Games and Economic Behavior 145 (2024) 451–466

Contents lists available at ScienceDirect

Games and Economic Behavior

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

Who's the deceiver? Identifying deceptive intentions in communication

Juan Francisco Blazquiz-Pulido [a,b,*], Luca Polonio [c], Ennio Bilancini [a]

[a] IMT School for Advanced Studies Lucca, Lucca, Italy
[b] University of Alicante, Alicante, Spain
[c] University of Milano - Bicocca, Milan, Italy

ARTICLE INFO

JEL classification:
C91
D83
D84
D91

Keywords:
Deception
Lying
Strategic communication
Sender-receiver game
Trust

ABSTRACT

Recognizing people's deceptive intentions when communicating is crucial to detect statements that may drive us to unintended harmful decisions. This paper studies individuals' intentions in games where players can tell the truth with deceiving purposes. In a preregistered experiment, we combine a sender-receiver game with possible strategic considerations and the associated belief elicitation questionnaire, with a sender-receiver game with no room for strategic considerations. We propose a new method that improves the identification of senders' intentions to deceive. Our findings reveal that relying solely on the strategic sender-receiver game and the elicited beliefs, as previously proposed in the literature, can lead to misinterpreting the actual intentions of a substantial proportion of senders. In particular, our new method helps discern actual deceivers from pessimistic truth-tellers and identifies senders who try to excuse their previous deceiving message. All in all, our method identifies more senders with deceptive intentions compared to previous methods.

1. Introduction

In many daily-life situations, an individual can exploit information asymmetries to benefit from an uninformed counterpart, such as legal and financial advising, medical recommendations, or political elections. In such situations, an informed "sender" chooses whether or not to convey a truthful message to an unknowing "receiver", whose action would impact the payoff to both parties. Hence deception, conceptually defined as a deliberate act to induce a belief that the deceiver considers false in another person, is an important component in the decision-making process of humans when they communicate.

However, deception can happen not only by sending false information but also by telling the truth in the expectation of being mistrusted. An example of the latter behavior is *paltering*, a tactic commonly applied in situations such as negotiation, where the receiver is aware of the possibility of the sender's deceptive intentions. For instance, let us consider a seller of used cars who would like to persuade a potential client to buy a low-priced car for which she has a large profit margin. The seller could actually tell the truth and say that the cheap car is a lemon, suggesting buying a more expensive car for which the seller's profit margin is lower. Hence, by telling this truth, the seller could actively try to deceive the buyer believing that the suggestion "do not buy a cheap car

* Corresponding author at: IMT School for Advanced Studies Lucca, Piazza S. Francesco, 19, 55100 Lucca, Italy.
E-mail address: jf.blazquizpulido@imtlucca.it (J.F. Blazquiz-Pulido).

https://doi.org/10.1016/j.geb.2024.02.006
Received 11 October 2022

Available online 20 March 2024

because it is a lemon" will be interpreted as a lie. Other real-life scenarios where truth-telling is commonly used deceptively under the expectation of being mistrusted are bluffs in poker or goalkeepers signaling the side of the net where they will dive in a penalty shoot-out.

From an ethical standpoint, one may say that deception is an act that should be avoided, as it has the potential to drive people to make decisions that lead to undesired outcomes. In this regard, it is fundamental to distinguish lying from deception: while lying is the act of claiming something that is not true, deception happens when the claim – irrespective of being true or false – generates in the receiver a wrong belief.

The literature on dishonesty has grown on the evidence that many people do not lie even in situations where it is fully private and beneficial to themselves (Gneezy, 2005). This behavior has been mostly modeled through a pure "lying aversion" component when communicating, that is, assuming that humans experience a disutility – incur a psychological cost – when telling a lie. Different explanations[1] for lying aversion have been investigated, ranging from intrinsic motivations such as guilt or preferences for doing "the right thing to do" to external factors such as social pressure or image concerns.

Importantly, this literature has focused on lying as the only means of deception, whereas truth-telling was typically considered free of deceptive intentions. An exception to this is Sutter (2009), who documented a potentially different reason for truth-telling in a laboratory experiment: some senders sent the true message in the expectation that the receiver would not follow their advice and then will choose the non-suggested option. Therefore, these apparently benevolent truth-tellers are actually expecting to mislead receivers by telling the truth. This behavior was denoted as *sophisticated truth-telling* or, alternatively, *sophisticated deception*.

In order to detect such behavior, Sutter (2009) included a belief elicitation questionnaire right after the sender had taken her decision in the sender-receiver game, asking senders for the option they expected the receiver to choose and the probability of their message being followed. By combining the message sent in the game with the option expected to be chosen by the matched receiver, he found that around 30% of the senders aimed to deceive by telling the truth. This result proved that just observing the message emitted is not enough to know the sender's deceptive intentions. This finding has been considered in posterior studies just to rule out[2] the possibility that senders use truth-telling as a deceptive method, either by increasing the number of available messages (Erat and Gneezy, 2012) or avoiding it by design (Fischbacher and Föllmi-Heusi, 2013).

This paper aims to clearly identify individuals' deceptive intentions when communicating in sender-receiver games. To this aim, properly capturing senders' thoughts while communicating is crucial. However, belief elicitation in this game could be problematic with the sole use of questions about the expected behavior of receivers since the sender may have incentives to hide the deceptive purposes of her message. A common assumption in the lying literature is that individuals have heterogeneous truth-telling preferences. Since the sender-receiver game has a strategic component, the action obtained (i.e., the message sent) depends on the sender's belief about the option chosen by the matched receiver. However, not only do the actions depend on the senders' deceptive intentions, but the elicited beliefs are also influenced by them. Then, although senders' decisions are driven by their beliefs about receivers' actions, solely employing a belief elicitation question right after the game (such as asking about the expected option chosen by the matched receiver) may pose challenges in identifying their deceptive intentions. This is because senders might provide heterogeneous answers to this belief question based on their aim to deceive.

Therefore, a subject with a deceitful intention in the sender-receiver game may have incentives to hide her selfish purposes from the experimenter in the belief elicitation task. This might lead senders to lie when asked about their beliefs in the game, thereby indicating an expectation about the receiver's action that differs from the actual motive behind their own action. On the one hand, this problem of capturing senders' intentions exacerbates with the difference in the incentives between the game and the belief elicitation task, reaching its maximum when beliefs are not incentivized. On the other hand, incentivizing the belief questions does not necessarily solve the problem since some senders may obtain a psychological benefit for keeping their deceptive intentions hidden from the experimenter that is larger than the provided incentive, e.g., if they strongly dislike being considered liars. Note that this problem might appear in any type of communication game where strategic concerns could play a role in the information transmission. To the best of our knowledge, this potential issue has not been previously discussed or considered within the lying literature.

In this paper, we propose a novel method to identify the sender's deceptive intentions for the whole spectrum of senders who aim to deceive in sender-receiver games with a strategic component, addressing the issues inherent in the previous approach. Specifically, we conducted a two-stage experiment where participants maintained their assigned roles throughout the study. Participants played the classical strategic sender-receiver game, followed by a belief elicitation task about their expectations of the counterpart's behavior. Additionally, they conducted a non-strategic version of the same game where misleading intentions were not allowed. This non-strategic version is equivalent to the classical sender-receiver game with the only difference being that the sender's message directly determines the option chosen for the payments. So in this game, the receiver assumes a passive role, merely receiving the message and observing the final selected option. By combining the messages from the two games, we were able to identify the selfish purposes behind the senders' messages in the strategic sender-receiver game. Then, by using the elicited belief about the expected option chosen by the matched receiver in the strategic game, we accounted for the different intentions that senders held during communication.

[1] For the sake of exposition, we present further relevant literature on the drivers investigated and how lying aversion is related to different situational or demographic factors in Section 1.1.

[2] A review of the different methods developed to avoid sophisticated deception is provided in Section 1.1.

To identify senders aiming to deceive in the strategic sender-receiver game and to differentiate them from those who do not, we present evidence of several novel behaviors that a sender might exhibit in such a game, which have not been previously assessed. First, some senders truthfully communicate due to a preference to do so (for example, they might be motivated by moral reasons), although they hold a pessimistic belief about people's trustworthiness. These individuals would be erroneously identified as deceivers if their intentions were solely determined by observing their actions in the strategic game and their beliefs about the expected receiver's choice. Specifically, we found that around 40% of the senders perceived as sophisticated truth-tellers solely using the belief question to discern their deceptive intentions are identified as pessimistic using our new approach, which is grounded in actions and further refined by the belief question. Second, senders could potentially provide false information during the game and subsequently offer a (false) justification to rationalize their previous self-interested action in response to the belief elicitation question. These subjects would appear to be benevolent liars (who aim to help the receiver with their message and lack deceptive intentions) when considering solely their actions (sending a false message) and their beliefs regarding the option chosen by the receiver in the strategic sender-receiver game (expressing an expected lack of trust since the receiver will choose the non-suggested option). In contrast, their intention in the game is actually deceptive. We found that around 70% of the senders who would be considered benevolent liars using the previous method in the literature are liars who provided an excuse in these questions to hide their deceiving intention. Third, there might be senders who sophisticatedly tell the truth to the receiver, expecting not to be trusted, and subsequently hide their deceptive intention once asked about their expectations of the receiver's choice. These subjects would appear to be honest truth-tellers if using the previous method in the literature, whereas their intention in the game was to deceive. We found that around 25% of the senders who would be considered honest using the previous method are actually suspicious of aiming to deceive through sophisticated truth-telling.

