d') and

$(\bar{\sigma}_S^i(1,d') - \bar{\sigma}_S^i(0,d'))$ is the *coin effect conditional on* d'. Hence, C_i is the equal-weighted average of conditional coin effects over all d' for which both $(1,d')$ and $(0,d')$ are observed.

Second, we consider the *average lie* for individual i:

$$L_i = \frac{\sum_{d'} \mathbf{1}\{i \text{ observes } d'\}(\bar{\sigma}_S^i(d') - d')}{|\{d|i \text{ observes } d\}|},$$

where $\bar{\sigma}_S^i(d')$ is the average message conditional on d'.

To better understand these statistics, consider the statistics implied by a few benchmark strategies.[30] The *truthful* strategy, employed by a sender with very strong lying aversion, yields $(L_i, C_i) = (0,0)$. The *always-8* strategy, which is the self-interested or material payoff-maximizing strategy in the message-payoff game and in the belief-payoff game under increasing second-order beliefs, yields $(L_i, C_i) = (3.5, 0)$. *Always-7* yields $(L_i, C_i) = (2.5, 0)$. This may be used by self-interested senders in the belief-payoff game if they have hump-shaped second-order beliefs. The strategy of messaging 8 following heads and 1 following tails, which we refer to as *(8,1)*, yields $(L_i, C_i) = (0, 7)$. This is the strategy of an infinitely deception-averse sender with increasing second-order beliefs. More generally, the family of strategies *(8,x)* for $x \in \{1, ..., 8\}$ yields $(L_i, C_i) = (\frac{1}{2}x - \frac{1}{2}, 8 - x)$. This family of strategies is used by purely deception-averse senders with varying degrees of deception aversion and increasing second-order beliefs. For fixed second-order beliefs, lower x implies greater deception aversion (Corollary 2).

Both C_i and L_i are signed. A subject for whom $C_i > 0$ ($C_i < 0$) exhibits a positive (negative) coin effect on average, and a subject for whom $L_i > 0$ ($L_i < 0$) tends to lie upward (downward) on average. Note that, by the L_i-measure, upward and downward lies cancel out, e.g. (8,1) implies $L_i = 0$ even though most messages are lies. Because both C_i and L_i are signed, random behavior will not systematically yield statistics of any

[30] These computations assume complete data—i.e. that every coin-die outcome is realized at least once. In practice, however, randomness in coin-die realizations will lead to incomplete data, and therefore slightly different statistics, even for a given pure strategy implemented without error. For example, a subject who always messages 8 and realizes all die outcomes will have $L_i = 3.5$, but a subject who always messages 8 and realizes all die outcomes other than 8, will have $L_i = 4$. Note, however, that (1) for strategies that depend on the coin only, or strategies that do not exhibit coin effects, C_i is the same for any data—complete or not; and (2) randomness will never result in $C_i \neq 0$ if the strategy does not involve any coin effect, nor $L_i \neq 0$ if the strategy does not involve lying.

particular sign, and noise will in fact attenuate the statistics toward zero. Conversely, if C_i or L_i tend to have a particular sign, this cannot be an artifact of unbiased noise.

Figure 5 plots (L_i, C_i) for all subjects. The left panel is for message-payoff (blue) and belief-payoff (green), and the right panel is for pure lying. From the left panel, we see that many subjects have positive coin effects in both message-payoff and belief-payoff, including a significant number with large coin effects, e.g. 24 subjects with C_i larger than 2. We see that very few subjects have sizable negative coin effects, e.g. only one subject with C_i less than -2. In pure lying (right panel), on the other hand, we see very few subjects with large effects. In fact, we see that the average coin effect of 0.43 reported in Section 4.1 is largely driven by two outlying subjects.

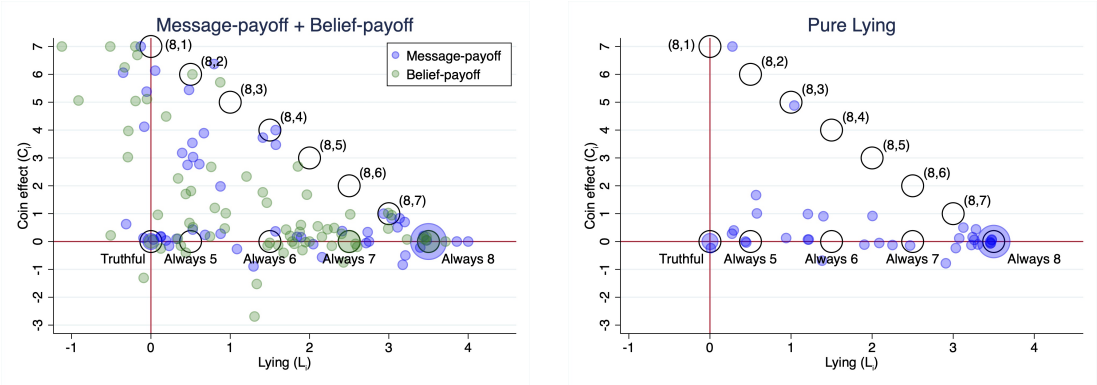


Figure 5: *Heterogeneity of individual coin effects and lying.*

To better appreciate differences across treatments, we plot the CDFs of individual coin effects C_i for each treatment in the left panel of Figure 6. This shows clearly that the distributions of individual coin effects in both message-payoff and belief-payoff stochastically dominate that in pure lying. This strengthens our previous conclusion that the positive coin effects we observe can be explained by deception aversion, but not so-called entitlement effects.

In the right panel of Figure 6, we plot CDFs of individual lying L_i and find that the distribution of lying in pure lying stochastically dominates that in the other treatments. Inspecting Figure 5, this difference is largely driven by a higher tendency to use the always-8 strategy in pure lying, as opposed to strategies that exhibit positive coin effects.

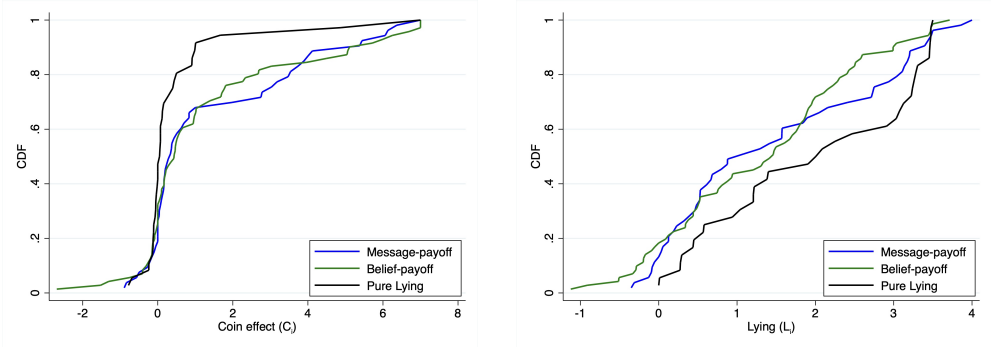


Figure 6: *Distribution of individual coin effects and lying.*

Inspecting Figure 5 more closely, there are clear clusters of truthful and always-8 strategies in all three treatments. In pure lying, no other strong clusters appear. In belief-payoff, there is a cluster of always-7, of similar magnitude as always-8. In message-payoff and belief-payoff, a good number of subjects use a strategy that is approximately (8,x) for different values of $x < 8$. In particular, there are several subjects close to (8,7), (8,4), and (8,1). There are also many subjects with clear positive coin effects, but who do not conform exactly to these simple strategies. As we show in the next section, these subjects use more complex strategies that feature both coin and die effects, suggesting a mix of both deception and lying aversion.

5 Structural model and estimates

Our structural model is based off of utility function (1). In our baseline specification, we assume that both lying and deception costs are linear:

$$\phi^g(m, \theta, d; \kappa^{lie}, \kappa^{dec}, \beta) = u^g(m; \beta) - \kappa^{lie}|m - d|/7 - \kappa^{dec}|\beta(m) - \theta|, \tag{2}$$

where $\kappa^{lie} \geq 0$ and $\kappa^{dec} \geq 0$ are the lying and deception cost parameters, respectively; and we recall that $u^M(m) = 0.32 + 0.04m$ in the message-payoff game and $u^B(m; \beta) = \beta(m)$ in the belief-payoff game. Note that lying costs depend on the *absolute* distance between the message and die outcome and deception costs depend on the *absolute* distance between the coin outcome and induced second-order beliefs.

In addition to the two cost parameters, which we will estimate, we must somehow

account for the unobserved second-order beliefs. We take an empirical approach, using the receivers' first-order beliefs to proxy for the senders' second-order beliefs. Because individual-level elicited beliefs are quite noisy for many subjects (Figure 2), we will take averages across subjects. However, the simple average across all subjects does not allow for any heterogeneity across individuals nor does it capture the fact that there are two distinct groups of subjects, those with monotonic beliefs and those with hump-shaped beliefs that peak at 7 (Figure 3). Hence, we take two averages: the average beliefs of subjects with hump-shaped beliefs and the average beliefs of all other subjects (in both cases pooling across message-payoff and belief-payoff). We then consider the family of belief functions $\beta_\alpha(\cdot)$ depicted in Figure 7: an α-weighted average of the two average belief functions, with α being the weight on the hump-shaped function. This family of belief functions is only one parameter, captures the important qualitative features in the data, and is closely related to the level k model of Online Appendix 9.1.[31]

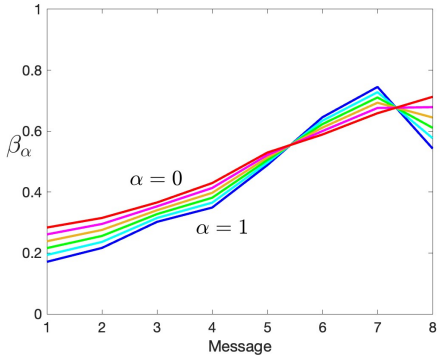


Figure 7: *Parametric family of second-order beliefs (β_α) based on the data.* $\beta_{\alpha=1}$ corresponds to the average beliefs of subjects with hump-shaped beliefs, $\beta_{\alpha=0}$ corresponds to the average beliefs of all other subjects, and $\alpha \in [0, 1]$ interpolates between the two functions.

