symmetry

MDPI

Article

Truth-Telling in a Sender–Receiver Game: Social Value Orientation and Incentives

Hanshu Zhang [1,2,*], Frederic Moisan [3], Palvi Aggarwal [2,4] and Cleotilde Gonzalez [2,*]

[1] School of Psychology, Central China Normal University, Wuhan 430079, China
[2] Department of Social and Decision Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[3] EM Lyon Business School, GATE UMR 5824, F-69130 Ecully, France; mailto:fmoisan@gmail.com or moisan@em-lyon.com
[4] Department of Computer Science, The University of Texas at El Paso, El Paso, TX 79968, USA; paggarwal@utep.edu
* Correspondence: hanshuzh@ccnu.edu.cn (H.Z.); coty@cmu.edu (C.G.)

Abstract: Previous research has discussed the effects of monetary incentives and prosociality on deceptive behavior. However, research has not comprehensively investigated the relationship between these two factors. In the current research, we introduce a repeated two-player sender–receiver binary choice task, where players in the role of senders or receivers receive asymmetric information regarding payoffs, offering the opportunity to explore the effects of economic incentives to lie according to the players' prosociality. In Experiment 1, players are paired to play the game as a sender or receiver online. We find that economic incentives determine the likelihood of deception from senders and the likelihood that receivers will deviate from the received suggestions. Moreover, prosociality is related to players' behavior: Prosocial senders send less deceptive messages and prosocial receivers choose options that benefit senders more. Furthermore, senders display consistent behavior when interacting with receivers, and they do not change their deceptive behavior even if detected by receivers. Experiment 2 further investigates how the players' behavior corresponds to their understanding and interpretation of the other players' actions, by pairing players with computer algorithms that display consistent probabilistic behaviors. We observe that senders deceive receiver algorithms by sending truthful messages when they expect the message not to be followed, and receivers follow the received messages by choosing the option that benefits "honest" sender algorithms. While we find a consistent result that prosocial senders send fewer deceptive messages than they should when telling the truth is costly, prosocial receivers are less considerate of sender payoffs in algorithms' interaction.

Keywords: deception; sender–receiver game; social preference; social value orientation; human–machine interaction

check for updates

Citation: Zhang, H.; Moisan, F.; Aggarwal, P.; Gonzalez, C. Truth-Telling in a Sender–Receiver Game: Social Value Orientation and Incentives. *Symmetry* **2022**, *14*, 1561. https://doi.org/10.3390/sym14081561

Academic Editor: John H. Graham and David A. Becker

Received: 5 July 2022
Accepted: 26 July 2022
Published: 28 July 2022

1. Introduction

Deception is the act of intentionally changing or suppressing information (i.e., by the sender) to cause behavior changes in another agent (i.e., the receiver) to benefit the sender (c.f., [1]). Current findings have identified two main effects on deceptive behavior (see [2], for a recent review): personal factors (e.g., gender, age, major, etc.) and situational factors (e.g., normative cues, reward size, etc.). While research on situational factors has identified that the sensitivity to other people's loss might result in less dishonest behavior in a sender–receiver game (e.g., [3]), it is unclear whether greater rewards lead to more dishonest behaviors (e.g., [4–6]). It appears that there might be a joint effect between personal and situational factors in driving people's deceptive behaviors. In addition, people try to infer other players' actions in deciding whether to behave deceptively (e.g., [7]), resulting in the possibility that deception may not always be the best option for decision-makers in dynamic environments in their sequential decisions (c.f., [8]). In the current research, we explore the influence of monetary reward on the players' behaviors as a function of their sensitivity to other players' losses or gains (represented by their prosociality) in a repeated sender–receiver game. In what follows, we will give a brief literature review on the influence of reward amount, social preference, and interactive play on players' deceptive behavior. Then we summarize the hypotheses in the current study. A detailed description of the sender–receiver game will be introduced in the second section.

1.1. Reward Amount

Typically, in a sender–receiver game, one participant, the sender, privately observes the true state of the world and then sends information to the receiver, who makes a selection that determines the payoffs of both players (e.g., [3]). In such a design, the economic incentives are usually misaligned between the players. Senders mostly gain more when they send a deceptive message, and receivers gain less when they trust the deceptive message. Notably, greater rewards incentivize dishonesty: senders send more deceptive messages to achieve higher monetary payoffs. Ref.[3] reported that senders' dishonest behavior was also sensitive to the resulting gains or losses to receivers. Senders sent less deceptive messages if that resulted in higher costs to the paired receiver. Likewise, ref. [9] reported that players would like to lie if others would benefit from their deceptive behavior.

To comprehensively investigate the influence of monetary incentives on deceptive behavior of senders, ref. [10] categorized lies based on the intentions of senders and explanations of their behavior. In their categorization, a lie can harm the liar but help the other person, i.e., "altruistic white lies". Likewise, a white lie can also make both sides earn more, i.e., "Pareto white lies". Based on this categorization, the research found that a large fraction of participants were reluctant to tell a Pareto white lie. Furthermore, there was also a group of participants willing to tell an altruistic white lie even if the lie would hurt them a bit. Together, the study suggests that the participants reacted differently when considering the monetary influences on both players.

1.2. Social Preference

One of the individual differences, represented by prosociality, is related to the player's sensitivity to how others are harmed by or benefit from deceptive behavior. In general, players with high prosociality are less likely to lie. For example, ref. [11] reported that generous decision makers who gave away a large share in the dictator game were also much less likely to lie in a sender–receiver game even if deception would benefit both participants. In another study, ref. [12] invited participants to play both the sender–receiver game and the prisoner's dilemma game. They found that players who were altruistic and cooperative were more likely to tell a Pareto white lie that boosts both players' payoffs and an altruistic white lie that involves a cost to the liar.

In addition, senders may also behave in a way to make them appear honest. For example, ref. [13] reported a task in which nuns were asked to report an observed die number that would decide their payoffs. Compared to the students who lied to increase their profits, the nuns lied to decrease their profits to appear honest even if no social preferences are involved. Similarly, participants also behaved in a way to make them appear honest when the experimenter observed potential undesirable behaviors. Ref. [14] found that when participants were offered the opportunity to lie for higher incentives to misreport and misreporting could not be detected, the aggregated report behavior was close to the expected truthful distribution. This suggests that lying costs are large and widespread. Other studies reported that senders were willing to tell the truth in fear of being caught. Experienced players who had participated in the deceptive game were more likely to lie for their own reward [5]. Moreover, ref. [6] showed that when the standard cheating game was modified to eliminate the concern of being exposed as a liar, the participants responded to the incentives to lie.

Although it has been accepted that a person's preference to lie depends on monetary incentives (e.g., [3,10]), it is still unknown how the innate social preference and the amount of reward would drive people's deceptive behavior together. One group of researchers proposes that there is a cut-off line for the influence of incentives on players' propensity to lie. For example, ref. [15] found that more participants partially lied when the observed results cannot be verified compared to situations where the observed results can be verified. Similarly, ref. [16] proposed that when lying induced a preferred outcome over truth-telling, a person's decision on whether to lie may be insensitive to the size of the monetary incentive.