Thanks to the identification of these new behavioral types, which were not recognized previously, we can more precisely distinguish between senders who intended to deceive with the message sent in the strategic sender-receiver game and those who did not have the intention to deceive the receiver. Overall, our new method identified that around 62.5% of the senders in our sample aimed to deceive in the strategic sender-receiver game, while only 49.1% of senders would be identified as deceivers using the previous method in the literature.

Although the literature on sender-receiver games is mostly focused on the sender's side,[3] the strategic component of this game makes the receiver's action and beliefs relevant to the senders' decision. Then, exploring beyond the receiver's decision and thoughts can provide insights into the senders' intentions. This is why we study the actions and beliefs of the receivers regarding the messages they get. In particular, we assess trustworthiness by examining how receivers act based on the suggestion provided by the sender. Notably, we observe a significant number of receivers who place trust in the message, regardless of the suggested option.

We also analyze receivers' beliefs about three important aspects of their decision in the game: the truthfulness of the message given, their belief in senders' expectations about their behavior, and whether they considered the message received when deciding their action. Besides, we elicited their belief about the truthfulness of the message also in the game where they are passive. We obtained that their belief about the message's truthfulness is independent of the receiver's role, that is, whether she takes an action or not in the game. Besides, we found that the receivers' expectation regarding the truthfulness of the received message differs from their belief about the sender's expectations of the receivers' trust. Finally, we found that only around 9% of receivers said that they ignored the message in choosing an option. These receivers chose the non-suggested option significantly more than the receivers who considered the message.

The rest of this paper is organized as follows: Subsection 1.1 reviews the related literature. Section 2 introduces the whole experiment, detailing the different tasks that form it, providing several pre-registered hypotheses, and displaying the details of the design and procedure used to identify senders' deceptive intentions. The experimental results of senders' actions in both games, their elicited beliefs, the behavioral types categorized and the number of senders with a deceptive intention identified using the previous and our new method, together with the results about receivers' trust and their beliefs are presented in Section 3. Section 4 concludes.

1.1. Related literature

Lying aversion has been extensively investigated in the past both theoretically and empirically. The main models in the literature attribute this behavior to a pure cost of lying (Gneezy et al., 2018), guilt aversion (Charness and Dufwenberg, 2006), image concerns (Khalmetski and Sliwka, 2019), or a combination of them (Kajackaite and Gneezy, 2017). More recently and from a broader perspective, moral preferences (i.e., a preference for doing what is considered "the right thing to do") have emerged as an explanation for the decisions made by individuals in different interactions (see Capraro et al. (2022) for a review on this novel topic), including dishonest behavior. An example is the finding by Thielmann et al. (2020) that the honesty-humility personality trait is related to several measures of pro-sociality, suggesting that honesty and different forms of pro-sociality share a common motivation.

Several studies in the dishonesty literature have shown that lying aversion appears to be sensitive to some situational factors, such as the monetary consequences of the lie for both the liar and the affected party (Gneezy, 2005; Hurkens and Kartik, 2009) or the time given to make the decision (Capraro, 2017; Lohse et al., 2018; Capraro et al., 2019). Besides, lying aversion seems to be related to demographic characteristics such as gender (Erat and Gneezy, 2012; Capraro, 2018), age (Glätzle-Rützler and Lergetporer, 2015), university studies (López-Pérez and Spiegelman, 2019), and occupation (Besancenot and Vranceanu, 2020).

[3] As far as we know, the only contribution that looks into the receiver's trust in sender-receiver games is Gylfason and Olafsdottir (2017). Trustworthiness has been analyzed in other types of settings, such as the distrust game McEvily et al. (2012) or the investment game Berg et al. (1995), which are less simple to administer than the sender-receiver game and then may trigger different reactions.

In the literature on dishonesty, the methods used to study how people aim to deceive have evolved over the years. After discovering that some senders tell the truth with deceptive intentions, many studies in the lying literature have explored different alternatives to rule out potential "strategic beliefs", thus avoiding the possibility of sophisticated truth-telling. With these changes in the design, they aim to rule out other forms of deception to focus on understanding the factors driving dishonesty. One approach proposed by Erat and Gneezy (2012) is expanding the message space of the game (from two to usually six possible messages). This method reduces the expected frequency of receivers following the message that a selfish sender would typically use, making sophisticated truth-telling a more preferred deceptive strategy than straightforward lying. However, Vanberg (2017) showed that in this different version of the game, there are participants who still aim to deceive by telling the truth. A different approach to avoid this behavior is to directly rule out any possible strategic concerns in the game by making the receiver passive (Biziou-Van-Pol et al., 2015; Capraro, 2017; Capraro et al., 2019), although this required a slight modification of the game framing. An alternative direction that a branch of the literature has taken is to change the framework completely, moving to the so-called cheating games (Fischbacher and Föllmi-Heusi, 2013; Abeler et al., 2014; Kajackaite and Gneezy, 2017). However, this variation implies losing the ability to identify dishonest behavior at the individual level, and it usually changes the victim of the lie, with the experimenter becoming a party to the interaction.

Furthermore, from a theoretical perspective, many contributions that model dishonesty explicitly rely on the psychological costs of lying (see Abeler et al., 2019; Gneezy et al., 2018; Khalmetski and Sliwka, 2019, and the references therein), considering truth-telling as an action free of deceptive intentions. An exception is Sobel (2020), who developed a model to characterize lying and deception in strategic communication settings, differentiating the two concepts. He showed that neither lying needs to be deceptive nor does deception require lying, formally characterizing the differences between both terms. Due to these differences, analyzing just the drivers of lying is not enough to know all the reasons that lead people to try to deceive in sender-receiver games. Then, the above-mentioned variations that the literature followed to avoid sophisticated deception cannot determine the whole spectrum of behaviors that can lead to deception in a sender-receiver game with a strategic component.

2. Experimental design and procedure

2.1. The experiment

We conducted a preregistered two-stage experiment: Before starting, we randomly assigned to each participant one single role (sender or receiver), which is preserved along the whole experiment. In the first stage, subjects started by playing the conventional sender-receiver game (also called deception game) created by Gneezy (2005), where strategic considerations can play a role (henceforth, the strategic game). Immediately after, senders answered two belief elicitation questions about the receiver's choice designed by Sutter (2009). In the second stage, subjects played a novel sender-receiver game in which no room for strategic considerations is allowed (henceforth, the non-strategic game).

Regarding the participants in the role of the receiver, in addition to their role in the two games, they answered a set of questions right after each of them. In particular, we elicited receivers' beliefs about the truthfulness of the message received in the two games and the senders' expected trustworthiness in the message received in the strategic one. Besides, we asked receivers whether they took into account the message obtained to choose an option in the strategic game. The instructions provided in the introduction of the experiment did not anticipate that more than one sender-receiver game would be played.[4] This was done to prevent undesired effects on senders' behavior and beliefs that may arise from forward-thinking senders considering the fact that they are playing two games of information transmission.

Notice that some studies have already used a sender-receiver game ruling out any strategic component by allowing senders to choose payments with their messages (Gneezy et al., 2013; Biziou-Van-Pol et al., 2015). However, to our knowledge, this is the first study with individual data of behavior in both a strategic and non-strategic sender-receiver game, preserving the same setting in both games.

2.1.1. The strategic sender-receiver game

The strategic sender-receiver game is a cheap talk game involving two types of players, the sender[5] and the receiver, each playing the game once. In this game, two options, labeled A and B, lead to conflicting monetary consequences for both players. The payments associated with each option are known just by the sender, who has to send to the receiver one of the two following predefined messages:

- Message 1: *"Option A will earn you more money than Option B."*
- Message 2: *"Option B will earn you more money than Option A."*

[4] Instructions stated that: "You will face a sequence of tasks that can involve another randomly picked participant." Section H in the Online Appendix contains the full set of instructions used in this experiment.

[5] To avoid any framing effect and have the best replication as possible of the main related literature, we named the sender as Person 1 and the receiver as Person 2. We keep using the sender and receiver labels throughout the text for the sake of clarity.

Once she gets the message, the receiver, who is fully unaware[6] of any information about the game's payoff structure[7] apart from the message received, has to choose which option is implemented for both players' earnings.

Moreover, right after emitting the message to the receivers, senders undergo a subsequent belief elicitation task. This task involves answering the same two belief questions created by Sutter (2009), asking about the option they expect the matched receiver to choose (question S_1) and the fraction of receivers they believe followed the sender's message (question S_2). The exact wording of the two questions is:

- Question S_1: *"Which option do you expect the receiver to choose?"*
- Question S_2: *"Out of 100 receivers, how many do you think follow the sender's message on average?"*

To have an experimental design as similar as possible to that of other contributions in the literature, we did not monetarily incentivize the answer to these belief elicitation questions. However, note that monetarily incentivizing the answers would not solve the problem of properly detecting senders' intentions, as explained in Section 1.

The standard approach in the literature to detect deceptive intentions in this game was to combine the sender's message with the answer provided to question S_1. In our new identification method, we adhere to this approach regarding how to use the answers to the belief questions. Specifically, we used the answers to the belief question S_1 in conjunction with additional behavior (as explained in Section 2.1.2) to identify deceptive intentions. As in the standard approach in the literature, the answers to S_2 are reserved for secondary analyses that investigate how the uncertainty in the expected trust affects the message sent in the strategic game.