Hence, we consider the utility function $\phi^g(m, \theta, d; \kappa^{lie}, \kappa^{dec}, \beta_\alpha)$ and assume a logit error structure with noise parameter $\lambda \geq 0$. For each individual i, we then estimate the

[31] Based on the level k model of Online Appendix 9.1, we think of $\beta_{\alpha=0}$ as the second-order beliefs of a level 1- or level 2-sender who believes the receiver is perfectly credulous (i.e. level 0 or 1), except the assumption of perfect Bayesianism is dropped in favor of belief updating more consistent with the data (i.e., overreaction to signals). Similarly, we think of $\beta_{\alpha=1}$ as the second-order beliefs of a level 3-sender who believes the receiver is level 2. Intermediate values of α interpolate between these two extreme second-order belief functions, and so higher α indicates a higher degree of strategic sophistication.

four parameters $(\kappa^{lie}, \kappa^{dec}, \alpha, \lambda)$ by maximizing the log-likelihood function

$$\mathcal{L}(\kappa^{lie}, \kappa^{dec}, \alpha, \lambda) = \sum_{t=1}^{40} ln \left(\frac{exp(\lambda \cdot \phi^g(m_t, \theta_t, d_t; \kappa^{lie}, \kappa^{dec}, \beta_\alpha))}{\sum_{m'=1}^{8} exp(\lambda \cdot \phi^g(m', \theta_t, d_t; \kappa^{lie}, \kappa^{dec}, \beta_\alpha))} \right),$$

where, for individual i in round $t \in \{1, ..., 40\}$, $(\theta_t, d_t) \in \{0, 1\} \times \{1, ..., 8\}$ is the coin-die realization and $m_t \in \{1, ..., 8\}$ is the message sent.

To determine which of κ^{lie} and κ^{dec} add explanatory power, we consider the full model $M_1 = (\kappa^{lie}, \kappa^{dec}, \alpha, \lambda)$ as well as three restricted models: $M_2 = (\kappa^{lie} = 0, \kappa^{dec}, \alpha, \lambda)$, $M_3 = (\kappa^{lie}, \kappa^{dec} = 0, \alpha, \lambda)$, and $M_4 = (\kappa^{lie} = 0, \kappa^{dec} = 0, \alpha, \lambda)$. We then perform standard likelihood ratio (LR) tests to determine the best-fitting model, penalizing models with more parameters.[32] If M_1 ($\kappa^{lie} > 0$ and $\kappa^{dec} > 0$) is the best model, we say that the individual is averse to both lying and deception. If M_2 ($\kappa^{lie} = 0$ and $\kappa^{dec} > 0$) is the best model, we say that the individual is only averse to deception. If M_3 ($\kappa^{lie} > 0$ and $\kappa^{dec} = 0$) is the best model, we say that the individual is only averse to lying. Otherwise, M_4 ($\kappa^{lie} = \kappa^{dec} = 0$) is the best model, and we conclude that the individual is averse to neither lying nor deception. Importantly, by using statistical tests, if a subject is classified as deception-averse (say), they are significantly so.

Game	Averse to:			
	Lying and deception ($\kappa^{lie} > 0$, $\kappa^{dec} > 0$)	Deception only ($\kappa^{lie} = 0$, $\kappa^{dec} > 0$)	Lying only ($\kappa^{lie} > 0$, $\kappa^{dec} = 0$)	Neither ($\kappa^{lie} = \kappa^{dec} = 0$)
Message-payoff	17% (12)	17% (12)	31% (22)	34% (24)
Belief-payoff	13% (11)	13% (11)	39% (32)	34% (28)
Pure lying	2% (1)	6% (3)	34% (17)	58% (29)

Table 4: *Summary of structural estimates.* This table summarizes the percentage of subjects (number of subjects in parenthesis) in each treatment who are best described by each model (after penalizing the number of parameters in the likelihood ratio sense). Hence, we estimate the percentage of subjects who are averse to lying and deception, deception only, lying only, and neither lying nor deception.

In Table 4, we give the percentage and number of subjects in each treatment who are best explained by each of the models. We do find slightly less deception aversion in belief-

[32] If $\mathcal{L}(M_1) > max\{\mathcal{L}(M_2), \mathcal{L}(M_3)\} + 3.841$ and $\mathcal{L}(M_1) > \mathcal{L}(M_4) + 5.991$, then M_1 is the best model. If $\mathcal{L}(M_2) > \mathcal{L}(M_3)$, $\mathcal{L}(M_1) < \mathcal{L}(M_2) + 3.841$, and $\mathcal{L}(M_2) > \mathcal{L}(M_4) + 3.841$, then M_2 is the best model. If $\mathcal{L}(M_3) > \mathcal{L}(M_2)$, $\mathcal{L}(M_1) < \mathcal{L}(M_3) + 3.841$, and $\mathcal{L}(M_3) > \mathcal{L}(M_4) + 3.841$, then M_3 is the best model. Otherwise, M_4 is the best model.

payoff compared to message-payoff, but the general patterns are similar: an incentive to deceive does *not* drive out deception aversion. Pooling across the two treatments, we find that that 15% of subjects are averse to both lying and deception, 15% are averse to deception only, 36% are averse to lying only, and 34% are averse to neither. Hence, 30% of subjects are significantly averse to deception.

As a placebo test, we fit the same model to the individual subject data in the pure lying treatment where there is no opportunity to deceive. Reassuringly, we find that only 8% (4 out of 50) are classified as deception-averse, a much smaller percentage than in the other treatments. We find that a similar percentage of subjects are purely lying-averse (34%) and a much higher percentage are neither lying- nor deception-averse (58%).

Robustness. In Online Appendix 9.3, we show that the results are robust to alternative structural models that vary in terms of the family of second-order beliefs considered as well as the shape of the lying and deception cost functions.

5.1 Individual subjects

To better visualize individual subject behavior and understand the implications of model (2), we plot the average messages, conditional on coin and die, for individual subjects, superimposed with the average messages predicted by the best-fitting model. Figure 8 gives plots for nine subjects in the message-payoff treatment, who are all well-captured by the model.

Referring to Figure 8, subject MP1 is classified as neither lying- nor deception-averse and therefore always messages 8 (always-8). MP2 is extremely lying-averse, and therefore always tells the truth (truthful). MP3 is extremely deception-averse, always messaging 8 when heads and 1 when tails ((8,1)). MP4 is a bit less deception-averse, always messaging 8 when heads and 2 when tails ((8,2)). MP5 is less deception-averse yet, always messaging 8 when heads and 4 when tails ((8,4)). This behavior, which we find in three subjects in the message-payoff treatment, is particularly interesting. We conjecture that, for these subjects, a message of 4 maximizes induced second-order beliefs among induced beliefs that favor tails. Hence, they are maximizing their payment, subject to not deceiving the receiver so much that their beliefs go "in the wrong direction." MP6 is only slightly deception-averse, always messaging 8 when heads and 7 when tails ((8,7)).

The remaining subjects, MP7-9, who exhibit both coin and die effects, are classified

as both deception- and lying-averse. They are ordered, from left to right, in order of decreasing deception aversion. It is interesting that the model can capture these rich patterns, including cases like MP9 where the sender's strategy involves very different "slopes" following heads versus tails.