Meanwhile, another research group emphasized that discussion of reward amounts should also include their monetary impact on paired partners. Ref. [9] reported that when people's dishonesty would benefit others, they were more likely to view dishonesty as morally acceptable and therefore felt less guilty about benefiting from cheating. Ref. [17] found that altruistic people lied less if their lie hurt their partners. However, when lying had no effect on others' payoffs, altruistic players were equally likely to lie compared to non-altruistic players. Ref. [18] tested the development of social preference and lying aversion among children. In the study, the children were given the opportunity to lie in order to achieve their preferred outcome. They were also aware that their choice would influence both their own payoff and their partner's welfare. The study indicated that lying was driven mainly by selfish motives and envy. Children with stronger social preferences were less prone to deceiving, even when lying would benefit others at no monetary cost.

1.3. Interactive Play

Typically, in the sender–receiver paradigm, the move of the receivers determines the payoff of the two players. Therefore, receivers must consider the perspective of the other player and then act accordingly. Given the interdependence between a sender and a receiver, ref. [3] discussed that senders expected the receiver to trust the message and 78% of the receivers actually followed the recommendations received. In another study, ref. [16] estimated that only 66% of the subjects actually followed the recommendation, resulting in conflicting evidence to suggest whether the receivers decided to follow the senders or not. Other research also reported that senders and receivers would understand their role in the game differently. Players who acted as senders and sent deceptive messages believed that they received truthful messages in the role of receivers instead [19], suggesting a discrepancy in believing whether the message was truthful or not. Ref. [20] found that when receivers did not have explicit information on the payoff alignment, their trust in the received message was largely influenced by the competition or cooperative context instructed. In general, receivers' reasoning implied that they believed that there was a payoff conflict between their desired selection and senders' recommendation, consistent with the game design.

Furthermore, ref. [21] reported that the receivers correctly anticipated the senders' motivation to lie and took this information into account when choosing whether to receive the message. In their employed experimental paradigm, the sender's payoff only depended on the message they sent, and the receiver's payoff depended on whether they followed a message. The results indicated that the receivers significantly reduced their trust in the message if the sender lied to them previously. Similarly, ref. [22] found that receivers punished deceptive messages from senders more often than truthful messages by choosing the option that gave both players a lower payoff. Ref. [23] also reported that when receivers were offered a chance to reduce the payoff of both participants to zero, their willingness to punish the sender was greater after a deceptive message (also see [24]). Whether receivers believe that message, in return, also impacts the likelihood of senders' deceptive behavior. Specifically, a sender may choose a message with the expectation that the receiver would not follow the sender's (true) message, resulting in a "sophisticated truth-teller" and potential deception by telling the truth [7]. Together, the ability of the player to infer the intent of their partner also plays an important role in the sender–receiver game.

1.4. Current Study

Previous studies have shown that prosocial behaviors of players are related to their responses to the effects of monetary incentives in sender–receiver games. Also, it is important to infer the choices of other players. To further explore the factors discussed above in sender–receiver games, we will study the effects of financial incentives and prosociality in repeated sender–receiver deception games.

In this research, we quantify the prosociality of senders and receivers and expect that senders and receivers with high prosociality would be more sensitive to the degree to which other people are harmed or benefited by deception. We use Social Value Orientation (SVO) as a measure of individual social preferences (see [25], for a review). SVO is a well-established measure of social preferences that has been used in the study of the prisoner's dilemma [26], fiscal honesty [27], pro-environmental preferences [28], and others. SVO gives a quantitative indicator of individual social preferences that ranges from selfishness to generosity towards another person [29,30]. SVO provides a quantification of the degree of social preference of participants in our current research.

As mentioned above, the players' actions also depend on inferring how their paired partner would choose. Such recursive thinking and assumptions on knowing what a player would know about the other player are known in psychology as "theory of mind" (c.f. [31]). For example, in a paper–rock–scissors game, if one player is temporarily choosing paper more often, the other player would choose scissors as a result. Likewise, we expect that if receivers do not select the recommended option, then senders would adapt by sending the message with the expectation that receivers would not follow their suggestion [7].

To examine the above assumptions, Experiment 1 employs a repeated sender–receiver game that includes two motivated scenarios in which one of the incentives for the senders is to lie. First, we expect that senders lie to achieve higher monetary rewards. Second, we expect that the senders who score higher in SVO would deceive less compared to senders who have a lower SVO score. Moreover, we expect that senders would infer the actions of the receivers on their recommendations. Specifically, we hypothesize that senders would send more truthful information to appear honest if their lies were previously caught by receivers. In Experiment 2, we instruct players to play with computer algorithms that mimic different propensities in the role of sender or receiver. This also enables us to investigate to what extent the strategy of players is influenced by partner behavior. We expect to find generalized results that are consistent with Experiment 1 and players are able to act according to their partners' strategy.

2. A Sender–Receiver Game of Deception

The sender–receiver game includes a binary choice task between a safe option and a risky option in which "nature" first randomly selects *state* behind a risky option according to probability p. The sender is privately informed about the outcome of the risky option, including the potential reward for the sender and the receiver. The sender selects and sends a message to the receiver, which can be true or false. When observing this message, the receiver chooses between safe and risky options. This selection determines the payoff for each of the two players. The game is formally represented in its extensive form in Figure 1 where the nature node (in the center) initially determines the state of the world (according to probability p), the sender nodes represent the sender's decision points (choosing between messages M1 and M2) for each state of the world (the sender knows the actual state of the world, whichever it is), and the receiver's nodes represent the receiver's decision points (choosing between risky and safe cards) for each message received by the sender. The receiver's uncertainty over the true state of the world is characterized by the dashed line in Figure 1, indicating that she cannot distinguish the two different decision points for any given message received from the sender (i.e., the message may be truthful or deceptive).

Note that the options are not explicitly identified as risky or safe. The players only know that there are two options (i.e., A and B), and they learn the outcomes from their choices. The risky option gives the receiver the payoff $x_{receiver}$ and the sender a payoff

x_{sender} with probability p; and a payoff $y_{receiver}$ to the receiver and y_{sender} to the sender with probability $1 - p$. The safe option yields a sure payoff of $z_{receiver}$ and z_{sender}, to the receiver and the sender, respectively. First, the selection of the true state behind the risky option is determined with the probability p; this information is privately observed by the sender, who then sends to the receiver one of two messages: "option A will earn you $x_{receiver}$ points while it will earn me x_{sender} points" (message M1) or "option A will earn you $y_{receiver}$ points while it will earn me y_{sender} points" (message M2). The selected message can be true or false based on the true state behind the risky option. Observing the received message, the receiver then chooses one of the two options, A or B. Importantly, the sender's payoff is also influenced by the option selected by the receiver. More specifically, if the risky option is selected, the sender earns x_{sender} or y_{sender} depending on the true state of the payoff. If, instead, the safe option is chosen, the sender earns z_{sender}.

To focus on the interesting cases, we assume that $y_{receiver} < z_{receiver} < x_{receiver}$ and $x_{sender}, y_{sender} < z_{sender}$ to create misaligned incentives between the sender and the receiver and, therefore, generate profitable opportunities for deception. Furthermore, to maintain the strategic component of the game, we assume the receiver's indifference between the two options, i.e., $p \times x_{receiver} + (1 - p) \times y_{receiver} = z_{receiver}$ or $p = \frac{z_{receiver} - y_{receiver}}{x_{receiver} - y_{receiver}}$. All the payoffs and the probability p are assumed to be common knowledge between both players.