Unlike most of the literature about sender-receiver games, we elicited receivers' beliefs after making their decision to know the reasons behind trustworthiness. We asked receivers two questions related to their beliefs about the message obtained: the first (question R_1) refers to the truthfulness of the sender's message, and the second (question R_2) to their belief about the sender's expectations of receivers following the message obtained, i.e., the second-order belief regarding the behavior answered by the senders to question S_1. By comparing the two questions, we can provide a measure of how suspicious receivers can be about a possible sophisticated truth emitted by the senders. The last question (question R_3) aims to know whether receivers considered the information provided by the sender's message to make their decision.

As in the senders' beliefs case, receivers' questions were not monetarily incentivized. The exact wording of the three questions is:

- Question R_1: *"Out of 100 senders, how many do you think sent a true message?"*
- Question R_2: *"Out of 100 senders, how many do you think expect the receiver to follow the message sent?"*
- Question R_3: *"Did you take into account the message received to choose an option?"*

2.1.2. The non-strategic sender-receiver game

The non-strategic game is a modified version of the strategic one where the receiver is passive, and the sender's message implements the option for payments. Except for this, the game remains as in the strategic version. This means that players' messages and the information provided are exactly as in the strategic game. The single difference is that, although the receiver observes the message emitted by the sender, she takes no action in the game. Notice that the sender was informed that her emitted message and the option selected for the payments in this game would be shown to the receiver.

Therefore, the strategic component that could drive senders' decision to tell the truth or lie is ruled out in this second game, remaining the factors related to other-regarding preferences and self-interested intentions that senders could have in the strategic game constant. Hence, with this non-strategic version of the game, we can capture senders' real intentions in the strategic game, ruling out any confound driven by strategic reasoning while controlling for any possible selfish or altruistic motive that could drive senders' communication. Note that we base our assumption that truth-telling preferences and selfish intentions do not vary across the two games of the experiment both on the full similarity between the two games and several features of our design explained in detail in Section 2.3.

Even though they cannot take any action to decide the option selected, receivers observe the sender's message in the non-strategic game. Hence, we can elicit their beliefs about the truthfulness of the message received to measure how the receiver's belief about the message's veracity depends on their role in the game. This question has the same wording as question R_1 in the strategic game, and no monetary incentives were provided. However, as receivers do not make a choice in this game, we cannot ask analogous questions to R_2 and R_3.

2.2. Identifying senders who aim to deceive

To identify whether a sender aims to deceive the matched receiver (either by telling the truth or lying) in the strategic sender-receiver game, it is necessary to know not only her expectations about the receiver's action, but also the purpose of the action taken. The combination of these three factors drives the behavior that the sender performs in the game.

When Sutter (2009) highlighted the importance of the interaction between the senders' decisions and their beliefs when communicating to receivers, he categorized the behavior of the senders in the classic sender-receiver game into four different behavioral

[6] Notice that the receiver's lack of information about the payoffs is also preserved after all decisions are taken. Receivers are only informed about their payoff once they receive their payment, preserving anonymity and hiding the veracity of the sender's message.

[7] The payoff distribution of each option in our experiment is reported in Table 2, along with other experimental procedure details presented in Section 2.3.

types: benevolent truth-tellers, who told the truth in the game expecting to be followed, plain liars, who lied expecting to be followed, sophisticated truth-tellers, who told the truth deceptively expecting not to be followed, and benevolent liars, who lied expecting not to be followed.

Although Sutter already recognized that other explanations could drive his categories, with our experimental design, we discovered that the previous method conflates senders with different deceptive intentions inside the same category. This implies that the number and the identity of senders considered deceivers in the strategic sender-receiver game could have been partially inaccurate.

To solve this problem, we combine the sender's action (that is, the message sent) in the two (strategic and non-strategic) sender-receiver games with the belief about the sender's expected option chosen by the matched receiver in the strategic game (acquired with the answer to question S_1). This approach offers a more precise method for identifying the intentions of senders in the strategic game, capturing behaviors driven by both deceptive and not deceptive intentions that could not be identified or measured using previous methods.

Table 1 schematizes our new method, providing a description of how we combine the sender's message in each of the two sender-receiver games with the sender's belief about her matched receiver's expected choice, ending up with eight possible behavioral types. In doing so, we improved the original method, which was solely based on beliefs to identify deceptive intentions in the strategic game, capturing four new behavioral categories.

Table 1
Characterization of senders' behavior using our new method.

Sender's behavioral type with our new method	Str game message	Belief S_1 answer	NS game message	Aim to deceive
Sophisticated truth-teller	Message 1	Option B	Message 2	Yes
Pessimistic truth-teller	Message 1	Option B	Message 1	No
Honest truth-teller	Message 1	Option A	Message 1	No
Incongruous truth-teller	Message 1	Option A	Message 2	Yes
Benevolent liar	Message 2	Option A	Message 1	No
Doubly liar	Message 2	Option A	Message 2	Yes
Plain liar	Message 2	Option B	Message 2	Yes
Incongruous liar	Message 2	Option B	Message 1	Yes

Notes: This table reports the behavioral typology of senders obtained with our new method. This method is based on the veracity of the message sent in the strategic and non-strategic sender-receiver game (message 1 is true whereas message 2 is false), and the expected option chosen by the matched receiver (i.e., the answer to the belief question S_1). The last column remarks the sender's deceptive intentions for each category.

Our method employs the action in the non-strategic game to account for the monetary-driven preferences that influence the sender's behavior in the strategic game, ruling out any confounding effect arising from strategic considerations. The last column of Table 1 indicates the behavioral types that aimed to deceive the receiver in the strategic sender-receiver game.

Aligned with our primary objective, we define the concept of *deceptive intention*[8] in cheap-talk games, serving as the foundation for our new method to identify a sender who aims to deceive.

Definition 1 *(Deceptive intention)*. A sender has a deceptive intention when transmitting a message to a receiver if, in any case, the aim of her message is leading the receiver to get an inferior outcome.

Our new method to identify deceptive intentions allows us to distinguish more precisely those senders who aimed to deceive with the message sent from those who did not. The branching of the behavioral types obtained using the previous method to identify deceptive intentions into the behaviors categorized with our new method is presented in Fig. 1. The first case involves the denoted sophisticated truth-teller category in the literature, considered a behavior with deceptive intentions. The way to identify this behavior was based on the fact that senders could tell the truth under the belief that the counterpart would not trust their message, then selecting the non-suggested option. However, there are two types of senders with different motives that could lead to this behavior: on the one hand, the real sophisticated truth-tellers (we preserve the label *sophisticated truth-tellers* when referring to them) who aim to deceive the receiver by telling the truth. These senders would be caught in our non-strategic game when lying to get the monetary benefit. On the other hand, the second type could be senders primarily motivated by their preferences regarding their actions when communicating. Then, these senders would choose to tell the truth according to their truth-telling preferences. If these players hold a pessimistic belief regarding people's trustworthiness, then their belief about the receiver's trust drives their answer in the belief elicitation question but is not connected with the message they sent in the strategic game. In this case, the player's intention in the strategic game is not to deceive, which is corroborated by their true message in the non-strategic game. We categorize these senders as *pessimistic truth-tellers*. An example of this behavior includes senders who are motivated by moral principles and experience a huge disutility in case they commit a lie, but that also believe people do not place much trust in others.

[8] The difference between this concept and deception is that the latter focuses on the actual outcome of the interaction, whereas the former deals with the sender's purpose. The two concepts and their definitions have been debated in the literature (see Sobel (2020) for an extensive conceptual and theoretical analysis).

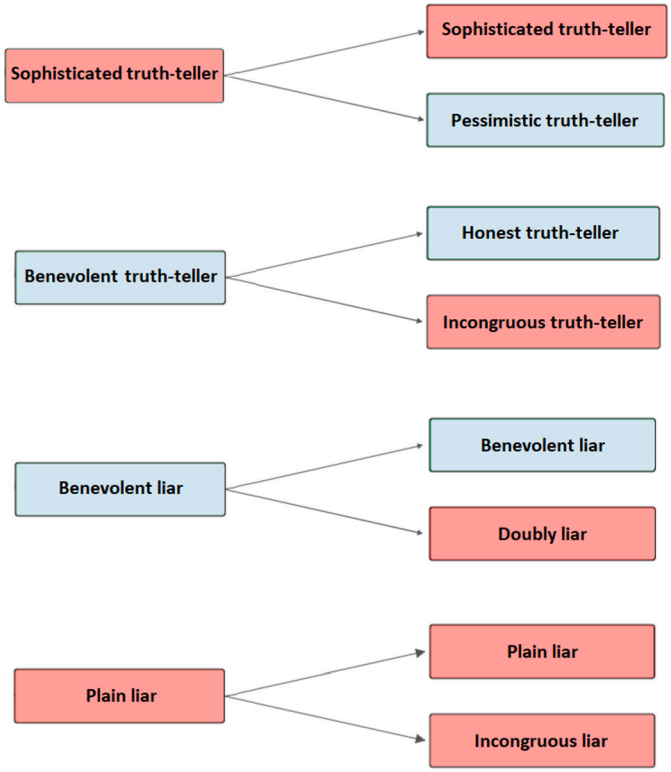


Fig. 1. Decomposition of the behavioral types categorized with the previous method that only uses a belief question to identify deceptive intentions in the strategic message (left nodes) into the behaviors categorized with our new method (right nodes) based on actions (both messages) and refined with beliefs. Behaviors in red (blue) are those who (do not) aim to deceive in the strategic game according to each identification method. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Furthermore, to preserve their deceptive intentions hidden, some senders may not express their real thoughts in the belief elicitation questionnaire. This could occur in cases of sophisticated truth-telling, where the sender's self-interested intention becomes even more challenging to identify, or in instances where a sender told a straightforward lie while expecting trust. These senders might subsequently provide false information in the belief questions to justify their prior deceptive behavior. Several reasons, such as social pressure or image concerns about the experimenter's opinion[9] could explain these two possible behaviors in the strategic game.