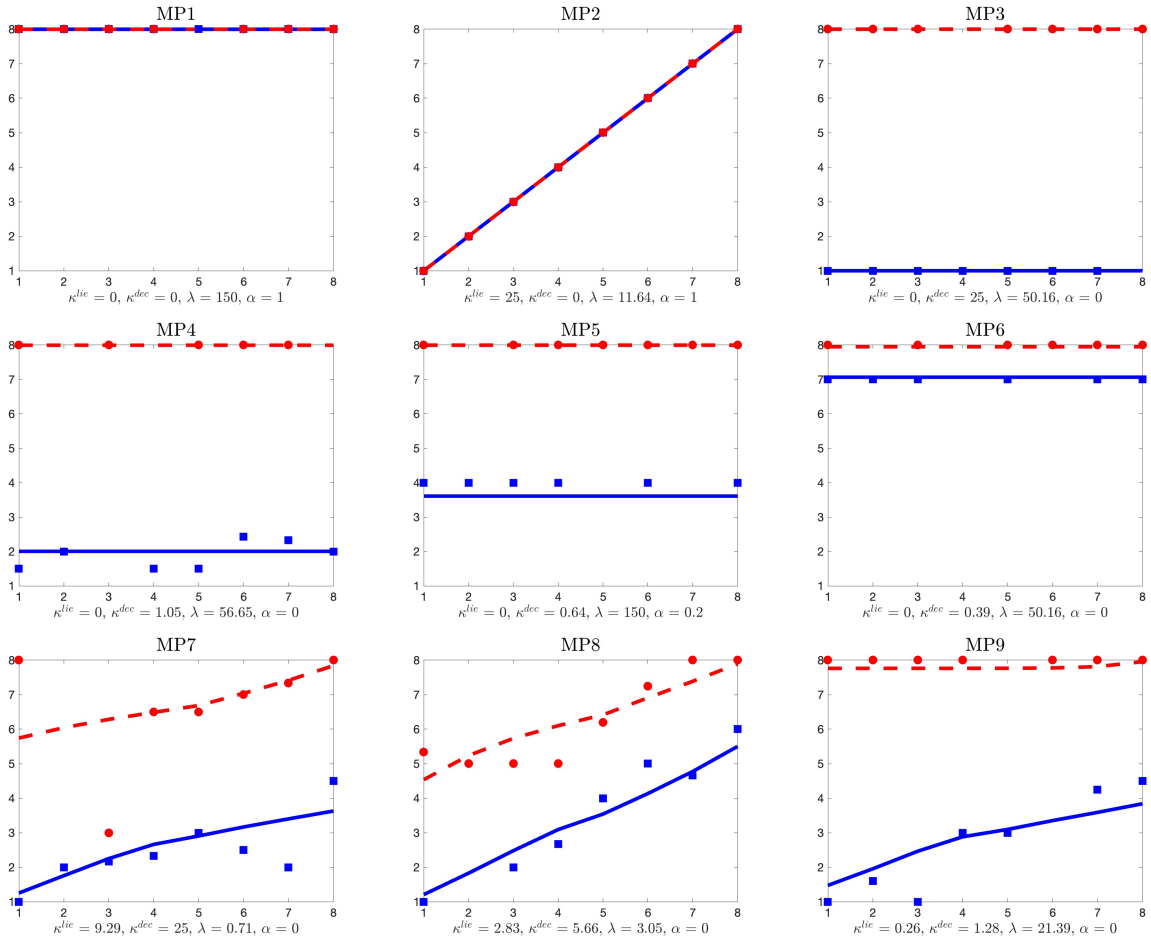


Figure 8: *Individual subjects (message-payoff).* Each panel gives an individual subject's average messages conditional on coin and die outcome, superimposed with the average messages predicted by the best-fitting model. Parameter estimates are given at the bottom of each panel.

Figure 9 gives plots for nine subjects in the belief-payoff treatment, who are also well-captured by the model. We do observe subjects in this treatment who are seemingly neither averse to lying nor deception (always-8), extremely lying-averse (truthful), or extremely deception-averse with increasing second-order beliefs ((8,1)), but such plots look identical to those in the message-payoff treatment, so we omit them here.

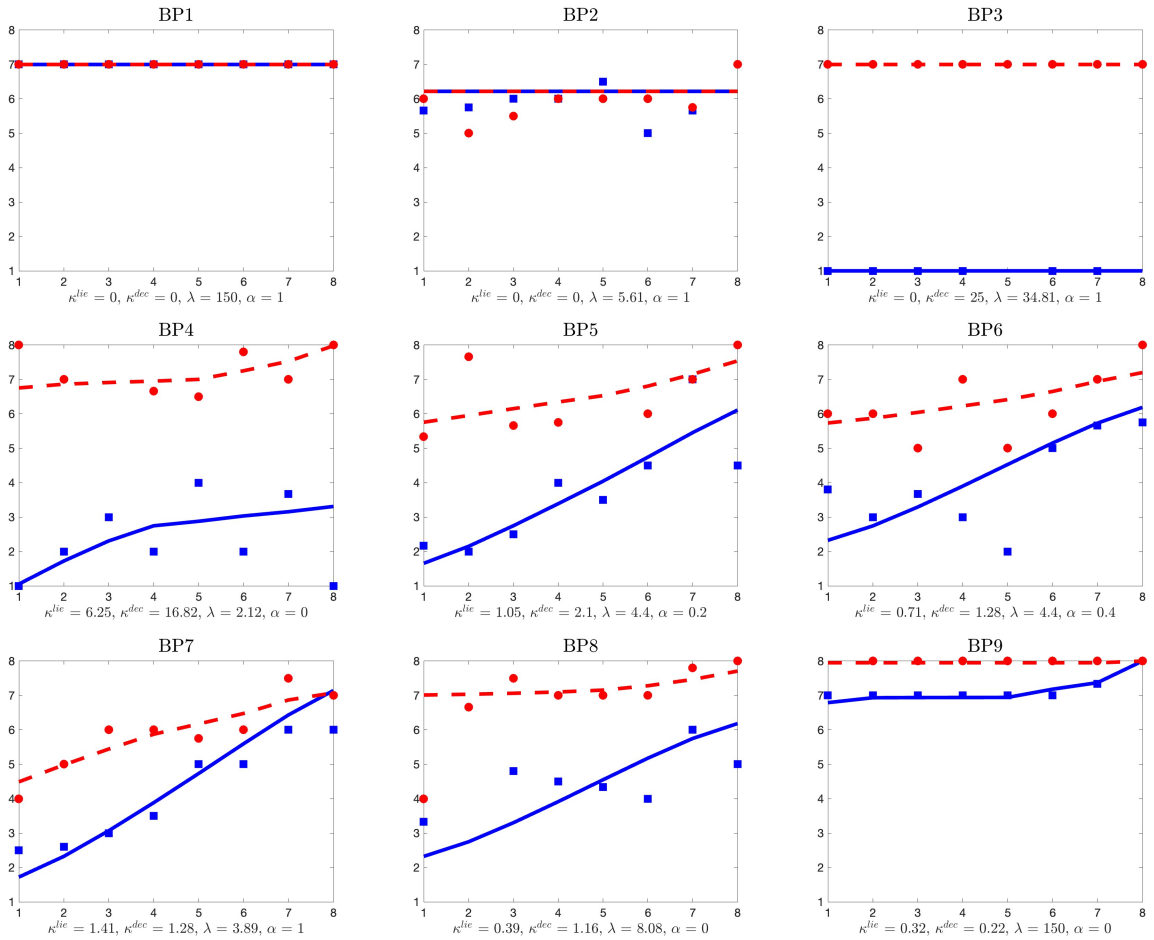


Figure 9: *Individual subjects (belief-payoff).* Each panel gives an individual subject's average messages conditional on coin and die outcome, superimposed with the average messages predicted by the best-fitting model. Parameter estimates are given at the bottom of each panel.

Referring to Figure 9, subject BP1 is classified as neither averse to lying nor deception, but is estimated to have hump-shaped second-order beliefs ($\alpha = 1$). She always messages 7 (always-7), which she believes maximizes the receiver's belief and thus her own payment. BP2 is also neither averse to lying nor deception, but tends to message 6 (always-6). This type of subject is not perfectly captured by the model as it does not allow for second-order beliefs that peak at 6. BP3 is extremely deception-averse, but with hump-shaped second-order beliefs ($\alpha = 1$). As such, she always messages 7 when heads and 1 when tails ((7,1)). The remaining subjects, BP4-9, are both lying- and deception-averse, arranged in order of decreasing deception aversion.

Unexplained patterns. The structural model provides a good description of the

data for many individual subjects, but it does not capture the complete richness. A small number of subjects exhibit die effects that cannot be captured qualitatively by the model, suggesting fixed or other non-linear lying costs (although these subjects are typically still classified as lying-averse). Another group of subjects exhibit clear positive coin effects, but are not classified as deception-averse. In our view, these subjects most likely *are* deception-averse, but the model is not flexible enough to capture the exact form in which it manifests. A third group of subjects exhibit patterns that suggest either a fixed cost of deception or result from what we believe are deliberate and intuitive, but ultimately misguided, deceptive heuristics. In Online Appendix 9.2, we discuss these patterns in detail.

6 Social preferences

By engaging in deception, the sender believes she is inducing inaccurate beliefs in the receiver, which in turn has negative payoff consequences for the receiver. An important question therefore is whether the effects we observe are due to deception aversion per se, or if they are fully driven by *social* or *other-regarding* preferences over final outcomes.[33]

To answer this question, we first generalize utility function (1) for the deception games (message-payoff and belief-payoff) to explicitly allow for social preferences:

$$\tilde{\phi}^g(m, \theta, d; \beta) = \tilde{u}(p_S^g(m; \beta), p_R^g(m, \theta; \beta)) - c^{lie}(m, d) - c^{dec}(\beta(m), \theta).$$

We have simply replaced material utility $u^g(m; \beta)$ with $\tilde{u}(p_S^g(m; \beta), p_R^g(m, \theta; \beta))$, where $p_S^g(m; \beta)$ and $p_R^g(m, \theta; \beta)$ are the sender-expected material payoffs in game $g \in \{M, B\}$ for the sender and receiver, respectively.[34] We impose no restrictions on material utility $\tilde{u}$, and therefore allow for *arbitrary social preferences*.[35] Noting that θ enters $\tilde{u}$ directly because it is payoff-relevant to the receiver, social preferences can give rise to coin effects

[33] It is important to emphasize, however, that the final outcomes the sender expects to induce depend on her second-order beliefs, so even a sender with social preferences, but not deception aversion, must be guided by her second-order beliefs.