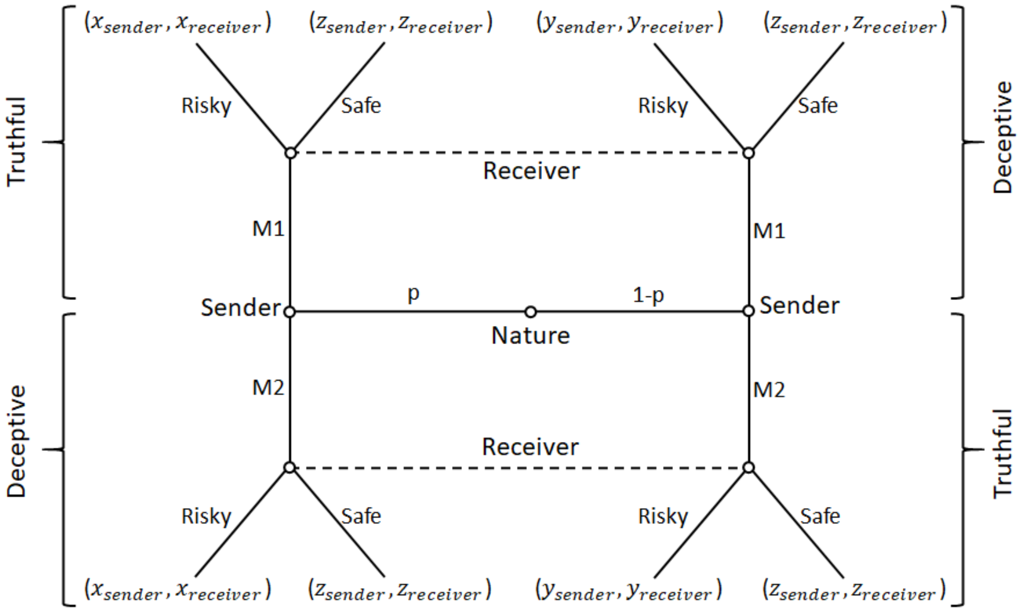

$(x_{sender}, x_{receiver})$
$(z_{sender}, z_{receiver})$
$(y_{sender}, y_{receiver})$
$(z_{sender}, z_{receiver})$
Truthful
Risky
Safe
Risky
Safe
Deceptive
Receiver
M1
M1
Sender
p
1-p
Sender
Nature
M2
M2
Receiver
Deceptive
Truthful
Risky
Safe
Risky
Safe
$(x_{sender}, x_{receiver})$
$(z_{sender}, z_{receiver})$
$(y_{sender}, y_{receiver})$
$(z_{sender}, z_{receiver})$

Figure 1. Sender–Receiver game such that $y_{receiver} < z_{receiver} < x_{receiver}$, $x_{sender}, y_{sender} < z_{sender}$, and $p = (z_{receiver} - y_{receiver})/(x_{receiver} - y_{receiver})$.

Some observations can be made from this deception game. First, the sender does not have a direct influence on the outcome of the game, which is determined solely by the choice of the receiver. The sender can only influence the receiver's choice by sending a true or deceptive message according to the state of nature p. Second, assuming that the players learn from the outcome feedback provided to both players over many rounds of this game, the receiver can learn whether the sender is sending false messages and the sender knows that the receiver can find this out. Third, the only way for the receiver to check the veracity of a message is to select the risky option. When choosing the safe option instead, the receiver cannot learn the results of the risky option.

In this context, the Nash equilibrium corresponds to the following: the sender sends a message independently of the true state of nature and the receiver chooses an option independently of the observed message. Note that the situation where the sender sends a truthful message and the receiver chooses an option consistent with this message is not stable because the sender would then be better off deviating by sending a deceptive message (see Appendix A.1 for the formal demonstration of the Nash equilibrium).

3. Experiment 1

To examine the effects of incentives and social preferences on the sender's deceptive intention, we measured the SVO of each participant and introduced a repeated deception game in which we manipulate the sender's incentives to lie ($x_{sender}, y_{sender} < z_{sender}$; Figure 1). Given that the deceptive message has no monetary effect on the receiver's payoff, we expect the senders to lie for their own benefit, and prosocial senders would lie less than individualistic senders. In addition, we expect that senders would adjust their strategies based on the receivers' choices accordingly.

3.1. Methods

3.1.1. Participants

Two hundred and forty subjects from Amazon Mechanical Turk (MTurk) were recruited for the task. Of these, 208 participants completed the entire task (age range: [20, 62], $N_{Male} = 106$), resulting in 104 pairs eligible for data analysis. Of the 104 pairs, we found that there were nine pairs in which the responses of at least one of the participants in the pair were not consistent with the SVO questionnaire (for more specific details regarding the method used, see Appendix A.2), thus their data were uninterpretable (c.f., [30]). These nine pairs were excluded from the analysis, resulting in 95 total pairs included in the final data analysis. Participants were informed that their earnings depended on their performance in the task (i.e., the cumulative points they received at the end of the task). The task took participants an average of 26 min to complete and participants who completed the task received an average of $5.04 payments based on their performance.

3.1.2. Design

The study consisted of a 2 × 2 mixed-subject experimental design. We manipulated the monetary incentives for the participants as shown in Table 1. Participants were assigned to the Cheap Truth or the Costly Truth condition (between-subjects), and in each condition, the risky options created two scenarios that decided whether senders had incentives to lie or not (within-subjects).

In both conditions, selecting the safe option will always result in 30 points for the sender and 20 points for the receiver; selecting a risky option will result in either 20 points or 0 points for the sender and 4 points or 36 points for the receiver, with an equal probability $p = 0.5$ (as discussed in the previous section, we assume that the state of risky option is decided with a random probability). Given that the expected value (EV) is 30 points for the safe option and 10 points for the risky option, the sender is economically motivated to send a message that would lead the receiver to choose the safe option under both conditions.

Note that when the risky option presents 4 points to the senders, choosing the safe option results in a higher payoff for both players. In this case, the sender has no (little) motivation to lie. Alternatively, when the risky option presents 36 points to the sender, choosing the risky option would result in higher payoffs for senders but not for the receiver, motivating senders to send deceptive information. In particular, the relative loss for the sender is 10 points in the Cheap Truth condition and 30 points in the Costly Truth condition. Therefore, the sender in the Costly Truth condition has larger monetary incentives to lie compared to the sender in the Cheap Truth condition.

Table 1. The payoff values for the experimental conditions Cheap Truth and Costly Truth.

	Payoffs Risky Option		**EV Risky**	**Payoffs Safe Option**
Condition	$(x_{sender}, x_{receiver})$	$(y_{sender}, y_{receiver})$		$(z_{sender}, z_{receiver})$
Cheap Truth	(20,36)	(0,4)	(10,20)	(30,20)
Costly Truth	(0,36)	(20,4)	(10,20)	(30,20)

Note. The risky option payoff $(x_{sender}, x_{receiver})$ or $(y_{sender}, y_{receiver})$ is decided with an equal 50% probability.

Interactive Sender–Receiver Game

We design a web-based, interactive two-player sender–receiver game. Figure 2 presents an example of the observed interface of the sender (named "messenger") and the receiver (named "decider") under the Cheap Truth condition. The sender (top panel) is presented with two randomly assigned cards colored red and blue, which were randomly assigned to the safe and risky options. In this example, the blue card corresponds to a safe option and the red card corresponds to the risky option. A safe option will provide 30 points with certainty to the sender and 20 points to the receiver. The risky option will provide 0 points with probability 0.5, 20 points otherwise to the sender; and 4 points with probability 0.5, 36 points otherwise to the receiver.