We denote as *incongruous truth-tellers* those senders who appear to be benevolent truth-tellers in the strategic game (since they tell the truth claiming to expect to be trusted) but lied when given the opportunity to determine the outcome of the game through their message (as in the non-strategic game). We hypothesize that the primary behavior underlying this category is that of sophisticated truth-tellers that aimed to conceal their deceptive intentions by providing false information when responding to the belief elicitation question. In the non-strategic game, they are forced to disclose their deceptive behavior to obtain the most beneficial option for themselves. If this selfish reason is the driver of their behavior, then the senders who fall in this category can be considered deceivers. The behavior of these individuals is therefore markedly different from that of *honest truth-tellers*, who genuinely convey the truth in both the strategic and non-strategic games without any intention to deceive the receiver.

An alternative explanation for this category is that senders could tell the truth in the strategic game driven by a high uncertainty about the receiver's choice since truth-telling preferences have little impact on their decision but are slightly preferred under uncertainty. Then, once they are in the non-strategic game, this uncertainty disappears, allowing them to deceive in order to gain a material benefit.

Our new method identifies individuals who tentatively provided false information in the belief questionnaire, providing an excuse for their observed spiteful behavior (claiming to expect not to be trusted). We denote them as *doubly liars*.[10] The deceptive intention

[9] Although we specify that anonymity both during and after the whole experiment was preserved, we acknowledge that we cannot reject the possibility that social image plays a role (Jordan and Rand, 2020). We clearly stated that the Prolific subjects' IDs (the only resource to track participants, although anonymously) would be used just for payments. Besides, running the experiment online in Prolific (whose regulation clearly explains that the experimenters can never identify the participants in the experiments) should be helpful in this concern. A novel result about cheating games in online settings goes in this direction, see Dickinson and McEvoy (2021).

[10] Despite we believe that the main reasoning behind this behavior is the one of lying in the questionnaire to screen a better image and hide their deceptive intentions, another possibility is that it is motivated by a reward gained if telling a lie, meaning that these senders are *lie lovers*. Note that both explanations align in terms of deceptive intentions, thus not affecting our identification process.

of these senders is captured in our non-strategic game, where they once again lie to attain a monetary benefit. Conversely, those *benevolent liars* who aimed not to deceive the receiver (they actually expected receiver's mistrust) consistently tell the truth in the non-strategic game.

Lastly, in our experiment, we can identify senders who deceive in the strategic game stating in the belief elicitation that they expected to be followed, and then subsequently, in the non-strategic game, tell the truth. Although this *incongruous liar* behavior seems odd, this combination could be driven by a change of behavior between the two tasks, meaning that these subjects regret their spiteful behavior in the previous condition. We admit that we cannot identify the drivers of this behavioral type. However, the results presented in Section 3 show that this behavior is uncommon, indicating that this category is of minor significance and did not pose an issue in our design.

To summarize, our two-game design allows us to recognize four new behavioral types according to their deceptive intentions that could not be detected just by using the data (action and belief answers) from the strategic game. We formulated several preregistered hypotheses[11] about the relevance of the new behavioral types categorized with our new method, emphasizing the importance of distinguishing them from the categories previously defined in the literature.

- **Hypothesis 1**: All the new behavioral types categorized with our new method are relevant, that is, none of the new categories is negligible.
- **Hypothesis 2**: The benevolent liar category identified with the previous method in the literature is mainly driven by doubly liars.
- **Hypothesis 3**: A significant number of subjects considered to be sophisticated truth-tellers using the previous method in the literature are pessimistic truth-tellers according to our new method to identify deceptive intentions.
- **Hypothesis 4**: The frequencies of the new behavioral types identified with our new method in our data cannot be fully driven by random behavior.

Crucially, evidence supporting Hypotheses 2 and 3, along with the potential presence of incongruous truth-tellers who are also deceivers, could alter the extent of observed deception in comparison to the results obtained using solely data from the strategic game (action and beliefs), as previously done in the literature.

2.3. Experimental procedure

To prevent compensation, reputation, or learning effects, each subject was randomly matched with a different participant for the second interaction. Also, we did not provide any feedback throughout the two tasks, not even the information regarding the first interaction outcome. Besides, anonymity was always preserved throughout the whole experiment, also to the experimenter.

To ensure incentive compatibility and avoid hedging and income effects, the payment of each subject was determined through the random selection of a single game outcome, either based on the payoffs of the option chosen in the strategic game by the receiver or the option selected in the non-strategic game through the sender's message. The belief elicitation task of the strategic game was not monetarily incentivized, as in Sutter (2009) and other contributions in this literature.

Table 2 presents the three between-subjects treatments we conducted, which differ in the incentives for the sender and the receiver (henceforth payment treatments). These are the ones used in the literature[12] (Gneezy, 2005; Sutter, 2009) to observe how sensitive the sender's decision is to both players' relative gains and losses. The two versions of the sender-receiver game (strategic and non-strategic) played by each subject had the same payoff scheme.

Table 2
Distribution of payoff and senders by treatment.

Treatment	Option	Sender payoffs	Receiver payoffs	Number of senders
Treatment 1 (T_1)	A	5	6	202
	B	6	5	
Treatment 2 (T_2)	A	5	15	180
	B	6	5	
Treatment 3 (T_3)	A	5	15	211
	B	15	5	

Notes: This table reports the distribution of payoffs (in points) and the total number of senders obtained in each payment treatment. Exchange rate: 10 points = 0.5 GBP.

[11] All these hypotheses and analyses, together with other important features of our experimental design such as the sample size and the recruitment criteria, were preregistered on Aspredicted.org. The link to the anonymized pre-registration file is https://aspredicted.org/N1N_KDP.

[12] We maintain the same payoff structure previously adopted in the literature by using experimental points. The exchange rate in this experiment was 10 points = 0.5 GBP.

Furthermore, the order of the two tasks (where a task includes the game and its subsequent belief elicitation questionnaire, if any) was randomized among subjects, resulting in two order conditions: some participants played first the strategic game followed by the subsequent belief elicitation, and then the non-strategic game (henceforth *normal order*), while others played first the non-strategic game and then the strategic game followed by its belief questionnaire (henceforth *inverted order*). Notice that, as explained in Subsection 2.1, receivers also answer a belief elicitation question right after the non-strategic game. Before starting the experiment, each participant was randomly assigned to a single role and payment treatment for both games, and an order condition. The diagram in Fig. 2 presents the procedure followed by senders who played in the normal order condition. For the procedure of senders in the inverted order and receivers, see Section F in the Online Appendix.

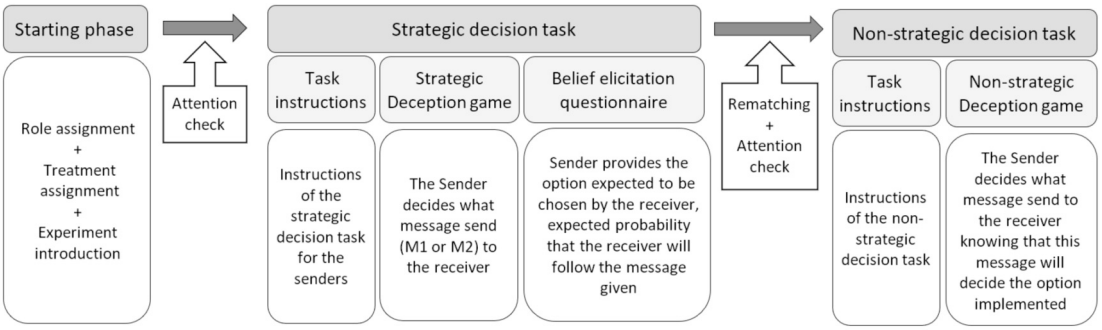


Fig. 2. Procedure of the senders who play in the normal order condition.

The experiment was conducted online, from mid-September to early October 2021, recruiting subjects through the experimental platform Prolific (Palan and Schitter, 2018). In total, we recruited 1.186 participants (593 assigned to the role of senders), all US nationals based in the US. 50.3% of the senders and 49.9% of receivers are females and the average age of the subjects is 30.12 (s.d. = 8.17) years old[13] for senders and 32.12 (s.d. = 8.42) for receivers.

As depicted in Fig. 2, an attention check was implemented immediately preceding the instructions of each decision task to ensure the quality of the data collected[14] and to exclude participants who did not pay attention. Participants obtained a show-up fee of 0.5 GBP for completing the experiment, and the payoffs of the games were paid through an extra bonus. On average, subjects were paid in total a number of points equal to 7.5 GBP per hour. The experiment took, on average, around 3 minutes to be completed.

3. Results

3.1. Behavior and beliefs in the strategic game

There are no statistically significant differences between the two order conditions regarding the sender's message transmitted (Pearson's Chi-squared test p-value = 0.22), the expected option chosen by the receiver (Pearson's Chi-squared test p-value = 0.62), and the expectation to be followed by the receiver[15] (Pearson's Chi-squared test p-value = 0.56). This lack of order effects is preserved in all our results. Then, also considering that the order of the two games was randomly assigned, we present the results aggregating order, and if not specified, the strategic game is referred to as the first game, whereas the non-strategic game is also called the second game. For the sake of exposition and to avoid repetitions, the robustness analyses controlling for the order condition are not provided in the main text but presented in brief in Section E of the Online Appendix.