[34] $p_S^M(m) = 0.32 + 0.04m$, $p_S^B(m; \beta) = \beta(m)$, and, for both $g \in \{M, B\}$: $p_R^g(m, 0; \beta) = 1 - \beta(m)^2$ and $p_R^g(m, 1; \beta) = 1 - (1 - \beta(m))^2$. The expressions for $p_R^g(m, \theta; \beta)$ are based on the binarized scoring rule.

[35] To fix ideas, there may be altruism, inequality aversion, or even spite, as in Fehr and Schmidt [1999]. The altruistic case in which the sender's utility is increasing in both sender and receiver payoffs is perhaps the most interesting, as it can most plausibly generate coin effects without deception costs.

even without deception costs.

Our approach is to compare behavior in the deception games to a suitable *control game* in which messages lead to the same sender-receiver allocations and incur the same lying costs, but there is no possibility to deceive—and therefore no deception costs. In other words, the control game should induce a utility function

$$\tilde{\phi}^0(m, \theta, d) = \tilde{u}(p_S^0(m), p_R^0(m, \theta)) - c^{0,lie}(m, d),$$

where $(p_S^g(m;\beta), p_R^g(m, \theta; \beta)) \approx (p_S^0(m), p_R^0(m, \theta))$ for all (m, θ) and $c^{0,lie}(m, d) \approx c^{lie}(m, d)$ for all (m, d). In such case, the utility difference $\tilde{\phi}^0(m, \theta, d) - \tilde{\phi}^g(m, \theta, d; \beta)$ approximately equals deception costs $c^{dec}(\beta(m), \theta)$, and therefore differences in behavior between deception and control games would be fully attributable to deception costs. We note that, because material utility $\tilde{u}$ drops out, this claim allows for arbitrary social preferences.

We argue that, *conditioning on tails* ($\theta = 0$), the pure lying game makes a good control for the deception games. Table 5 shows the sender-expected material payoffs to both sender and receiver as a function of the message m, given second-order beliefs β and that the coin is tails. Note that the receiver's payoff of $1 - \beta(m)^2$ in the deception games is based on the binarized scoring rule (BSR). If β is strictly increasing, all games feature a similar qualitative tradeoff between sender and receiver material payoffs, with lower messages being more generous to the receiver at the expense of the sender. The quantitative tradeoffs are also plausibly similar, though this is sensitive to β. In the case that $\beta(m) = 0.32 + 0.04m$ (the truthful/Bayesian benchmark), sender payoffs are identical across all games. Moreover, the marginal changes in receiver payoffs as a function of the message are also similar across games: $(1-\beta(m)^2)-(1-\beta(m+1)^2) \approx 0.04$ for $m \in \{1, 2, ..., 7\}$.[36]

Game	Sender's payoff (p_S^g)	Receiver's payoff (p_R^g)
Message-payoff	$0.32 + 0.04m$	$1 - \beta(m)^2$
Belief-payoff	$\beta(m)$	$1 - \beta(m)^2$
Pure lying	$0.32 + 0.04m$	$0.68 - 0.04m$

Table 5: *Sender's expected material payoffs for both sender and receiver* conditional on tails.

[36]The exact values range from 0.030 to 0.050: 0.030, 0.034, 0.037, 0.040, 0.043, 0.046, and 0.050.

Under Assumption 1, following tails, lower messages in the deception games imply lower deception costs. Hence, within our framework and assuming the validity of the control, if the sender sends lower messages following tails in the deception games (relative to the control), this can only be explained by deception aversion. Intuitively, following a given coin-die realization $(\theta = 0, d)$, the sender's message in the pure lying game reveals her optimal sender-receiver allocation in the absence of deception costs. If the sender chooses lower messages in the deception games following the same coin-die realization, this can only mean that she is changing her behavior to avoid deception costs.

In Figure 10, we plot the average messages sent after each die outcome, conditional on tails. The left panel compares message-payoff to pure lying, and the middle panel compares belief-payoff to pure lying. We find that, compared to pure lying, average messages following tails are much lower (and statistically so) in the deception treatments, which can only be explained by deception aversion within our framework.

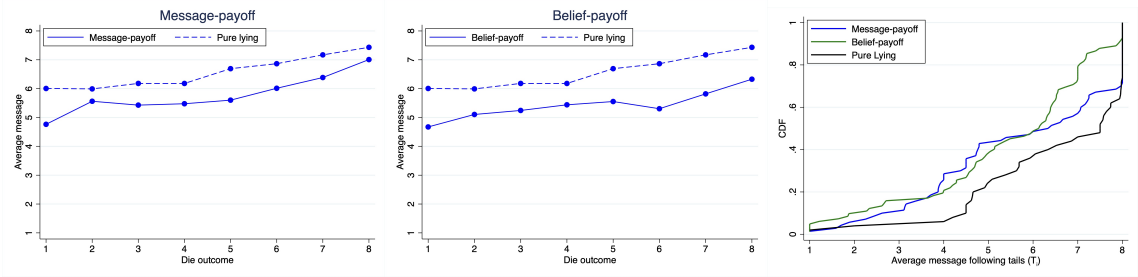


Figure 10: *Average messages conditional on tails—comparison to pure lying.*

To confirm that this is not driven by a small handful of subjects, we compute the average message following tails for individual i as $T_i = \frac{\sum_{d'} \mathbf{1}\{i \text{ observes } (0,d')\}(\bar{\sigma}_S^i(0,d'))}{|\{d|i \text{ observes } (0,d)\}|}$, where $\bar{\sigma}_S^i(0,d')$ is the average message conditional on $(0,d')$. In the right panel of Figure 10, we plot CDFs of all subjects' T_i for each treatment. We find that the distribution of T_i in pure lying stochastically dominates that in both deception treatments, confirming that, indeed, there is a strong tendency to message lower numbers following tails in the deception games, consistent with deception aversion.

Remark 7. An alternative control game would be an exact replication of one of the deception games, except that the BSR gives the receiver a chance at $\$X$ instead of \$10. By setting $X = 0$, one could turn off payoff consequences for the receiver and therefore social

preferences as well. However, this would be a poor control because the deception cost associated with any degree of deception, as measured by $|\beta(m)-\theta|$, may depend directly on X.[37] Hence, it is important that our control game approximates the same mapping from messages to sender-receiver allocations as in the deception games. In this way, our control is similar to that employed by Gneezy [2005], where he compares behavior in a lying game to a control dictator game with the same set of feasible allocations.

Remark 8. By explicitly allowing for arbitrary social preferences, the analysis of this section does not preclude (nor provide evidence for) social preferences. Some coin effects may be driven by social preferences, but we have shown that they cannot explain a significant feature of the data that can be explained by deception aversion.

Remark 9. One possible threat to the validity of our control is that the similarity of the sender-expected receiver payoffs across the deception and pure lying games is based on the sender's understanding of the fine details of the BSR, beyond just understanding that it is incentive-compatible. However, we conjecture that not fully grasping the fine details of the receiver's payoffs should, if anything, make the sender less willing to sacrifice her own (certain) payoffs to increase the receiver's (uncertain) payoffs.[38] This force would push an otherwise altruistic sender to choose higher numbers in the deception games, making our analysis even more conservative.

Finally, we also look for other indications of social preferences in our data. Recall that only in the belief-payoff game does the receiver have any influence over the sender's payoffs, where in fact the sender's payoff *equals* the receiver's reported belief. Therefore, receiver altruism should lead to higher belief reports in round 41 of the belief-payoff treatment. However, we find that (1) within the belief-payoff treatment, average beliefs are not higher in round 41 when the report effects the sender's payoff than in the questionnaire when belief elicitation is purely hypothetical; and (2) average beliefs in the belief-payoff treatment are no higher than in the other treatments (whether incentivized or not). Hence, social preferences seem to have little effect on the receivers' reported

[37] Our framework does not preclude that deception costs depend directly on X. We have not made such dependence on X explicit in the notation because we do not vary X. However, we may replace $c^{dec}(\beta(m),\theta)$ with $c^{dec}(\beta(m),\theta,X)$, and all results go through as stated.

[38] Consistent with this conjecture, Huang et al. [2023] show experimentally that the proposer's offer in an ultimatum game is less generous when there is incomplete information about the responder's valuation.

beliefs. We present this analysis in Online Appendix 9.4.

7 Conclusion

Conveying private information is central to almost every economic and social interaction. In many cases, there are opportunities both to *lie* by misreporting the truth and to *deceive* by inducing inaccurate beliefs about some payoff-relevant state. While a large and influential literature has documented lying aversion, there is very little existing evidence for aversion to deceiving others, which may be just as empirically relevant.

In part, the lack of studies documenting deception aversion is unsurprising as the distinction between lying and deception has only recently been formalized (Sobel [2020]). Perhaps the bigger issue, however, is that deception depends on second-order beliefs, a complex and fundamentally unobservable object. This makes studying deception difficult without making strong auxiliary assumptions or eliciting second-order beliefs directly. In this paper, we take a different approach, introducing a simple game and theoretical framework that allows us to identify deception aversion—and separate it from lying aversion—with minimal assumptions on second-order beliefs. We find strong evidence for deception aversion that cannot be explained by social preferences and is robust to varying strategic features of the game.

Our work is an early step toward a systematic study of deception aversion, and suggests many directions for future work. In particular, the "coin effect," which captures an individual's degree of deception aversion and can be correlated with other individual-level characteristics, may find broad application. For instance, it can be used to study the relationship between deception aversion and payoff consequences, paralleling one of the major themes within the lying aversion literature. This would help to refine the exact form of deception costs, informing theoretical applications. Another direction would be to document lying and deception aversion in diverse populations (e.g. children and adults, religious and secular, etc.). As lying and deception aversion may be culturally mediated, it would also be interesting to conduct cross-cultural studies. Finally, unlike existing approaches, we identify deception aversion from choice data alone—without an assumption of equilibrium or the need to elicit beliefs. This suggests that our results may be adapted to study deception "in the field" using observational data.