The lottery is drawn first, and only the sender observes the resulting payoff (i.e., highlighted in bold): the sender would receive 0 points, and the receiver would receive 4 points if the risky option is selected. After receiving this information, the sender is asked to select one of the two messages to send to the receiver. The message selected can be true or false based on the sender's observed risky option payoff; in this example:

False: "The red card will earn you 36 points and earn me 20 points."
True: "The red card will earn you 4 points and earn me 0 points."

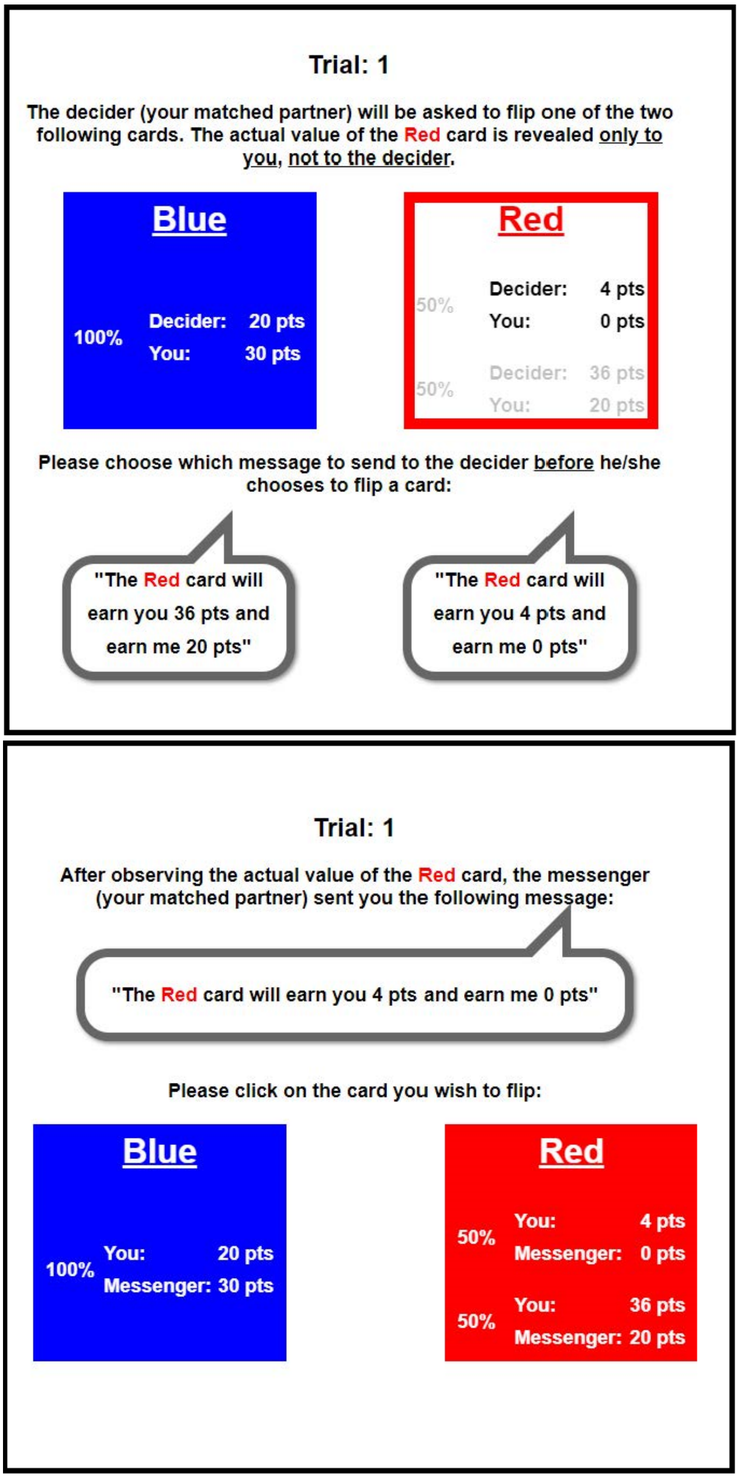
The receiver (bottom panel) observes the same cards together with the message sent by the sender (e.g., a true message in this example). The receiver then chooses one of the two cards. The receiver would determine whether the sender sent the true message or not if the risky card was flipped.

After the receiver selects a card, both players receive feedback on their choices and the earned results of the sender and receiver. Figure 3 shows the feedback received in this example. After viewing the feedback, players can move on to the next trial. Thus, based on the feedback, the sender would learn whether the receiver follows the recommendation of the message sent or not, and the receiver would learn whether the sender tells the truth or not if the risky option is chosen.

Social Value Orientation Scale

We used the SVO slider scale as the reliable SVO measure of social preferences [25,30,32]. Figure 4 gives an example of one of the 15 items of the SVO measure. In this item, a player moves a continuum of the joint payoff slider to assign points to the self and to another player. For example, if the slider is moved all the way to the right, that would assign 85 points to the self and 85 points to the other matched player. If the slider is moved all the way to the left, then that would assign 100 points to the self and 50 points to the other matched player. In the example of Figure 4, the player receives 93 points, while the other matched player receives 68 points based on the indicated slider position.

This SVO measure includes six primary items and nine optional secondary items that consist of different payoff allocation structures. The primary elements define whether the player is altruistic, prosocial, individualistic, or competitive; the secondary elements help differentiate the motivations of prosocial individuals: whether they search to minimize inequality or maximize joint gains. Similarly to one of our previous studies [26], only primary items were included in our analyses. We refer readers who are interested in the measurement of secondary items to previous research by Murphy and colleagues (e.g., [25,30]).



Trial: 1

The decider (your matched partner) will be asked to flip one of the two following cards. The actual value of the Red card is revealed only to you, not to the decider.

Blue

100% Decider: 20 pts
You: 30 pts

Red

50% Decider: 4 pts
You: 0 pts

50% Decider: 36 pts
You: 20 pts

Please choose which message to send to the decider before he/she chooses to flip a card:

"The Red card will earn you 36 pts and earn me 20 pts"

"The Red card will earn you 4 pts and earn me 0 pts"

Trial: 1

After observing the actual value of the Red card, the messenger (your matched partner) sent you the following message:

"The Red card will earn you 4 pts and earn me 0 pts"

Please click on the card you wish to flip:

Blue

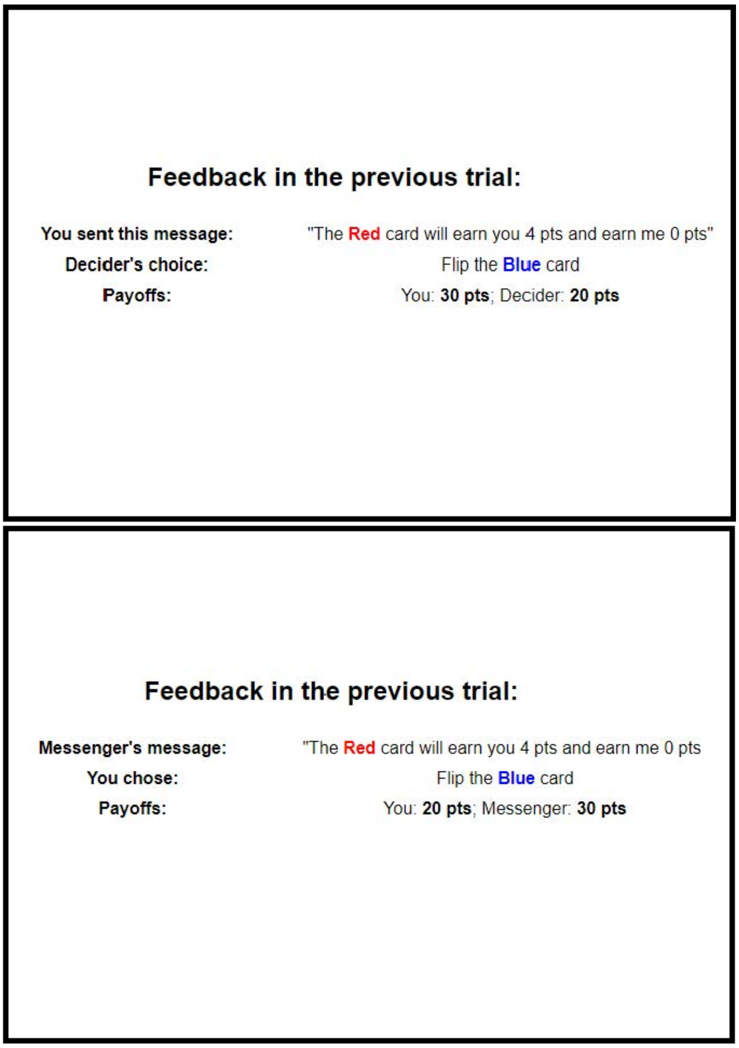
100% You: 20 pts
Messenger: 30 pts

Red

50% You: 4 pts
Messenger: 0 pts

50% You: 36 pts
Messenger: 20 pts

Figure 2. An illustration of the experiment trial. Top panel shows the options of the sender. Bottom panel shows the information given to the receiver. The two options are randomly assigned colored cards. In this given example, the safe option is the blue card and the risky option is the red card.



Feedback in the previous trial:

You sent this message:	"The **Red** card will earn you 4 pts and earn me 0 pts"
Decider's choice:	Flip the **Blue** card
Payoffs:	You: **30 pts**; Decider: **20 pts**

Feedback in the previous trial:

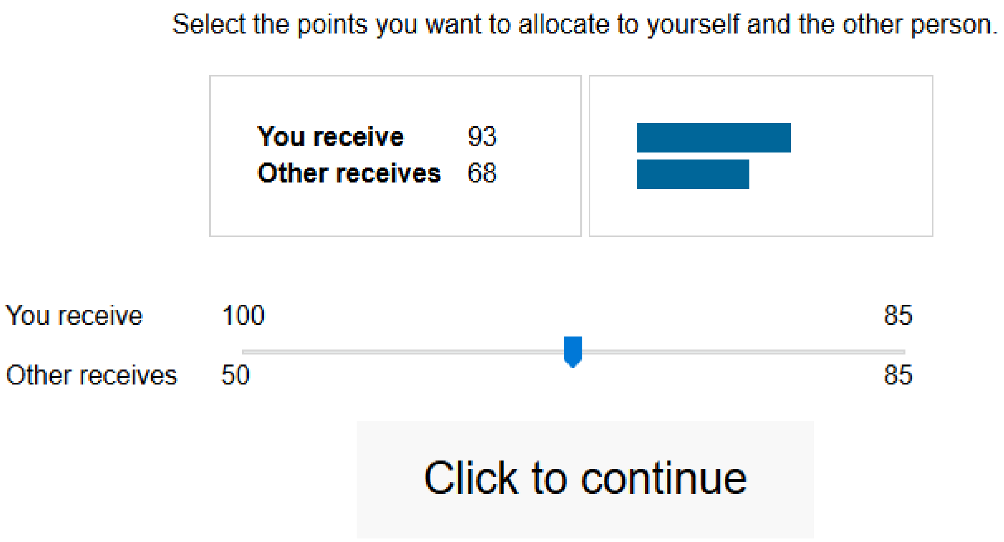
Messenger's message:	"The **Red** card will earn you 4 pts and earn me 0 pts
You chose:	Flip the **Blue** card
Payoffs:	You: **20 pts**; Messenger: **30 pts**

Figure 3. An illustration of the feedback after a choice trial the Cheap Truth condition. Top panel shows the feedback received by the sender. Bottom panel shows the feedback received by the receiver

The response to primary items yields a continuum score of the player's social preference. Based on the average of allocations for herself (A_s) against another player (A_o), each player's SVO (α) is obtained by (Note that we used the ratio instead of the angle that $tan(\alpha) = SVO$ for measuring SVO scores as reported by [30]. There are no qualitative differences in drawing conclusions between the two measures, also see discussion in [29]):

$$\alpha = \frac{A_o - 50}{A_s - 50}. \tag{1}$$

In general, SVO scores α on the primary items indicate prosociality (i.e., the willingness to assign reward points to the other player at the cost of their own received points). As shown in Figure A1, the distribution of the SVO scores suggests that the participants in this study generally fall into two categories: individualism and prosociality. This bimodal pattern is consistent with previous studies (e.g., [26,30]). The α in our study falls into the range of [−0.29,1.8]: trying to minimize the other player's payoff and maximize the other player's payoff, respectively. The three boundaries that separate categories are: 1.55, 0.41, −0.2, mapping from categories in [30].



Select the points you want to allocate to yourself and the other person.

You receive	93
Other receives	68

You receive	100	85
Other receives	50	85

Figure 4. An example item of the SVO measurement in the experiment.

3.1.3. Procedure

Participants were asked for informed consent according to the protocol approved by the Institutional Review Board at Carnegie Mellon University. Then, all participants were asked to read the same general task instructions before being directed to complete a brief demographic survey about their age, gender, and education level.

The participants were then asked to respond to the SVO items. Each participant was instructed to select the allocation of points for 15 items between themselves and an anonymous MTurk participant who signed up for this study. In addition, one of these 15 chosen allocations would be randomly selected to determine the payoff for the player and the other participants in the first part of the experiment. The participants were exclusively informed that they would not be matched again with the same participant in the following parts of the experiment and that the selected allocation would be unknown to the other participants.

After finishing the SVO survey, participants were notified that they would be matched with another online participant available in the MTurk pool. A participant waited to be matched with another participant for as long as 10 min. If no other participant was found, the participant was thanked and paid for the waiting time. If another participant was available, a match was made. The participants did not receive any information on the other player with whom they were matched. After being matched, they were randomly assigned a role (i.e., sender or receiver) and started the interactive sender–receiver game based on the monetary incentive conditions they were assigned, where the sender made a move first. The pair of players went through 60 trials of the game. The sender–receiver game was the only part of the experiment in which the two participants had real-time interaction.

Finally, all participants were instructed to complete an open-ended question about their strategy and their feelings towards the paired participant in the game. The participants were paid according to the points they accumulated on the task.

3.2. Results

Given that the current study extends our game from a traditional one-shot game to repeated measures, we first examine whether participants understand the game by calculating the results in the first trial. Our results indicate that in the Costly Truth condition, 15 out of 24 senders lied in cases where they were motivated (i.e., observed risky outcome (0,36)), significantly higher than the group compared to 5 out of 25 senders who were not motivated to lie (20,4), $\chi^2(1) = 7.48$, $p = 0.006$. In contrast, in the Cheap Truth condition, only 6 of 27 senders chose to lie when the observed risky outcome was (0,4), not statistically different from the scenario in which they observed that the risky outcome was (20,36) (6/19), $\chi^2(1) = 0.14$, $p = 0.71$. This suggests that our incentive manipulation is effective for the first trial in the designed sender–receiver game, even without the receivers' feedback.