Table 3 presents, for each payment treatment in our sample, the frequency of false messages sent in the strategic game and whether senders expected to be followed.

Table 3
Distribution of false messages and senders' expectations by treatment.

	T_1	T_2	T_3
Send the false message	102/202 (50.5%)	63/180 (35.0%)	127/211 (60.2%)
Expect to be followed	156/202 (77.2%)	144/180 (80.0%)	156/211 (73.9%)

Note: This table reports the frequency of false messages sent in the strategic task (Row 1) and the distribution of senders who expected to be followed (Row 2) by payment treatment.

[13] For more information about the sociodemographic characteristics of the subject pool, see Section G in the Online Appendix.

[14] Peer et al. (2017) reported the high quality of the data in Prolific relative to other alternatives.

[15] This value is determined by combining the message sent and the stated belief about the option that the sender expected the receiver to choose.

As observed in the first row, the payment treatment affects the message sent in the strategic game overall, i.e., T_1 vs. T_2 vs. T_3 (Pearson's Chi-squared test p-value < 0.001), and there is a statistically significant[16] difference between treatments T_1 and T_2 and between T_2 and T_3 (Pearson's Chi-squared test p-value < 0.001), but not between T_1 and T_3 (Pearson's Chi-squared test p-value $= 0.18$). However, as noticed in the second row, we find no difference among the three payment treatments regarding the expectations to be followed, neither altogether (Pearson's Chi-squared test p-value $= 0.36$) nor in pairwise comparisons (Pearson's Chi-squared test p-values are 1 for T_1 vs. T_2, 1 for T_1 vs. T_3, and 0.586 for T_2 vs. T_3).

The effect of payment treatment on the message sent combined with the lack of significant difference in expected trust between treatments T_2 and T_3 points to the fact that, keeping the cost of lying for the counterpart fixed, the higher the own material relative gain (with respect to the other possible option), the more incentives a sender has to tell a lie. An analogous argument explains the differences found between payment treatments T_1 and T_2, so keeping the own monetary relative gains fixed, the higher the relative loss for the counterpart (concerning the possible other option), the more difficult it is to tell a lie. A possible explanation for the lack of a significant difference between treatments T_1 and T_3 is that the positive and negative effects observed in the previous pairwise comparisons counterbalance each other due to their opposite directions. This occurs because, even though the sender's relative gains are increased, resulting in a rise in false messages, the receiver's relative losses also increase, causing a reduction in false messages.

3.2. Measuring who aims to deceive in the strategic game

To capture the senders' intentions when communicating in the strategic sender-receiver game, it is necessary to identify the sender's expected trust and how it interacts with the sender's action in the game. First, we provide the results obtained with the method used in the literature, which is based on the interaction between the message sent and the sender's expectation about the receiver's action. As explained in Section 2.2, this method relies on the fact that the answers to this question would accurately display the sender's beliefs when transmitting the message, independently of the sender's deceptive intentions. The frequencies of each behavioral type categorized using the previous method in the literature for each payment treatment in our sample are presented in Table 4.

Table 4
Distribution of behavioral types in the previous method by treatment.

Sender's behavioral type with the previous method	T_1 N = 202	T_2 N = 180	T_3 N = 211	Overall N = 593
Benevolent truth-teller	36.1%	55.6%	28.4%	39.3%
Sophisticated truth-teller	13.4%	9.4%	11.4%	11.5%
Plain Liar	41.1%	24.4%	45.5%	37.6%
Benevolent liar	9.4%	10.6%	14.7%	11.6%

Notes: This table reports the frequency of behavioral types categorized with the previous method in the literature (in rows) obtained in our online sample by treatment (in columns). The first row values (N) display the total number of senders per payment treatment.

We observe payment treatment effects consistent with the results presented in the previous subsection, indicating differences altogether (Fisher's exact test p-value < 0.001) and in all the pairwise comparisons but between T_1 and T_3 (Fisher's exact test p-values are lower than 0.001 for T_1 vs. T_2, and T_2 vs. T_3 and 0.50 for T_1 vs. T_3). Furthermore, we find no order effects in the results obtained using the previous method in the literature (Fisher's exact test p-value $= 0.558$). Additionally, we obtained results similar to those in Sutter (2009) regarding the message sent and the sender's expectations to be followed in the strategic game. However, we observed different frequencies for each behavioral type compared to the previous method in the literature. For the sake of exposition, we refer the reader to Section B in the Online Appendix for a more detailed comparison of the results obtained in our sample with those from previous contributions in the literature.

The previous results, together with the reasons explained in Subsection 2.2, highlight the importance of assessing whether this method correctly discerns senders' deceptive intentions and whether the previous behavioral types capture all the relevant rationale behind senders' communication in the strategic game. To reach this goal, we created a new method to identify deceptive intentions, revealing previously unmeasured behavior. The frequencies obtained for all behavioral types categorized with our new method, aggregating and controlling by payment treatments, are presented in Table 5.

We do not observe any order effects in the frequencies of the categorized behavioral types obtained with our new method, neither aggregating payment treatments (Fisher's exact test p-value $= 0.27$) nor controlling for them (Fisher's exact test p-values are 0.71 for T_1, 0.75 for T_2, and 0.39 for T_3). Besides, none of the frequencies of each category[17] are different across orders. This is why the frequencies in Table 5 and the following results are presented aggregating both order conditions. We also remark that all our results are similar when considering only the normal order condition.

[16] The results of all payment treatment comparisons are presented after correcting for multiple testing using Bonferroni's method. Notice that results are almost identical without correcting for multiple comparisons.

[17] The only behavioral type mainly driven by one of the two orders is the incongruous liar. In particular, 24/36 senders who fall in this category played in the inverted order.

Table 5
Distribution of behavioral types in our new method by treatment.

Sender's behavioral type with our new method	T_1	T_2	T_3	Overall
Sophisticated truth-teller	17 (8.4%)	7 (3.9%)	17 (8.1%)	41 (6.9%)
Pessimistic truth-teller	10 (5.0%)	10 (5.6%)	7 (3.3%)	27 (4.6%)
Honest truth-teller	57 (28.2%)	86 (47.8%)	34 (16.1%)	177 (29.8%)
Incongruous truth-teller	16 (7.9%)	14 (7.8%)	26 (12.3%)	56 (9.4%)
Benevolent liar	3 (1.5%)	5 (2.8%)	10 (4.7%)	18 (3.0%)
Doubly liar	16 (7.9%)	14 (7.8%)	21 (10.0%)	51 (8.6%)
Plain liar	68 (33.7%)	36 (20.0%)	83 (39.3%)	187 (31.5%)
Incongruous liar	15 (7.4%)	8 (4.4%)	13 (6.2%)	36 (6.1%)
Total	202 (34.1%)	180 (30.4%)	211 (35.6%)	593 (100%)

Notes: This table reports the frequency of behavioral types obtained with our new method, using the message in the strategic and non-strategic games and the stated expected option chosen by the receiver in the former.

Similar to the previous results regarding payment treatment effects, we found a significant effect both considering all treatments altogether (Fisher's exact test p-value lower than 0.001) and in all the pairwise comparisons but between T_1 and T_3 (Fisher's exact test p-values are 0.007 for T_1 vs. T_2, 0.12 for T_1 vs. T_3, and lower than 0.001 for T_2 vs. T_3).

The aggregated frequencies obtained show that a significant fraction of senders ($170/593 \simeq 28.7\%$) falls into one of the new behavioral types captured using our method. Besides, these new categories are relevant. To test Hypothesis 1 (none of the new behavioral types is negligible), we used both individual binomial tests and joint Fisher's exact tests (testing the corresponding hypothesis that they are a fraction below 1% of the data). We consider as the total number of subjects in our analysis either all senders (individual binomial test p-values lower than 0.001, joint Fisher's exact test p-value $= 0.007$) or those who fall in one of the new behavioral types identified with our method (individual binomial test p-values lower than 0.001, joint Fisher's exact tests p-value $= 0.015$). Besides, we observe significant differences in how the senders are assigned to a new behavioral type between the different categories from the previous method in the literature (joint Fisher's exact test p-value < 0.001).

As it can be observed by comparing the corresponding frequencies in Tables 4 and 5, the results obtained using our new method vary substantially for the behavioral types identified using the previous approach in the literature. In Fig. 3, we decompose[18] the number of senders in each behavioral type identified with the previous method in the literature on the number of senders in each behavioral type obtained with our new method.

Our second hypothesis is also supported by the data: the benevolent liar behavior from the previous method in the literature is mainly driven (51/69, 74%) by senders who probably lied in the belief elicitation questionnaire to excuse the false message sent, in line with Hypothesis 2. Note that this fraction of deceivers would not be identified by just combining the belief elicitation question with the message sent in the strategic sender-receiver game.

We found that a significant number of subjects (27/68, 40%) categorized as sophisticated truth-tellers using the previous method in the literature are just pessimistic, supporting Hypothesis 3. Note that this fraction of senders will be incorrectly identified as deceivers if we use the previous method in the literature to identify deceptive motives.

Moreover, we found that a considerable number (56/233, 24%) of senders designated as benevolent truth-tellers with the previous method in the literature are suspected of being deceivers according to our new method. In particular, they could be sophisticated truth-tellers who, in the belief elicitation question, hid their deceptive intentions of the strategic game by misreporting that they were expecting to be trusted. However, with our experimental design, we would identify the deceptive intentions in their message since they consequently lied in the non-strategic game.