References

Johannes Abeler, Daniele Nosenzo, and Collin Raymond. Preferences for truth-telling. *Econometrica*, 2019. (document), 9, 1.3, 9.2

Despoina Alempaki, Valeria Burdea, and Daniel Read. Deceptive communication: Direct lies vs. ignorance, partial-truth and silence. *Journal of Political Economy: Microeconomics*, 2025. 12

Kai Barron and Tilman Fries. Narrative persuasion. *Working paper*, 2024. 40

Pierpaolo Battigali and Martin Dufwenberg. Guilt in games. *American Economic Review*, 2007. (document)

Pierpaolo Battigalli, Gary Charness, and Martin Dufwenberg. Deception: The role of guilt. *Journal of Economic Behavior and Organization*, 2013. (document)

Daniel Benjamin. Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations*, 2019. 9.1

Juan Francisco Blazquiz-Pulido, Luca Polonio, and Ennio Bilancini. Who's the deceiver? identifying deceptive intentions in communication. *Games and Economic Behavior*, 2024. 12

Gary Charness and Martin Dufwenberg. Promises and partnership. *Econometrica*, 2006. (document)

Syngjoo Choi, Chanjoo Lee, and Wooyoung Lim. The anatomy of honesty: Lying aversion vs. deception aversion. *Working paper*, 2025. (document), 14

Vincent Crawford and Joel Sobel. Strategic information transmission. *Econometrica*, 1982. (document)

Martin Dufwenberg and Martin A. Dufwenberg. Lies in disguise – a theoretical analysis of cheating. *Journal of Economic Theory*, 2018. 8

Rain Eilat and Zvika Neeman. Communication with endogenous deception costs. *Journal of Economic Theory*, 2023. (document), 1.2

Tore Ellingsen, Magnus Johannesson, Sigve Tjøtta, and Gaute Torsvik b. Testing guilt aversion. *Games and Economic Behavior*, 2010. (document)

David Ettinger and Philippe Jehiel. A theory of deception. *American Economic Journal: Microeconomics*, 2010. 15

David Ettinger and Philippe Jehiel. An experiment on deception, reputation and trust. *Experimental Economics*, 2021. 15

Agata Farina, Guillaume Frechette, Alessandro Ispano, Alessandro Lizzeri, and Jacopo Perego. The selective disclosure of evidence: An experiment. *Working paper*, 2024. 14

Ernst Fehr and Klaus Schmidt. A theory of fairness, competition and cooperation. *The Quarterly Journal of Economics*, 1999. 35

Urs Fischbacher and Franziska Follmi-Heusi. Lies in disguise–an experimental study on cheating. *Journal of the European Economic Association*, 2013. (document)

John Geanakoplos, David Pearce, and Ennio Stacchetti. Psychological games and sequential rationality. *Games and Economic Behavior*, 1989. (document)

Rajna Gibson, Carmen Tanner, and Alexander F. Wagner. Preferences for truthfulness: Heterogeneity among and within individuals. *American Economic Review*, 2013. 9, 18

Uri Gneezy. Deception: The role of consequences. *American Economic Review American Economic Review*, 2005. (document), 6

Uri Gneezy, Bettina Rockenbach, and Marta Serra-Garcia. Measuring lying aversion. *Journal of Economic Behavior and Organization*, 2013. 10

Uri Gneezy, Agne Kajackaite, and Joel Sobel. Lying aversion and the size of the lie. *American Economic Review*, 2018. (document), 1.2, 1.3, 9.2

Ben Greiner. Subject pool recruitment procedures: Organizing experiments with orsee. *Journal of the Economic Science Association*, 2015. 2

Elizabeth Hoffman, Kevin McCabe, Keith Shachat, and Vernon Smith. Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 1994. 2

Tanjim Hossain and Ryo Okui. The binarized scoring rule. *The Review of Economic Studies*, 2013. 2, 1.1, 2

Jennie Huang, Judd Kessler, and Muriel Niederle. Fairness has less impact when agents are less informed. *Experimental Economics*, 2023. 38

Sjaak Hurkens and Navin Kartik. Would i lie to you? on social preferences and lying aversion. *Experimental Economics*, 2009. 11

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 2011. (document)

Navin Kartik. Strategic communication with lying costs. *The Review of Economic Studies*, 2009. 13

Kiryl Khalmetski and Dirk Sliwka. Disguising lies - image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics*, 2019. 8

Kyungin Kim. Endogenous market segmentation for lemons. *The RAND Journal of Economics*, 2012. (document)

John C. Raven. Mental tests used in genetic studies: The performances of related individuals in tests mainly educative and mainly reproductive. *Unpublished master's thesis, University of London.*, 1936. 2

Joel Sobel. Lying and deception in games. *Journal of Political Economy*, 2020. (document), 1.3, 7

Matthias Sutter. Deception through telling the truth?! experimental evidence from individuals and teams. *Economic Journal*, 2009. (document)

Michael Thaler, Mattie Toma, and Victor Yaneng Wang. Numbers tell, words sell. *Working paper*, 2025. 40

Joseph Taoyi Wang, Michael Spezio, and Colin F. Camerer. Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. *American Economic Review*, 2010. 6, 3.1, 9.1

8 Appendix

8.1 Omitted proofs

Proof of Proposition 1. In the message-payoff game, messaging T is strictly dominant for any (θ, d), so it must be that $\sigma_S(T|\theta, d) = 1$ in any PBE. Hence, observing message T, the receiver will not update beliefs. In the belief-payoff game, suppose there is a message $m^{'}$ sent with positive probability in equilibrium after which the receiver forms beliefs $b^{'} \neq p_\Theta(1)$ and therefore reports $b = b^{'}$. Without loss, we may take $b^{'} > p_\Theta(1)$, implying the existence of another message $m^{''}$ sent with positive probability in equilibrium after which the receiver forms belief $b^{''} < p_\Theta(1)$ and therefore reports belief $b = b^{''}$. But this cannot be an equilibrium because it is strictly better for the sender to message $m^{'}$ than $m^{''}$. Hence, there can be no belief-updating in equilibrium, which only occurs if $(\theta = 0)$-senders and $(\theta = 1)$-senders use strategies that induce the same distribution over messages. □

Proof of Proposition 3. Suppose for purposes of contradiction that $c^{lie}(\sigma_S^*(\theta, d), d^{'}) \leq c^{lie}(\sigma_S^*(\theta, d), d)$ and $c^{lie}(\sigma_S^*(\theta, d^{'}), d) \leq c^{lie}(\sigma_S^*(\theta, d^{'}), d^{'})$. By definition of $\phi^g(\cdot)$, we have that $c^{lie}(\sigma_S^*(\theta, d), d^{'}) \leq c^{lie}(\sigma_S^*(\theta, d), d) \implies \phi^g(\sigma_S^*(\theta, d), \theta, d^{'}) \geq \phi^g(\sigma_S^*(\theta, d), \theta, d)$. By definition of $\sigma_S^*(\cdot)$, we have that $\phi^g(\sigma_S^*(\theta, d^{'}), \theta, d^{'}) > \phi^g(\sigma_S^*(\theta, d), \theta, d^{'})$ and thus that $\phi^g(\sigma_S^*(\theta, d^{'}), \theta, d^{'}) > \phi^g(\sigma_S^*(\theta, d), \theta, d)$. A symmetric argument yields $\phi^g(\sigma_S^*(\theta, d), \theta, d) > \phi^g(\sigma_S^*(\theta, d^{'}), \theta, d^{'})$, a contradiction. □

9 Online Appendix

9.1 A simple model of boundedly rational beliefs

The tendency for most subjects to form monotonically increasing beliefs and for some subjects to have hump-shaped beliefs that peak at 7 (as documented in Section 3) is broadly consistent with level k models in the spirit of Wang et al. [2010]. In such a model, level 0-senders are *truthful* and level 0-receivers are *credulous*—believing that every message was sent truthfully.

The level 1-receiver, believing she faces a truthful level 0-sender, is also credulous. Hence level 0- and level 1- receivers will form beliefs that are monotonically increasing in the message due to the positive correlation between Θ and D. The level 1-sender, best responding to a level 0-receiver, will always message 8. This trivially maximizes payoffs in the message-payoff and pure lying games, and is also the payoff-maximizing action in the belief-payoff game (due to the level 0-receiver's monotonically increasing beliefs). Now consider the level 2-receiver. Under the standard level k model, she believes she faces a level 1-sender who always message 8. However, this does not account for "off-path" messages that would be sent by other types. Hence, we suppose that the level 2-receiver believes she faces a large mass $(1-\epsilon)$ of level 1-senders and a small mass ϵ of truthful level 0-senders.

Applying Bayes' rule, this pins down the first-order beliefs b^k of the level k-receiver for $k \in \{0,1,2\}$. Assuming the level k-sender believes she faces a level $(k-1)$-receiver, this also pins down the second-order beliefs β^k of the level k-sender as $\beta^k = b^{k-1}$ for $k \in \{1,2,3\}$. We summarize these beliefs in Table 6.