Our dependent variable on the analysis is the choice made by each player. In the case of the sender, we averaged the proportion of deceptive messages given the observed outcome for the risky choice on every 15 trials. In this way, we examined whether senders changed their behavior over all 60 trials in four blocks. Likewise, we coded each of receivers' decisions made as being risky or safe, and again averaged that proportion over every 15 trials, in four blocks. We used JASP [33] for repeated ANOVA analysis with individual SVO score as a covariate and Greenhouse–Geisser for sphericity correction. We report post hoc pairwise comparisons with the Bonferroni p value adjustment. The data and instructions for the experiments have been made publicly available in the Open Science Framework (OSF) and can be accessed at https://osf.io/fmbv7/ (accessed on July 27, 2022).

3.2.1. Lying Propensity and Risk Taking

Sender's Lying Propensity

Figure 5 presents the proportion of deceptive messages sent by the sender based on the result of each block according to the experimental condition. First, the sender sends more deceptive messages when it receives more benefits from lying: the proportion of deceptive messages is higher when the outcomes observed by the sender are (20,36) or (0,36) than when they are (0,4) or (20,4). In contrast to our initial hypothesis that the sender should tell the truth when the risks in the option are (0,4) or (20,4) as the safe option is more beneficial to both players, the sender still sends deceptive information. In addition, we observe a much larger proportion of deceptive messages sent by the sender in the Costly Truth condition than in the Cheap Truth condition. Therefore, the sender lies more when it is more costly to tell the truth. By separating the sender's behavior into four blocks, we also find that the sender's behavior is relatively consistent across four blocks.

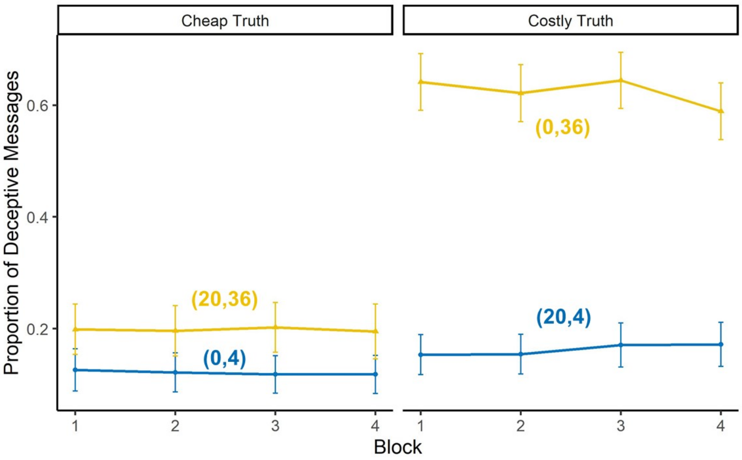


Figure 5. The proportion of deceptive messages sent among senders. The error bars indicate the standard error within each group. See the text for details.

To test these observations, we performed repeated ANOVA analysis on the proportion of deceptive messages sent by senders. The results indicated that there was only a main effect on the influence of risky option payoff on senders' deceptive behavior, i.e., the observed outcomes ($F(1, 92) = 27.11, p < 0.001, \eta_p^2 = 0.23$). The proportion of deceptive messages sent by the sender was higher when the results were (20,36) or (0,36) than when the results were (0,4) or (20,4), $t = 7.27, p < 0.001, d = 0.93$. Furthermore, there was a significant interaction between the observed results and the truth condition ($F(1, 92) = 22.67, p < 0.001, \eta_p^2 = 0.20$). There was a main effect of the truth condition ($F(1, 92) = 28.21, p < 0.001, \eta_p^2 = 0.25$), as senders in the Costly Truth condition sent more deceptive messages than in the Cheap Truth condition, $t = 5.31, p < 0.001, d = 0.74$; as well as their social preference ($F(1, 92) = 6.97, p = 0.01, \eta_p^2 = 0.07$). For further interpretation of the influence of the covariate, we ran the linear regression on SVO against the senders' deceptive behavior. The results indicated that SVO was a negative predictor of the proportion of deceptive messages sent by senders ($t(188) = 3.05, p = 0.003$). For each unit increase in the sender's SVO, there was a 0.26 decrease in the proportion of deceptive messages sent.

Post hoc comparisons of the interaction further indicated that on average senders sent more deceptive messages when the observed outcome was (0,36) compared to the scenario when the observed outcome was (20,4) in the Costly Truth condition, $t = 8.63, p < 0.001, d = 1.55$, whereas such difference was not significant in the Cheap Truth condition. In addition, senders in the Cheap Truth condition also sent a less deceptive message when the risky option benefited senders more, i.e., (20,36), compared to a similar beneficial scenario (0,36) in the Costly Truth condition.

Receiver's Risk Taking

Figure 6 presents the proportion of risky cards chosen by the receiver based on the message received from the paired sender for each block according to the experimental condition. First, the receivers reacted to the message from the sender: the risky option was selected more often when the message received was (20,36) or (0,36) than when the messages were (0,4) or (20,4), indicating that the receivers still followed the senders' recommendations. Furthermore, the figure also indicated that receivers selected the risky option less in the Costly Truth condition when informed that it was more profitable (0,36) compared to the profitable message in the Cheap Truth condition (20,36). Similarly to our observation of the behavior of senders, we also found that the selections were consistent across blocks.

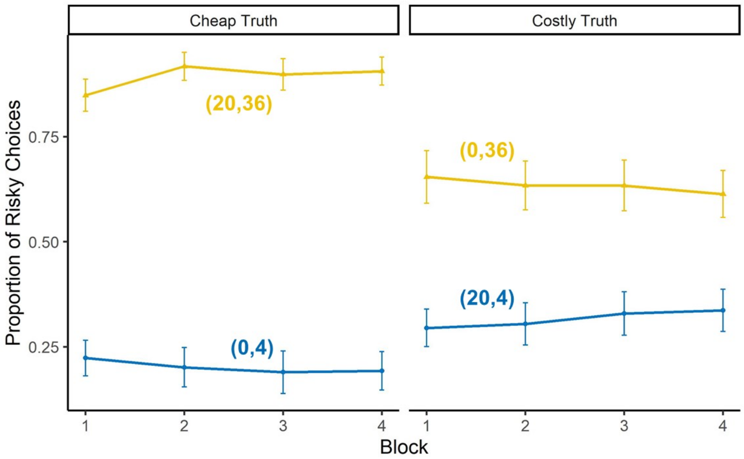


Figure 6. The average proportion of cards with risk to the receivers. Error bars indicate the standard error within each group. See the text for details.

The results of the ANOVA analyses (there were 21 receivers' choices excluded from the analysis because they only received one type of information, thus we lack evidence of the within-subject variance, thus leaving 74 subjects in the analysis for this section) indicated that, there was a significant main effect for the received message ($F(1, 71) = 44.83, p < 0.001, \eta_p^2 = 0.39$); receivers selected more risky cards when informed that the risky card would bring them more profits compared to the less beneficial message ($t = 11.02, p < 0.001, d = 1.66$). There was also a significant interaction between the message received and the truth condition ($F(1, 71) = 25.24, p < 0.001, \eta_p^2 = 0.26$). Averaged across different blocks, the post hoc analysis indicated that all pairwise comparisons were significant. In addition, there was a three-way interaction between the block, the message received, and the truth condition $F(2.59, 183.61) = 2.93, p = .04, \eta_p^2 = 0.04$. Social preference was significantly related to their risk-taking behavior, $F(1, 71) = 8.82, p < 0.001, \eta_p^2 = 0.15$. Likewise, we ran the linear regression on the SVO against the receivers' proportion of the risky choice. SVO was also a negative predictor of the proportion of risky choices selected by the receivers ($t(166) = 3.03, p = 0.003$). For each unit increase in the receiver's SVO, there was a 0.31 decrease in the proportion of risky choices made.