Our further analyses point out that the uncertainty-driven alternative for this behavior proposed in Section 2.2 is not driving the behavior of these senders in our data since the sender's uncertainty about the receiver's trust (measured using question S_2) has no significant effect in explaining the message sent in the strategic game. This is analyzed by conducting several regressions to test the role of the sender's uncertainty (modeled using various functions to measure the distance to the full uncertainty point) in explaining the message sent in the strategic game. In none of the regressions[19] conducted, this variable has a significant effect, neither after controlling by the belief about the action of the matched receiver (S_1, which is statistically significant) nor not doing so. This evidence suggests that the uncertainty-driven alternative cannot explain our results, and those senders identified as incongruous truth-tellers cannot be considered to be plain honest but actually sophisticated truth-tellers who aim to deceive the receiver in the strategic sender-receiver game. It is important to note that, using the previous method in the literature, these senders will be incorrectly identified as subjects not intending to deceive with the message sent.

[18] Note that the percentages in the left balloons are the frequency of senders for each behavioral type categorized with the previous method in the literature over the total number of senders in our sample, whereas the percentages in the right balloons are the frequency of senders of the behavioral types detected with our new method over the ones in the correspondent parent category of the previous method.

[19] We used the absolute, the square and the linear function in our regression analyses. Section A in the Online Appendix contains the detailed results of all these regressions and the distance functions used.

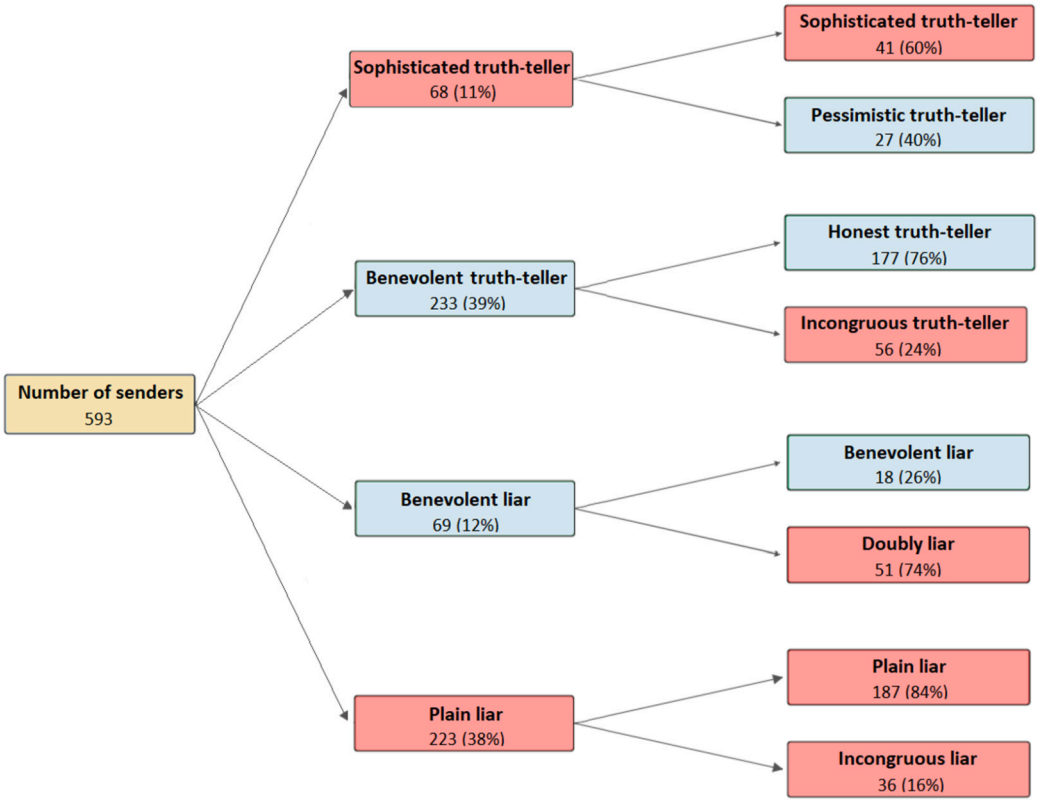


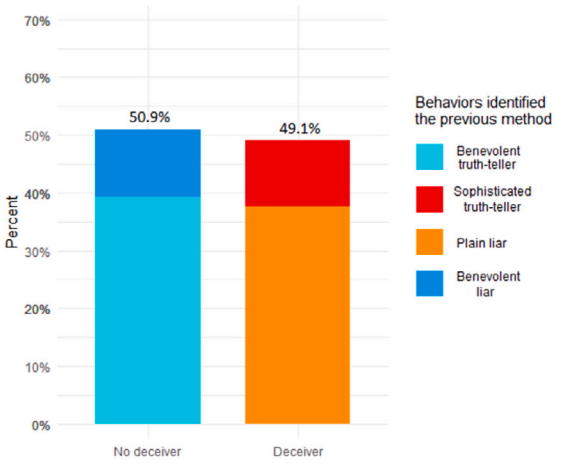
Fig. 3. Decomposition of the behavioral types categorized with the previous method into the new behaviors categorized with our new method. Percentages in parentheses are the relative frequencies with respect to the respective parent node. Behaviors in red (blue) are those who (do not) aim to deceive in the strategic game according to each identification method.

Furthermore, we conducted additional analyses to examine the role of S_2 in our findings. Firstly, we analyzed how our main results are influenced by senders who expressed complete uncertainty about the receivers' behavior (i.e., $S_2 = 50$), denoted as *fully uncertain*. After discarding them (ending up with 477 senders), there is no statistically significant difference for each behavioral type categorized with our new method (Pearson's Chi-squared test p-value = 0.92). Moreover, the fraction of fully uncertain senders obtained in our sample is not significantly different from the results observed previously in the literature (compared to Sutter (2009), Pearson's Chi-squared test p-value = 0.27). All these analyses are presented in more detail in the Online Appendix (see Section C) together with further checks where we expand the threshold to wider values of S_2. This provides additional evidence supporting the notion that uncertainty is not a driving factor behind our main results. Furthermore, we considered senders as *incoherent* if the combination of their answers to the two belief questions does not align with the message sent in the strategic sender-receiver game. All our main results remain preserved after ruling them out (ending up with 499 senders). Besides, upon excluding both incoherent and fully uncertain senders (ending up with 383 participants), our main findings are preserved or even strengthened, resulting in a higher relative proportion of doubly liars (which is increased to 95% of the number of senders categorized as benevolent liars with the previous method). For the sake of readability, we leave all these robustness analyses to Section C in the Online Appendix.

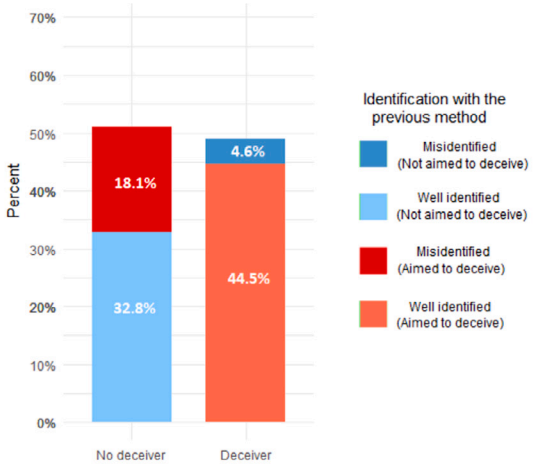
Confirming the result obtained using the previous approach in the literature, most senders either plainly tell the truth or lie in both sender-receiver games. However, the fraction of incongruous liars is lower than that of incongruous truth-tellers (proportion test p-value = 0.048). This is also in line with our previous arguments that this behavior has little impact on our experimental data. Notice that this behavior is diminished in our robustness analyses after ruling out fully uncertain or incoherent senders. In our view, this points out the lack of relevance of this behavior in our data.

Taking all the previous results together, we compare the fraction of senders that acted with a deceptive intention in the strategic sender-receiver game identified with the previous method in the literature and those using our new method. This comparison is graphically represented in Fig. 4. Specifically, Figs. 4a and 4d show the fraction of senders with deceptive intentions identified using the previous method in the literature and our new approach. Fig. 4c represents the transition between the two methods, displaying how the behavioral types categorized with our new approach would be distributed using the previous method in the literature. After all, 62.5% (371/593) of the senders in our total sample[20] aimed to deceive the receiver according to our new method. This

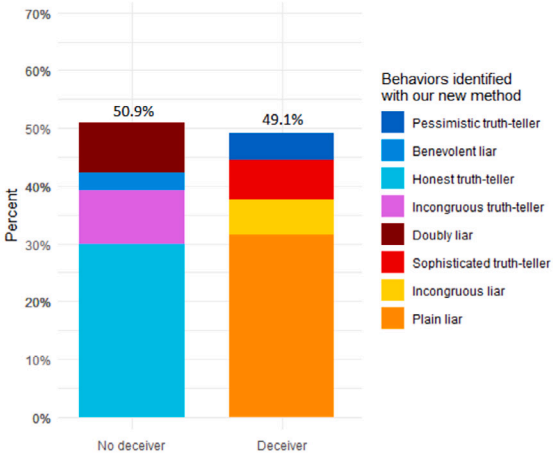
[20] Even if we observe a trend apparently suggesting that younger participants aim to deceive more than older ones, and that women aim to deceive more than men, we obtained no significant effect of age (t-test p-value = 0.097) or gender (t-test p-value = 0.061) in the fraction of senders identified with a deceptive intention



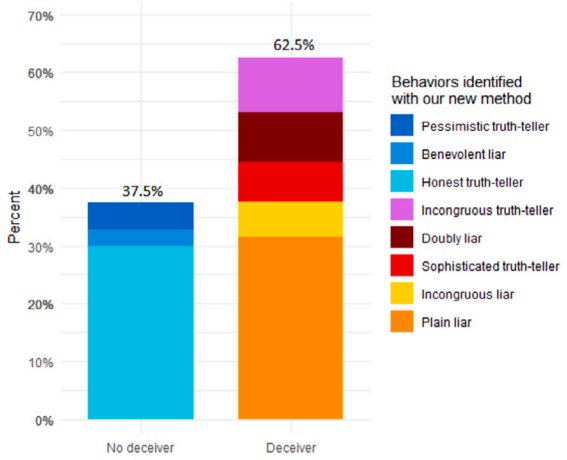
(a) Senders' behavioral types according to the method in the literature (Str. message, S_1).