Intuitively, when the level 2-receiver observes any message other than 8, she knows it came from a truthful level 0-sender, and so her beliefs are strictly increasing in messages 1 through 7 due to the positive correlation between Θ and D. Instead, when the level 2-receiver observes message 8, she knows it may have come from a level 1-sender who always messages 8, independent of the coin. A message of 8 may therefore be a relatively

weak signal that the coin was heads, depending on $(1-\epsilon)$—the fraction of level 1-senders she believes she faces. If $(1-\epsilon)$ is sufficiently high, then her beliefs will be hump-shaped with a peak at 7. Using the expressions from Table 6, the exact condition for hump-shaped beliefs is $b^{k=2}(8;\epsilon) < b^{k=2}(7;\epsilon) \iff (1-\epsilon) > \frac{1}{21}$. Similarly, because $\beta^k = b^{k-1}$, the level 3-sender's second-order beliefs will be hump-shaped if and only if $(1-\epsilon) > \frac{1}{21}$.

k	Sender's second-order beliefs $\beta^k(m;\epsilon)$	Receiver's first-order beliefs $b^k(m;\epsilon)$
0	–	$0.32 + 0.04m$
1	$0.32 + 0.04m$	$0.32 + 0.04m$
2	$0.32 + 0.04m$	$\frac{\epsilon(0.08+0.01m)+(1-\epsilon)\mathbf{1}\{m=8\}}{\epsilon 0.25+2(1-\epsilon)\mathbf{1}\{m=8\}}$
3	$\frac{\epsilon(0.08+0.01m)+(1-\epsilon)\mathbf{1}\{m=8\}}{\epsilon 0.25+2(1-\epsilon)\mathbf{1}\{m=8\}}$	...

Table 6: *Level k beliefs.*

We take the fact that most receiver-subjects have monotonic beliefs and that the second-largest group have hump-shaped beliefs as strong evidence for the type of credulity embodied in the level k model. Of course, no subjects conform to any of the level k-types perfectly. For instance, an exactly level 0- or level 1-receiver would form beliefs that correspond to the *truthful/Bayesian* benchmark (black line) in Figures 2 and 3. The only systematic deviation we observe, however, is consistent with the well-known empirical regularity of overreaction to signals (see, for example, Benjamin [2019]),[39] so we do not take this is as evidence against the credulity mechanism.

We conclude this section with two remarks.

Remark 10. The payoffs to the sender in the message-payoff game, i.e. $u^M(m) = 0.32 + 0.04m$, exactly coincide with those the level 1-sender believes she faces in the belief-payoff game, i.e. $u^B(\beta^{k=1}(m)) = \beta^{k=1}(m) = 0.32 + 0.04m$. Hence, if level k is indeed the true model generating beliefs, then the message-payoff and belief-payoff games are similar for subjects with low levels of sophistication.

Remark 11. In the message-payoff game, the receiver's beliefs do not matter for the sender's best response. As such, all senders of level $k \geq 1$ will send message 8 with

[39] Focusing on the average beliefs (blue) in Figure 2, relative to the truthful/Bayesian benchmark, we see that subjects tend to over-update in the "tails direction" following messages 1-4 and in the "heads direction" following messages 5-8.

probability 1. In the belief-payoff game, however, because the sender is incentivized to maximize the receiver's beliefs, the level 3-sender will message 7 when her second-order beliefs are hump-shaped. Indeed, senders of even higher level, if they believe they face a distribution of lower level receivers, may message lower numbers yet. Hence, the level k model captures the intuitive idea that a message of 8 is a "likely story" or "too good to be true," and so lower messages may be used by more sophisticated subjects.[40] This leads to the prediction that senders will send lower messages in the belief-payoff game.

9.2 Unexplained patterns in sender strategies

A small number of subjects exhibit die effects that cannot be captured qualitatively by the structural model, suggesting fixed or other non-linear lying costs. In Figure 11, we show a few examples. Subject MP10 tells the truth for die outcomes 5-8, and otherwise messages 8. This suggests a *fixed cost* associated with lying, such that it is optimal to tell the truth if it leads to good outcomes and tell a *maximal lie* otherwise. BP10 tells the truth for die outcomes 5-8, and messages $d+4$ for die outcomes $d \in \{1, ..., 4\}$. This is consistent with two tiers of fixed costs: one associated with any degree of lying, and another that applies only if the distance between the true die outcome and message exceeds 4. BP11 messages 6 for die outcomes 1-6, and tells the truth otherwise. This can be explained by incorporating image concerns—i.e. "reputation for honesty" cost in the spirit of Gneezy et al. [2018] and Abeler et al. [2019] (see also footnote 18). That is, if the subject thinks that messaging 7 and higher signals that she is lying with high probability, she may avoid such messages unless they are true. Alternatively, it may also be that the subject has an "aspiration level" and wishes to guarantee at least the payoff associated with 6.

There may also be some unexplained deception aversion. If the structural model perfectly captured subjects' utility functions and second-order beliefs, then we would expect (up to some randomness) no coin effects among subjects classified as non-deception-averse—those for which the best model involves $\kappa^{dec} = 0$. However, among such subjects, we identify three (all from belief-payoff) with C_i exceeding 1.20 (the next highest

[40] In other words, lower messages may constitute a more credible "narrative" that the coin is heads. This relates to the nascent experimental literature on narratives, e.g. Barron and Fries [2024] and Thaler et al. [2025].

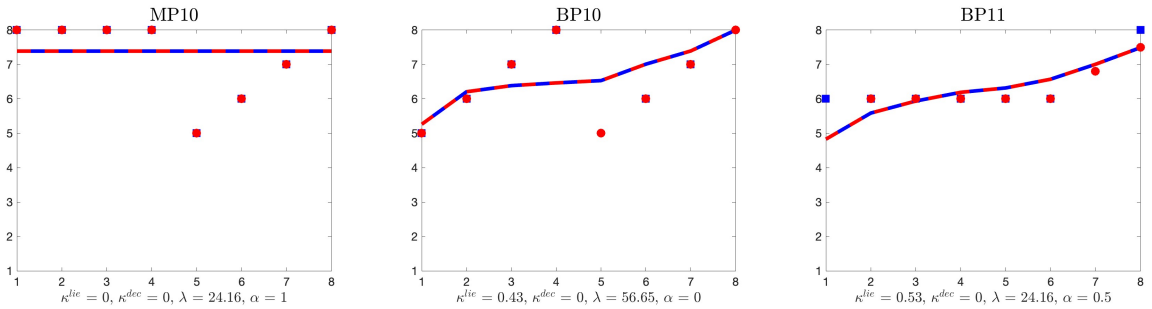


Figure 11: *Unexplained lying aversion—individual subjects.*

being 1.01). We give plots for these subjects in Figure 12. These subjects have clear positive coin effects, but the model provides a poor fit and does not identify them as deception-averse. In our view, these subjects most likely *are* deception-averse, but the model is not flexible enough to capture the exact form in which it manifests. These examples suggest that the structural estimates may understate the degree of deception aversion, further highlighting the value of the reduced-form analysis.[41]

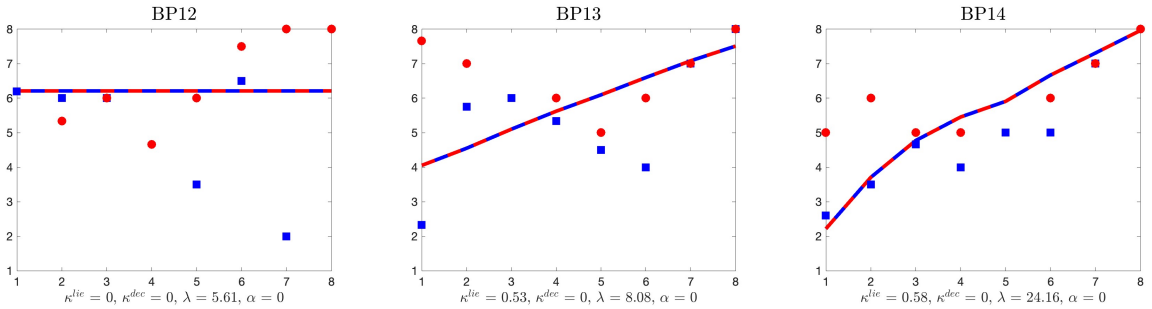


Figure 12: *Unexplained deception aversion—individual subjects.*

We observe some other patterns in a handful of individual subjects' data that we believe result from deliberate and intuitive strategies, but are still not well captured by the structural model. Overall, our sense is that more subjects in the belief-payoff treatment display such patterns, perhaps because the game's higher complexity leads to greater use of simple heuristics. In Figure 13, we give the data for three such subjects, all from the belief-payoff treatment.

[41] Note also that, because the structural model assumes monotonically increasing or hump-shaped second-order beliefs that peak at 7, it classifies the small number of subjects with negative coin effects as non-deception-averse. In principle, however, such subjects may be deception-loving or deception-averse with *decreasing* second-order beliefs. Of course, they may also be confused or careless.