To further interpret the three-way interaction, separate analyses with the block and received message repeated ANOVA were then conducted on the Costly Truth condition and the Cheap Truth condition, respectively. For the choices in the Costly Truth condition, there was a main effect of the message received, $F(1, 30) = 4.26, p = 0.048, \eta_p^2 = 0.124$. When the message received indicated that the outcome was (0,36), receivers selected more risky options compared to the scenario when they received the message indicating that the risky option would give them only 4 points, $t = 3.74, p < 0.001, d = 0.78$. For choices in the Cheap Truth condition, there was a main effect of the received message, $F(1, 40) = 65.20, p < 0.001, \eta_p^2 = 0.62$. Similarly, receivers selected more risky options when informed that the observed outcome was (20,36) compared to the scenario when the message received indicated that it would only give them 4 points, $t = 12.61, p < 0.001, d = 2.75$. Thus, we can conclude that the influence of the messages received on the proportion of risky choices depends on the truth condition.

3.2.2. Interactive Play

Next, we examine the ability of the players to infer other players' actions. We examine the hypothesis that senders would switch from sending deceptive messages to telling truth if their deceptive messages have been "checked" by receivers.

To analyze this type of behavior change, we marked the trials where senders sent deceptive messages and receivers chose the risky card; thus lies were detected. In addition, we also collected information on whether the senders continued with their deception or switched to sending a truthful message instead. To ensure that we had enough observations ($N \geq 15$) in which the receivers were able to catch the lies of the senders, we only included 16 pairs in our analysis.

Following the work of [34], we define this type of deceptive message adjustment of the senders after being caught as

$$p_{adjustment} = p(sub\,deception|detected) - p(sub\,deception|not\,detected)$$

This gives us the difference in conditional probability in subsequential deceptive messages after their lies were detected compared to the scenarios in which lies were not detected. If the senders changed their behavior, we would expect that $p(sub\,deception|detected)$ is statistically different from $p(sub\,deception|not\,detected)$. Figure 7 shows the calculated proportion of our included pairs. The one-sample t-test against the baseline of 0 also indicated that there was no statistical difference in the adjustment proportion in the Costly Truth condition, $t(10) = 0.42, p = 0.69$; nor in the Cheap Truth condition, $t(4) = 0.67, p = 0.54$. Therefore, the senders did not change their deceptive behavior after their lies were detected.

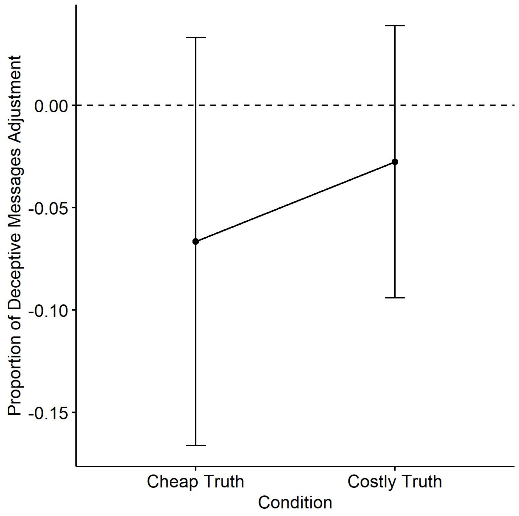


Figure 7. The average proportion of the adjusted proportion of lies. Error bars indicate the standard error within each group.

3.3. Discussion

In Experiment 1, both senders and receivers acted according to the economic incentives that provided more benefits to themselves. Consistent with previous findings, we found that senders lied when there was a greater monetary incentive to lie. Under both conditions, senders sent deceptive messages regarding the payoffs for the risky option when the safe option brought more payoffs to them but not to the receivers. Specifically, as in [3], we found that the senders were sensitive to misaligned incentives between the two players. Senders lied more when they were to suffer a relative loss of 30 points by telling the truth in the Costly Truth condition, compared to a relative loss of 10 points in the Cheap Truth condition. The senders also lied when it would have been more beneficial to both players if the senders had told the truth and the receivers had followed the recommendation. Based on the senders' responses to the survey we collected after the sender–receiver game, we provided two possible explanations for this behavior: One possibility was that senders tried to reinforce the receivers' choice of the safe option by telling receivers that the risky option was not profitable; another was that senders lied with the expectation that their messages would not be followed, and thus the risky option would be chosen in return. To explore these possibilities, in Experiment 2, we investigated the influence of receivers' actions on senders' lying behavior in further detail.

Our results also indicate that the receivers reacted to the senders' messages. The receivers were willing to choose the risky option when the message received indicated that the risky option would benefit them more. The selection was also modulated by the amount of misaligned incentives for senders. When senders had more motivation to lie under Costly Truth conditions, receivers were less willing to follow the message, thus choosing the risky option less often. Furthermore, receivers also chose the risky option more often when the sender's message suggested that the safe option would benefit both players. Therefore, the sender's economic incentive not only affected the sender's behavior but also the receiver's behavior, suggesting that the receivers were able to use their ability to infer the sender's lying motivations (e.g., [20,21]).

One major question we asked before was whether there was a cutoff line for the effect of reward amount on players' prosocial behavior. To answer this question, we designed two scenarios in our repeated sender–receiver game where senders had motivations to lie for higher monetary payoffs and receivers decided whether to benefit both players. Our results indicate that players' prosociality, measured by SVO, is independent of the observed outcome (for senders) and received message (for receivers). In this way, our work seems to be consistent with the previous work that once the sender decides to lie, the lying propensity is independent of monetary incentives [16]. In addition, our work also suggests that the likelihood of selecting the safe option that brings more benefits to both players is influenced by whether the players are prosocial or not, rather than by messages indicating which option is more profitable. There are two potential explanations for the connection between social preferences and risky choices. One is that prosocial receivers simply were risk averse and preferred certain payoffs; another is that prosocial receivers were economically indifferent to the options and they tried to benefit the senders (i.e., the senders' expected value is 20 points regardless of the senders' message). To further investigate the connections of prosociality and risk aversion, we would include a measure for players' risk preference in Experiment 2.

Our study further examined players' behavior change in the repeated sender–receiver game. First, the block effect was not the main effect in either analyzing the sender or the receivers' behavior. Therefore, although we designed a repeated deception game, the players played a consistent strategy in the 60 trials. More importantly, we also found that senders did not reduce their likelihood of lying after their deceptive message was checked by receivers. Together, our results implied that players did not adapt to their partners' strategy. To further examine the ability of players to infer their partners' strategy, we designed computer algorithms playing as senders and receivers in Experiment 2.

In summary, we investigated the effects of economic incentives and social preferences on the behavior of the players in Experiment 1. Our results indicate that prosocial senders sent fewer deceptive messages and prosocial receivers chose the safe option more often. While our analysis of receivers' performance indicated their ability to infer senders' high lying propensity in the Costly Truth condition, we did not find evidence suggesting that players changed their behavior during the game, and senders switched to truth-telling after their lies of being caught. To further examine the extent to which players will revise their strategies during the interactive sender–receiver game, we designed Experiment 2, which explored more directly the influence of the other player's strategy.