(b) Number of misidentified deceivers and non-deceivers from the previous to our new method.



(c) Distribution of the behavioral types of our new method if identified using the previous method.



(d) Senders' behavioral types according to our new method (Str. message, S_1, NS message).

Fig. 4. Comparison of the number of senders acting with a deceptive intention captured using the previous method in the literature and our new method.

proportion is significantly higher than the 49.1% (291/593) of senders identified using the previous approach in the literature (Pearson's Chi-squared test p-value < 0.001).

This result occurs since the fraction of pessimistic truth-tellers (misidentified as deceivers using the previous approach) is significantly lower than the proportion of senders (both liars or truth-tellers) who are not properly identified with their behavior in the strategic task and whose deceptive intentions are detected with our new approach notwithstanding their misreporting in the belief questions that may hide their deceptive purpose. This can be observed graphically in Fig. 4b, which represents the number of senders misidentified (concerning their deceptive intentions) using the previous approach in the literature compared to our new method.

Finally, we conducted several tests to check for the possibility that any possible random behavior[21] drives the results from the categorization obtained using our novel method. First, we checked if the new behavioral types[22] proposed were solely driven by

using our new method. We refer the reader to Section G in the Online Appendix for a more comprehensive study on the effect of age or gender on senders' deceptive intentions.

[21] We are aware that a small amount of random behavior can always be captured in an experiment, and our aim with these analyses is not to state that any player could have been making all her decisions randomly. Our argument is to provide quantitative evidence that, if any, random behavior is not the factor that promotes the results we have already discussed.

[22] As a robustness check, we conducted analogous analyses taking all the non-plain categories, that is, all but honest truth-tellers and plain liars. All results are preserved, and the main results are presented in Section D in the Online Appendix.

random behavior: the results show that the frequencies of the four new behavioral types are jointly different from their respective average (which is equal to 7.17%, proportion test p-value = 0.009). Besides, we can reject the null hypothesis that each new behavioral type is equally likely to happen in our data (Fisher's exact test p-value = 0.003), which would be another potential indication of random behavior.

Moreover, we checked for the possibility that a fraction of senders playing randomly drives the observed results. To do so, we made a battery of tests,[23] one for each noise level (from 1% to 100%) for each of the new behavioral types proposed. We used proportion tests for the hypothesis that a percentage of the senders' is spread uniformly over the four new behavioral types (as it would be if driven by random behavior). We used individual binomial tests for each category to check what levels of randomness could be explained by the data we obtained.

To summarize the results of these randomness tests, we found that there is no level of randomness that can explain the frequencies we captured in the categories obtained with our new method. Specifically, there is no percentage of senders who play randomly that can explain the data in the four new behavioral types, i.e., the intersection of the levels for which we cannot reject the hypothesis of equal distribution across categories is empty.

Therefore, we consider the combination of these two outcomes as an indication that if there is any fraction of senders in the sample who act randomly, their behavior is not driving the results we obtained. Hence, Hypothesis 4 is confirmed, and the new behavioral types proposed are not driven by random behavior.

3.3. The receiver side

As found in previous contributions, most receivers trusted the message given. In particular, around 87% of the receivers (513/593) chose the option recommended by the sender's message, trusting[24] her recommendation. Another possible explanation for this high rate of receivers who follow the sender's recommendation is that receivers ignore the message obtained, choosing an option without considering the information received. To check if this confound is driving our trust results, we asked receivers about it using an unincentivized question. We found that only around 9% of the receivers (55/593) answered that they did not consider the information obtained. This result provides additional evidence supporting that senders' messages were considered by receivers when deciding, so trust drove receivers' actions. Alternatively, our data shows that receivers took into account the message acquired since, if they did not, they would have randomized the chosen option. However, as it has been shown before, the distribution of their choices is far from random.

Moreover, we elicited receivers' beliefs about the message's truthfulness in the two games. On average, receivers believed that the message obtained was true around two-thirds of the time, both in the strategic and the non-strategic game. Besides, those who did not choose the option suggested in the message believed that the probability of the message being true was lower than the belief of those who followed the message recommendation (Wilcoxon signed-rank test p-value < 0.001).

Comparing the belief in the truthfulness of the message between the two games allows us to examine whether the receivers' role in the game influences their perception of the message's veracity. Truthfulness beliefs in both games were not significantly different (Wilcoxon signed-rank test p-value = 0.229). An explanation for why the two games' beliefs are similar is that most of the receivers just followed the option suggested by the message received without considering its possible falsity since this was the only source of information about the payments they had to make a decision. An interesting avenue for future investigation is to study how the receivers' role in the game affects their beliefs about the message's veracity in settings where receivers have more information than the sender's suggestion.

Lastly, we elicited second-order beliefs from receivers to investigate whether they suspected that senders were employing sophisticated truth-telling behavior. This higher-order reasoning could be a factor influencing senders[25] not to use this deceptive method if they anticipated receivers' suspicions. We obtained that the answers to the first two belief questions (R_1 and R_2) were significantly different (Wilcoxon signed-rank test p-value = 0.003). This result, indicating differences between the belief about the message's truthfulness and the belief about senders' expectations of receivers' trust in the message, could be driven by the fact that some receivers expected non-plain behavior in the strategic game, in particular, sophisticated deception. We cannot provide more information about the drivers of this surprising result, but we consider it an interesting path for the future in order to know better the reasons why receivers trust more than what is expected from an equilibrium viewpoint in sender-receiver games.

4. Conclusion

In sender-receiver games where strategic considerations are allowed, senders can deceptively tell the truth to exploit a mistrustful recipient, implying that having a deceptive intention does not necessarily coincide with lying. Therefore, the sender's beliefs about the receiver's trustworthiness become crucial for identifying deceptive intentions in these games of information transmission. In this paper, we propose a new method to identify the deceptive intentions of senders in a sender-receiver game characterized by a conflict of interest, where senders may convey truthful information with the intention to deceive. With our new method, we are

[23] The results of the whole battery of statistical tests are reported in Section D in the Online Appendix.

[24] Receivers' trust was independent of the option suggested in the message received (Pearson's Chi-squared test p-value = 0.39). In our experiment, we obtained significantly higher trust rates than previous results in the literature, which ranged around 72% (proportion test p-value < 0.001).

[25] A part of the sender-receiver games literature shows how bounded rationality models such as the level-k reasoning (see Kawagoe and Takizawa (2009); Wang et al. (2010)) or the cognitive hierarchy model (see Li et al. (2022)) can explain senders' behavior.

able to identify four additional behavioral types that would be misclassified in terms of their deceptive intentions using the previous approach in the literature.

We conducted a two-stage experiment composed of two tasks: First, a replica of the classic sender-receiver game in the literature where strategic considerations are allowed, followed by a belief elicitation questionnaire about expected receivers' behavior. Second, an equivalent non-strategic version of the same game with no room for misleading intentions in the message transmitted.

To capture the senders' selfish intentions in the strategic game, we mainly use actions, i.e., the messages sent in the two games. Then, we refine the identification with the answer to the belief question used in the literature, discerning between different deceptive motives. This differs from the previous method, which only uses the belief about the expected receiver's action in the strategic game to infer the sender's intention in the message sent. With our new method, we find that around 40% of the senders who would be considered sophisticated deceivers – using the previous method in the literature – are actually not aiming to deceive with their message. These subjects hold a pessimistic belief about others' trust, being their truth-telling imputed not to a deceptive intention, consistently with their behavior in the non-strategic game.

Our results show that, in games of information transmission where strategic considerations are allowed, relying only on a question about the sender's belief can be insufficient to correctly identify senders' deceptive intentions. The reason is that those senders who aim to deceive in the game, either by telling a plain lie or by sophisticatedly telling the truth, might have incentives to lie also in a belief elicitation question that reveals to the experimenter their deceptive purpose. Then, these senders might opt for excusing their selfish behavior, preserving their deceptive intentions hidden. Our results show that around 74% of the senders in our data who appear to be benevolent liars (liars who do not aim to deceive) are actually deceivers who misreported their beliefs to justify the false message sent. Besides, around 25% of the senders who appear to be honest are actually suspicious to have aimed to deceive the receiver by telling the truth sophisticatedly.

We measured how many senders are differently identified (regarding their deceptive intentions) using the previous and our new method. According to our new method, about 62.5% of the senders in our sample aimed to deceive in the strategic sender-receiver game. This proportion is higher than the 49.1% implied by the previous method in the literature. This finding suggests that the single use of a belief elicitation questionnaire right after a sender-receiver game (together with the message's veracity) is not enough to correctly capture the sender's deceptive intentions when strategic considerations are allowed.