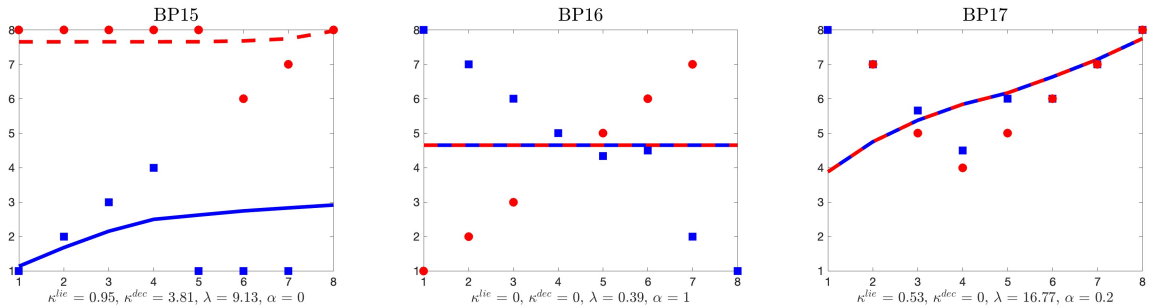


Figure 13: *Other unexplained patterns.*

Referring to Figure 13, subject BP15 is classified as deception-averse, but exhibits a pattern that cannot be explained by the structural model. Following heads, this subject tells the truth for die outcomes 6-8, but messages 8 otherwise; following tails, the subject tells the truth for die outcomes 1-4, but messages 1 otherwise. Her overall behavior suggests a heuristic of "tell the truth if deception is below some threshold; minimize deception otherwise." Because this subject exhibits both coin and die effects, we know that both lying and deception costs are needed to rationalize her behavior. This pattern can be explained by a combination of lying costs and a *fixed cost* from deceiving above some threshold. BP16 tells the truth following heads and tells the opposite of the truth following tails, i.e. $m = 9 - d$. Recalling that incentives are to convince the receiver that the coin was heads, this subject is "telling the truth in the good state; saying the opposite in the bad state." Hence, while misguided in this context, this may be an application of an intuitive heuristic, as opposed to confusion about the game per se. BP17 tells the opposite of the truth for die outcomes 1-4 and tells the truth for die outcomes 5-8. Hence, she applies a similar heuristic: "tell the truth if (she believes) it leads to good outcomes; say the opposite otherwise."

9.3 Robustness to alternative structural models

We consider three alternative structural models and conduct the same exercise from Section 5 to type individual subjects. The first two models are the same as in Section 5, except we use alternative specifications for the second-order belief function. Model (A1) uses the average beliefs of *all* receiver subjects in message-payoff and belief-payoff

treatments.[42] Model (A2) uses the "fictitious" second-order beliefs $\beta(m) = 0.32 + 0.04m$ that are implied by truthful/Bayesian benchmark (which coincides with the second-order beliefs of the level 1 or 2-sender from the model of Online Appendix 9.1). Finally, model (A3) is the same as the model from Section 5, except the lying and deception costs are assumed to be quadratic: $c^{lie}(m,d) = \kappa^{lie}((m-d)/7)^2$ and $c^{dec}(\theta, \beta(m)) = \kappa^{dec}(\beta(m) - \theta)^2$, respectively. Table 7, which is directly comparable to Table 4, summarizes the results. The alternative models imply similar distributions of types. In particular, the estimated fraction of individuals who are averse to deception varies only between 28 to 32% across all four models, and so we conclude the results are robust.

Model	Game	Averse to:			
		Lying and deception ($\kappa^{lie} > 0$, $\kappa^{dec} > 0$)	Deception only ($\kappa^{lie} = 0$, $\kappa^{dec} > 0$)	Lying only ($\kappa^{lie} > 0$, $\kappa^{dec} = 0$)	Neither ($\kappa^{lie} = \kappa^{dec} = 0$)
(A1)	Message-payoff	14% (10)	16% (11)	31% (22)	39% (27)
	Belief-payoff	15% (12)	12% (10)	48% (39)	26% (21)
	Pure lying	4% (2)	4% (2)	34% (17)	58% (29)
(A2)	Message-payoff	16% (11)	16% (11)	31% (22)	37% (26)
	Belief-payoff	13% (11)	13% (11)	40% (33)	33% (27)
	Pure lying	2% (1)	6% (3)	34% (17)	58% (29)
(A3)	Message-payoff	6% (4)	29% (20)	26% (18)	40% (28)
	Belief-payoff	9% (7)	21% (17)	27% (22)	44% (36)
	Pure lying	6% (3)	10% (5)	26% (13)	58% (29)

Table 7: *Summary of structural estimates—Robustness.*

[42] That is $\beta(1, 2, ..., 8) = (0.256, 0.285, 0.343, 0.416, 0.521, 0.610, 0.686, 0.700)$, which is strictly monotonic in messages.

9.4 Comparing beliefs across treatments: no evidence for receiver altruism

	Questionnaire			**Round 41**	**R41 vs. Question.**
	MP vs. BP	PL vs. BP	MP+PL vs. BP	MP vs. BP	BP
	(1)	(2)	(3)	(4)	(5)
Belief-payoff	2.942	1.770	2.524	3.397	
	(2.040)	(1.737)	(1.700)	(5.019)	
Round 41					-5.735*
					(3.242)
Observations	1368	1138	1768	152	820

Table 8: *Comparing beliefs across treatments.* In columns (1)-(3), we regress the belief reported in the questionnaire on indicators for the belief-payoff treatment, conditioning on the die outcome (the 8 die dummies are omitted); columns (1), (2), and (3) compare belief-payoff to message-payoff, pure lying, and message-payoff and pure lying pooled together, respectively. In column (4), we compare beliefs in round 41 between belief-payoff and message-payoff treatments by regressing the belief reported in round 41 on an indicator for belief-payoff (recall that beliefs are not elicited in round 41 of pure lying), conditioning on the die outcome. All magnitudes are small and insignificant: beliefs are no higher in the belief-payoff treatment when higher belief reports from the receiver increase the sender's payoff. In column (5), focusing on the belief-payoff treatment, we compare beliefs in round 41 (when beliefs are incentivized) to beliefs in the questionnaire (when they are not incentivized) by regressing beliefs on an indicator for round 41, conditioning on die outcome. We find that beliefs are 5.7 p.p. *lower* in round 41 compared to the questionnaire, indicating that reported beliefs are not higher when reporting higher beliefs increases the sender's payoff. Standard errors are clustered at the subject level.

9.5 Experimental instructions

We present the experimental instructions for the message-payoff treatment. Instructions for the other treatments were designed to be as similar as possible, and are available from the authors on request.

Experimental Instructions

The experiment will take approximately **50 minutes**, and you will receive **$10** just for completing the experiment. In addition, you may receive additional payments based on your choices and the choices of other participants.

Please turn off your cellphones now. The session will be fairly short, so please try to be attentive throughout. Please feel free to ask questions at any point during the instruction period or during the experiment itself.

The experiment will have **one main section**, followed by a short **questionnaire**.

The main section, which we will describe now, has **41 rounds** in total.

Rounds 1-40

In each of the first 40 rounds, you will privately observe the outcome of a **coin flip**, which will be either HEADS or TAILS. Both are equally likely (50-50 chance of HEADS or TAILS).

After observing the outcome of the coin, you will also privately observe the outcome of a **roll of an 8-sided die**. The die is *weighted*:

- If the coin was HEADS, you are *more likely to roll* ***higher*** *numbers.*
- If the coin was TAILS, you are *more likely to roll* ***lower*** *numbers.*
- If the coin was HEADS, the exact probabilities of die outcomes are given in the following table:

Die outcome:	1	2	3	4	5	6	7	8
Probability:	9%	10%	11%	12%	13%	14%	15%	16%

- If the coin was TAILS, the exact probabilities of die outcomes are given in the following table:

Die outcome:	1	2	3	4	5	6	7	8
Probability:	16%	15%	14%	13%	12%	11%	10%	9%

In each of the first 40 rounds, you will flip a new coin and roll a new die. The outcomes in each round will be completely independent of previous rounds' outcomes AND completely independent of other participants' outcomes. For the entire experiment, no participant (other than you) will ever observe the outcomes of your coin flips or your die rolls.

In each round, after seeing your coin flip and die roll, you will **send a message**. Your message will say:

"The outcome of my die roll is ____"

where you must fill the blank with any number 1, 2, 3, 4, 5, 6, 7, or 8.

No one will see any of your messages until *after* all 40 rounds. In round 41, **one round** from rounds 1-40 will be randomly selected (all equally likely), and the message you sent in that round will be shown to another randomly selected participant.

We will describe the exact details later, but the **payment** you receive for sending messages will depend *only* on your randomly selected message, and you will tend to *earn more money* by reporting higher die outcomes.

NOTE: You may report any number you wish, independent of the actual outcome of the die roll.

Round 41

In round 41, you will observe the message sent by a randomly selected participant (other than you) in a randomly selected previous round (1-40). After observing the message (1, 2, 3, 4, 5, 6, 7, or 8), you will **report your guess** for the probability the participant's coin was HEADS in the randomly selected round. That is, you will be asked:

A participant (other than you) and a round are randomly selected.

In this round, the participant sent the message:

"The outcome of my die roll is ____"

What is your guess for the probability this participant's coin was HEADS in this round?

You will then enter your numerical guess between 0 and 100 via a slider.

NOTE: If you had no information, your best guess would be 50% (since the coin gives an equal chance of HEADS and TAILS). However, the message sent by the participant may give you additional information, allowing you to make a better guess.

Payments

Payment for sending messages (Rounds 1-40). Recall that, in round 41, one of your messages will be randomly selected and shown to another participant (other than you).

Depending on your message that is randomly selected, *you will receive* \$10 with some probability. This probability depends *only* on your message. The exact probabilities are given in the following table:

Randomly selected message:	1	2	3	4	5	6	7	8
Your probability of receiving \$10:	36%	40%	44%	48%	52%	56%	60%	64%

For example,

- If your randomly selected message is 2, you will receive \$10 with probability 40%.
- If your randomly selected message is 6, you will receive \$10 with probability 56%.

Hence, you will tend to *earn more money* by reporting higher die outcomes.

Payment for making a guess (Round 41). Recall that, after observing the message sent by a randomly selected participant (other than you), you will report your guess for the probability that the participant's coin was HEADS.

It is in your best interest to simply tell us your best guess. To ensure this, we use the following **payment procedure**:

Your guess is a number X between 0 and 100 that gives the probability that you believe the participant's coin was HEADS.

(1) If the participant's coin was HEADS, then you receive \$10 with probability $\left[100 - \frac{(100-X)^2}{100}\right]\%$

(2) If the participant's coin was TAILS, then you receive \$10 with probability $\left[100 - \frac{X^2}{100}\right]\%$

This procedure may seem complicated, but don't worry! You do not need to understand the details. You just need to know that, if you believe that the probability of HEADS is $X\%$, you will maximize the probability of earning \$10 by reporting exactly $X\%$.

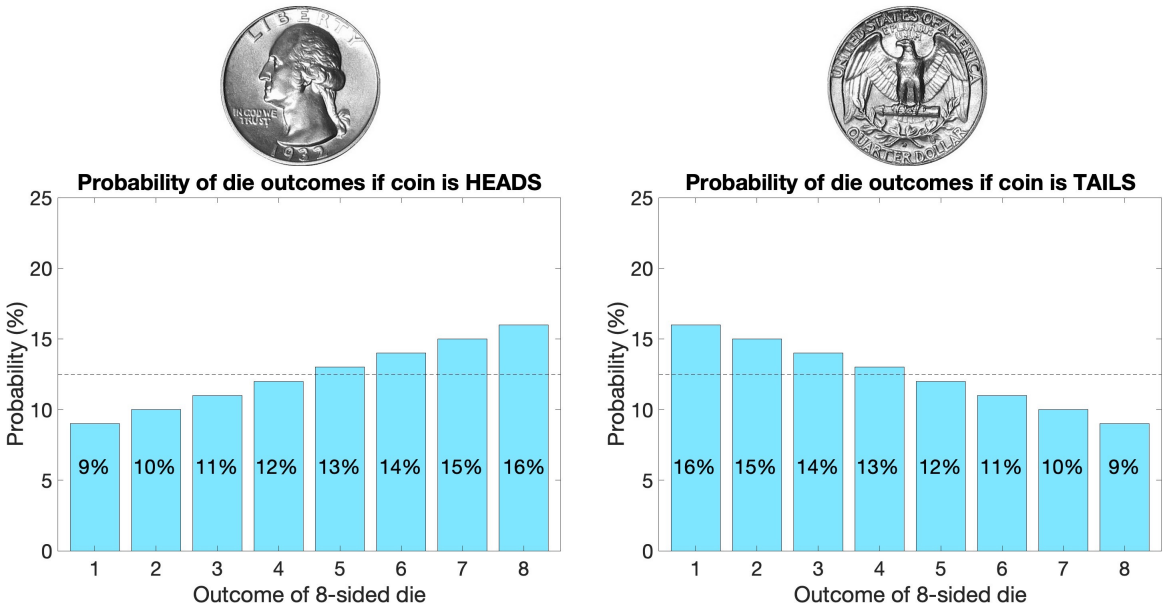
Hence, *whatever your best guess is, it is in your best interest to simply tell us.*

[Subjects are also provided with the following 1-page summary sheet]

Summary

Rounds 1-40. In each round, ...

- You will observe the outcomes of a **coin** and an **8-sided die**
  - If the coin is HEADS, you are more likely to roll higher numbers
  - If the coin is TAILS, you are more likely to roll lower numbers



- After observing the outcome of the coin and die, you will send a message reporting the outcome of the die, but you may send whatever message you like. *No participant (other than you) will ever observe the outcomes of your coin flips or your die rolls.*
- One of your 40 messages will be randomly selected and shown to another participant in round 41.

Round 41. You will observe the message sent by a randomly selected participant (other than you) and make a guess for the probability that the participant's coin was HEADS in the round when they sent that message.

Payment.

- Sending a message. By reporting higher die outcomes, you will increase the probability of receiving $10. The exact probabilities are given on Page 3 of the instructions.
- Making a guess. You will maximize your probability of earning $10 by simply reporting your exact best guess.

9.6 Experimental screenshots

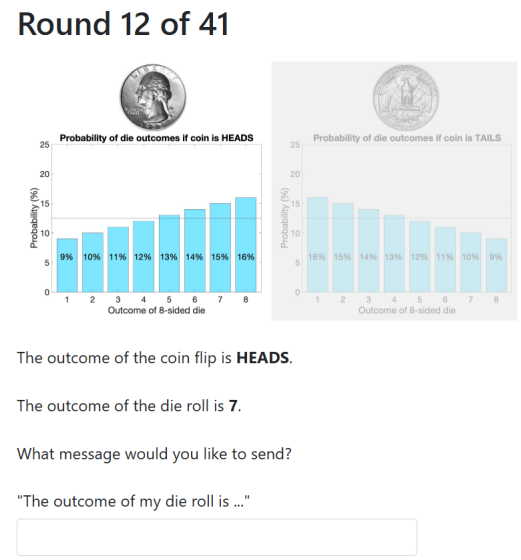


Figure 14: *Rounds 1-40 (heads).*

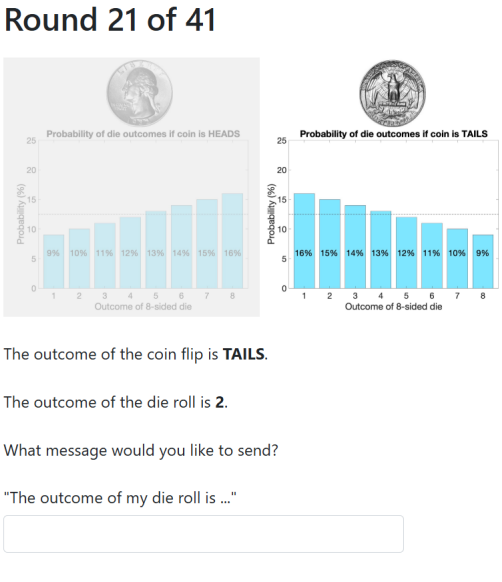


Figure 15: *Rounds 1-40 (tails).*

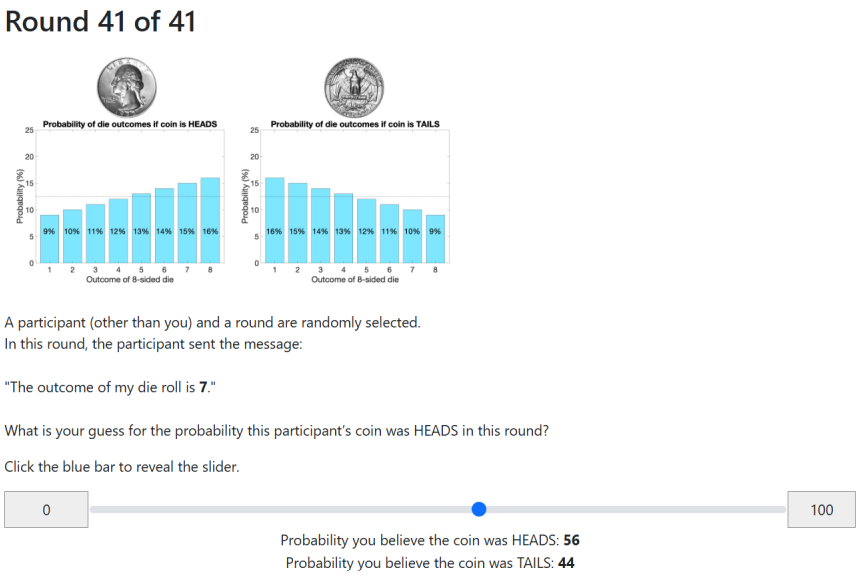


Figure 16: *Round 41 (message-payoff and belief-payoff).*

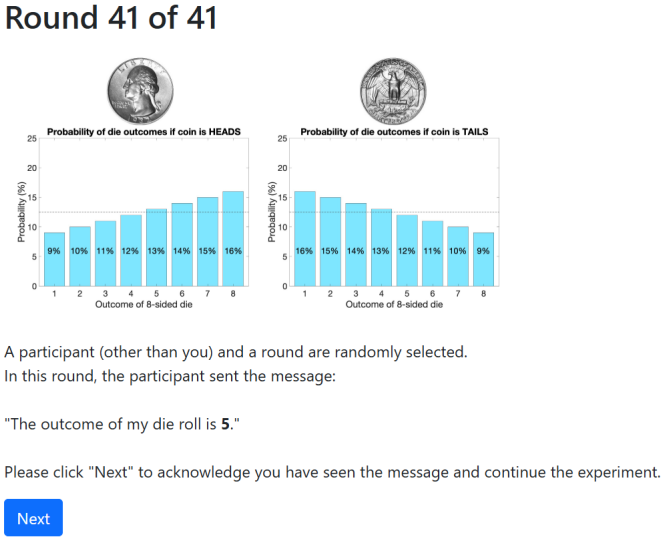


Figure 17: *Round 41 (pure lying).*