An interesting further research step in this regard is to inquire to what extent the heterogeneity implied by our new method is driven by a pure aversion to lying and to what extent it is driven by other sources of benefit from truth-telling, such as moral values or concerns for social image. Another promising avenue for future investigations is to study how the distribution of behavioral types obtained with our new method varies in different communication settings, such as allowing for evasive lying (Khalmetski et al., 2017) or providing a context with a richer language where partial truths are possible (Alempaki et al., 2019).

We also provide information about receivers' behavior and beliefs in the sender-receiver game, something little explored before. As in previous contributions, we obtained that receivers had substantial trust in the senders, and that this trust was independent of the message received. Regarding their stated beliefs, we found that two-thirds of the receivers believe the message obtained is true. Interestingly, we found this belief about the message's truthfulness independent of whether receivers have an active or passive role in the game. An explanation for this result is that the lack of information receivers have in this game might drive this "blind faith". Alternatively, receivers may fail to recognize the strategic and non-strategic nature of the games. Besides, around 9% of the receivers ignore the message to choose an option, and those who ignore it mistrust it significantly more. Our receivers' data also indicates that some of them might expect senders to tell the truth while expecting not to be followed. Since some receivers seem to mistrust and others to trust but expect sophisticated deception, we consider an interesting avenue for future research to analyze how the receiver's role and information about the game details affect their trust and beliefs about the message received.

To our knowledge, this is the first study to analyze deception in sender-receiver games where strategic considerations are allowed using an online setting. A few recent contributions provided information about lying behavior in games where no strategic concerns are possible, such as cheating games Dickinson and McEvoy (2021) or modified sender-receiver games where the opponent is an external institution that takes no action Janezic (2020). However, these games do not allow senders to act strategically, equalizing by design deception to dishonesty.

A word of caution needs to be said about quantifying the spectrum of behaviors concerning deceptive intentions in sender-receiver games. As we noted in Subsection 2.2, despite we have reasons to think that deceptive intentions are driving those truth-tellers whose stated beliefs are incongruous with their spiteful behavior in the non-strategic game, we cannot fully differentiate those who are aiming to deceive and hiding their intentions by misreporting their beliefs to the experimenter from those who are telling the truth driven by their uncertainty about the receiver's trust. Besides, we cannot exclude the possibility that some senders change their preferences according to whether the sender-receiver game has a strategic component or not. Another limitation of our study is that the density of some categories we detected could depend on the message space, the two-option design, or the conflict of interest between the incentives of senders and receivers in our experiment.

Declaration of competing interest

The authors declare no competing conflicts of interest.

Data availability

Data will be made available on request.

Acknowledgments

The authors gratefully acknowledge financial support from the Italian Ministry of Education, University and Research (MIUR) through the PRIN project Co.S.Mo.Pro.Be. “Cognition, Social Motives and Prosocial Behavior” (grant n. 20178293XT) and from the IMT School for Advanced Studies Lucca through the PAI project Pro.Co.P.E. “Prosociality, Cognition, and Peer Effects” (grant n. PAI2018 PROCOPE). Moreover, the authors gratefully acknowledge the financial support from the Italian Ministry of University and Research (MIUR) through the PRIN–PNRR project “Strategic Thinking Development in an Ever-Changing World” (grant n. P2022TALJF) funded by the European Union – Next Generation EU. We sincerely thank two anonymous referees and the editor for their constructive comments and helpful suggestions that helped us to greatly improve the paper. We also thank Valerio Capraro, Carlos Cueva, and Marta Serra-Garcia for many helpful suggestions and invaluable discussions. Further, we thank Vincent Crawford, Chiara Nardi, Axel Ockenfels, Matteo Ploner, Joel Sobel, Stefan Trautmann, and Emmanuel Vespa, as well as seminar participants at the IMT School for Advanced Studies Lucca and the University of Alicante for their useful feedback and comments. Many valuable comments were also received from participants in numerous conferences.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.geb.2024.02.006.

References

Abeler, J., Becker, A., Falk, A., 2014. Representative evidence on lying costs. J. Public Econ. 113, 96–104.
Abeler, J., Nosenzo, D., Raymond, C., 2019. Preferences for truth-telling. Econometrica 87 (4), 1115–1153.
Alempaki, D., Doğan, G., Saccardo, S., 2019. Deception and reciprocity. Exp. Econ. 22 (4), 980–1001.
Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games Econ. Behav. 10 (1), 122–142.
Besancenot, D., Vranceanu, R., 2020. Profession and deception: experimental evidence on lying behavior among business and medical students. J. Econ. Behav. Organ. 179, 175–187.
Biziou-Van-Pol, L., Haenen, J., Novaro, A., Liberman, A.O., Capraro, V., 2015. Does telling white lies signal pro-social preferences? Judgm. Decis. Mak. 10 (6), 538–548.
Capraro, V., 2017. Does the truth come naturally? Time pressure increases honesty in one-shot deception games. Econ. Lett. 158, 54–57.
Capraro, V., 2018. Gender differences in lying in sender-receiver games: a meta-analysis. Judgm. Decis. Mak. 13 (4), 345–355.
Capraro, V., Halpern, J.Y., Perc, M., 2022. From outcome-based to language-based preferences. J. Econ. Lit.
Capraro, V., Schulz, J., Rand, D.G., 2019. Time pressure and honesty in a deception game. J. Behav. Exp. Econ. 79, 93–99.
Charness, G., Dufwenberg, M., 2006. Promises and partnership. Econometrica 74 (6), 1579–1601.
Dickinson, D.L., McEvoy, D.M., 2021. Further from the truth: the impact of moving from in-person to online settings on dishonest behavior. J. Behav. Exp. Econ. 90, 101649.
Erat, S., Gneezy, U., 2012. White lies. Manag. Sci. 58 (4), 723–733.
Fischbacher, U., Föllmi-Heusi, F., 2013. Lies in disguise-an experimental study on cheating. J. Eur. Econ. Assoc. 11 (3), 525–547.
Glätzle-Rützler, D., Lergetporer, P., 2015. Lying and age: an experimental study. J. Econ. Psychol. 46, 12–25.
Gneezy, U., 2005. Deception: the role of consequences. Am. Econ. Rev. 95 (1), 384–394.
Gneezy, U., Kajackaite, A., Sobel, J., 2018. Lying aversion and the size of the Lie. Am. Econ. Rev. 108 (2), 419–453.
Gneezy, U., Rockenbach, B., Serra-Garcia, M., 2013. Measuring lying aversion. J. Econ. Behav. Organ. 93, 293–300.
Gylfason, H.F., Olafsdottir, K., 2017. Does Gneezy’s cheap talk game measure trust? J. Behav. Exp. Econ. 67, 143–148.
Hurkens, S., Kartik, N., 2009. Would I lie to you? On social preferences and lying aversion. Exp. Econ. 12 (2), 180–192.
Janezic, K.A., 2020. Heterogeneity in lies and lying preferences. Working paper.
Jordan, J.J., Rand, D.G., 2020. Signaling when no one is watching: a reputation heuristics account of outrage and punishment in one-shot anonymous interactions. J. Pers. Soc. Psychol. 118 (1), 57.
Kajackaite, A., Gneezy, U., 2017. Incentives and cheating. Games Econ. Behav. 102, 433–444.
Kawagoe, T., Takizawa, H., 2009. Equilibrium refinement versus level-k analysis: an experimental study of cheap-talk games with private information. Games Econ. Behav. 66 (1), 238–255.
Khalmetski, K., Rockenbach, B., Werner, P., 2017. Evasive lying in strategic communication. J. Public Econ. 156, 59–72.
Khalmetski, K., Sliwka, D., 2019. Disguising lies - image concerns and partial lying in cheating games. Am. Econ. J. Microecon. 11 (4), 79–110.
Li, X., Özer, Ö., Subramanian, U., 2022. Are we strategically naïve or guided by trust and trustworthiness in cheap-talk communication? Manag. Sci. 68 (1), 376–398.
Lohse, T., Simon, S.A., Konrad, K.A., 2018. Deception under time pressure: conscious decision or a problem of awareness? J. Econ. Behav. Organ. 146, 31–42.
López-Pérez, R., Spiegelman, E., 2019. Do economists lie more? In: Dishonesty in Behavioral Economics, pp. 143–162.
McEvily, B., Radzevick, J.R., Weber, R.A., 2012. Whom do you distrust and how much does it cost? An experiment on the measurement of trust. Games Econ. Behav. 74 (1), 285–298.
Palan, S., Schitter, C., 2018. Prolific. ac—a subject pool for online experiments. J. Behav. Exp. Finance 17, 22–27.
Peer, E., Brandimarte, L., Samat, S., Acquisti, A., 2017. Beyond the turk: alternative platforms for crowdsourcing behavioral research. J. Exp. Soc. Psychol. 70, 153–163.
Sobel, J., 2020. Lying and deception in games. J. Polit. Econ. 128 (3), 907–947.
Sutter, M., 2009. Deception through telling the truth?! Experimental evidence from individuals and teams. Econ. J. 119 (534), 47–60.
Thielmann, I., Spadaro, G., Balliet, D., 2020. Personality and prosocial behavior: a theoretical framework and meta-analysis. Psychol. Bull. 146 (1), 30.
Vanberg, C., 2017. Who never tells a lie? Exp. Econ. 20 (2), 448–459.
Wang, J.T.Y., Spezio, M., Camerer, C.F., 2010. Pinocchio’s pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games. Am. Econ. Rev. 100 (3), 984–1007.