4. Experiment 2

To examine whether the sender and receiver are able to adapt to their paired players' strategies, we created three different computer bots (computer algorithms) that played predefined probabilistic actions in the role of sender or receiver. We expect that players' behavior also depends on the actions of paired players. For example, receiver bots would induce senders to be truthful about the messages by acting against their message. Similarly, deceptive bot senders would induce the receiver to deviate from the actions of the received messages. In addition, we assume that there is a learning effect as players interact more with computer bots in the repeated game, thus knowing their strategies.

4.1. Methods

4.1.1. Participants

A total of 371 participants signed up for the MTurk study. Among these participants, 61 of them failed the attention check (explained below), while another 35 participants did not complete the task. In total, we have 265 participants (Age range: [18,71], $N_{male} = 149$, $N_{female} = 114$, $N_{not\ revealing} = 2$) who completed the study and only these participants were eligible to receive the payment. Participants received \$2 as the base payment in addition to the bonus based on the points they accumulated during the task (25 points equal to 1 cent). After applying the same SVO transitivity check as in Experiment 1 (see Appendix A.2), we excluded another 38 participants among these 265 participants, and this left us 227 participants in the data analysis, 108 participants played the role of a sender, and 119 participants played the role of a receiver.

4.1.2. Design

As in Experiment 1, the incentive to lie (i.e., Cheap Truth and Costly Truth; Table 1) is the between-subjects in Experiment 2. Additionally, Experiment 2 designs three different computer algorithms (bot-type). The bot-type is a within-subjects factor that determines the probability with which the bot takes certain behaviors in its role as the sender or receiver.

Sender Bot

The three bots in the sender role are designed so that each exhibits lying behaviors with different proportions. As observed in Experiment 1, the sender's motivation to lie depends on the risky option's payoff: if the risky option brings the receiver 4 points ((0,4) or (20,4), Table 1; $p = 0.5$), the sender has less motivation to lie because the safe option (30,20) is more beneficial than the risky option for both players. However, if the risky option brings the receiver 36 points (i.e., (20,36) or (0,36)), the sender has more motivation to lie, as the safe option (30,20) is more beneficial for the sender than it is for the receiver. Given the results of Experiment 1, we design the bots as shown in Table 2. The table presents three sender bots that vary in the probability of sending deceptive messages when the observed result is (0,36) or (20,36). A *Truthful* bot sends deceptive messages at low levels of 20% rates, a *Deceptive* bot sends a deceptive message at high levels of 80% rates, and a *Random* bot sends a deceptive message randomly (i.e., 50%). If the observed result is (0,4) or (20,4) the sender bot sends deceptive messages at 10% rates. With these rates, we expect human receivers to follow the message of a truthful bot more often than the message of the deceptive bot.

Table 2. The proportion of deceptive messages for the Cheap Truth and Costly Truth experimental on the payoff of the risky option.

Condition	**$(x_{sender}, x_{receiver})$**	$(y_{sender}, y_{receiver})$		
		Truthful Bot	**Random Bot**	**Deceptive Bot**
Cheap Truth	(0,4) $p = 0.1$	(20,36) $p = 0.2$	(20,36) $p = 0.5$	(20,36) $p = 0.8$
Costly Truth	(20,4) $p = 0.1$	(0,36) $p = 0.2$	(0,36) $p = 0.5$	(0,36) $p = 0.8$

Note. The risky option payoff $(x_{sender}, x_{receiver})$ or $(y_{sender}, y_{receiver})$ is decided with an equal 50% probability.

Receiver Bot

The three receiver bots are designed to manipulate the proportions in which the receiver relies on the message sent by the human sender by choosing the risky option. When the human sender sends a message indicating that the risky option is more profitable (36 points) than the safe option (20 points), a *Following* bot follows the sender's message at high level 80% rates; a *Defecting* bot only follows the message at low level 20% rates; and a *Random* bot follows the message at 50% rates. Given the human sender's ability to infer bots' lying propensity, we hypothesize that they will send more deceptive messages to the Following bot than to the Defecting bot, since humans are expected to understand that their benefits depend on whether their suggestions are followed by the receiver or not.

Social and Risk Preference

Following Experiment 1, we also implement the SVO slider scale to measure the social preferences of human players. The results of Experiment 1 suggest a main effect of individual prosociality on sender and receiver behavior. Therefore, we expect that the player's prosociality will play a generalized role in driving the sender's motivation to lie and the receiver's preference in selecting the safe option. Recall that in Experiment 1, our results did not distinguish whether the prosocial sender chose the safe option due to their risk-averse preference or their willingness to benefit receivers. Therefore, we collect information on the players' risk preferences in Experiment 2: All players are instructed to rate their level of willingness to take risks. Players are asked: "Rate your willingness to take risks in general" on a 10-point scale, with 1 equal to completely unwilling and 10 completely willing [35].

4.1.3. Procedure

Participants went through the same procedure as in Experiment 1, the only difference is that they were paired with a bot instead of another human in the sender–receiver game. Participants were asked for their informed consent, a brief demographic survey, and their ratings of risk preferences.

Participants were also asked to complete the SVO survey (SVO distribution depicted in Figure A2). They were notified that they would select the allocation of points for SVO items for the player and another partner. Although we did not match their selected payoffs for the bot player, participants as the human player were still paid for their selected item, randomly chosen as in Experiment 1. In the sender–receiver task, the instructions were updated so that the participants knew they would play the game with computer algorithms. They were exclusively notified that they would interact with three different types of bots (computer algorithms) for each session, respectively, in the task, but were not given information regarding the exact behavior patterns of the bots.

Furthermore, we inserted an attention check trial that was randomly located between the first trial and the 20th trial (exclusively in the first session). The attention check was the same as a regular trial except that participants were instructed to flip a specific card rather than making their own trial judgment. Receivers were asked to flip the card on the right side (ignoring the information), while senders were instructed to send the truthful information.

Participants interacted with each type of bot during a 60 trials session. Each participant completed three sessions, for a total of 180 trials, with the order of the bots in each session being decided randomly. To minimize carryover effects, participants were also notified each time before the session started. Finally, all participants were asked to complete a survey about their strategy in each session. Participants were paid based on the points they accumulated during the task.

4.2. Results

For the collected responses from senders and receivers, we performed ANOVA analyses (Greenhouse–Geisser correction for sphericity where appropriate) and post hoc comparisons similar to those in Experiment 1. Given that the bots type was introduced as a within-subject factor, we separated our analysis for the Cheap Truth and the Costly Truth condition, respectively. To examine the adaption of participants to computer bot strategies, we also averaged the behavior of participants every 15 trials (i.e., four blocks) in each interactive session. The initial analysis indicates that the risk preference is not related to SVO for both senders ($r(106) = 0.08, p = 0.42$) and receivers ($r(117) = -0.09, p = 0.34$).

4.2.1. Human Sender

Cheap Truth

Figure 8 (top panel) describes the proportion of deceptive messages sent by the sender based on the type of bots encountered in the Cheap Truth condition. First, when the receiver bot followed senders' message most of the time (i.e., the Following bot), we found the same pattern as in Experiment 1: senders lied more on the risky outcome (20,36) where they were motivated to and they lied less on the risky outcome (0,4) where they were not motivated. Interestingly, when the Defecting bot deviated from the followed message, senders sent a deceptive message where they were not motivated instead. In addition, there is also an increased proportion of deceptive messages, suggesting that they understood that their message would not be followed by the receivers during their interaction. Finally, when the Random bot randomly followed the received message, there was no clear difference in senders' proportion of deceptive messages based on the two observed outcomes for the risky option.

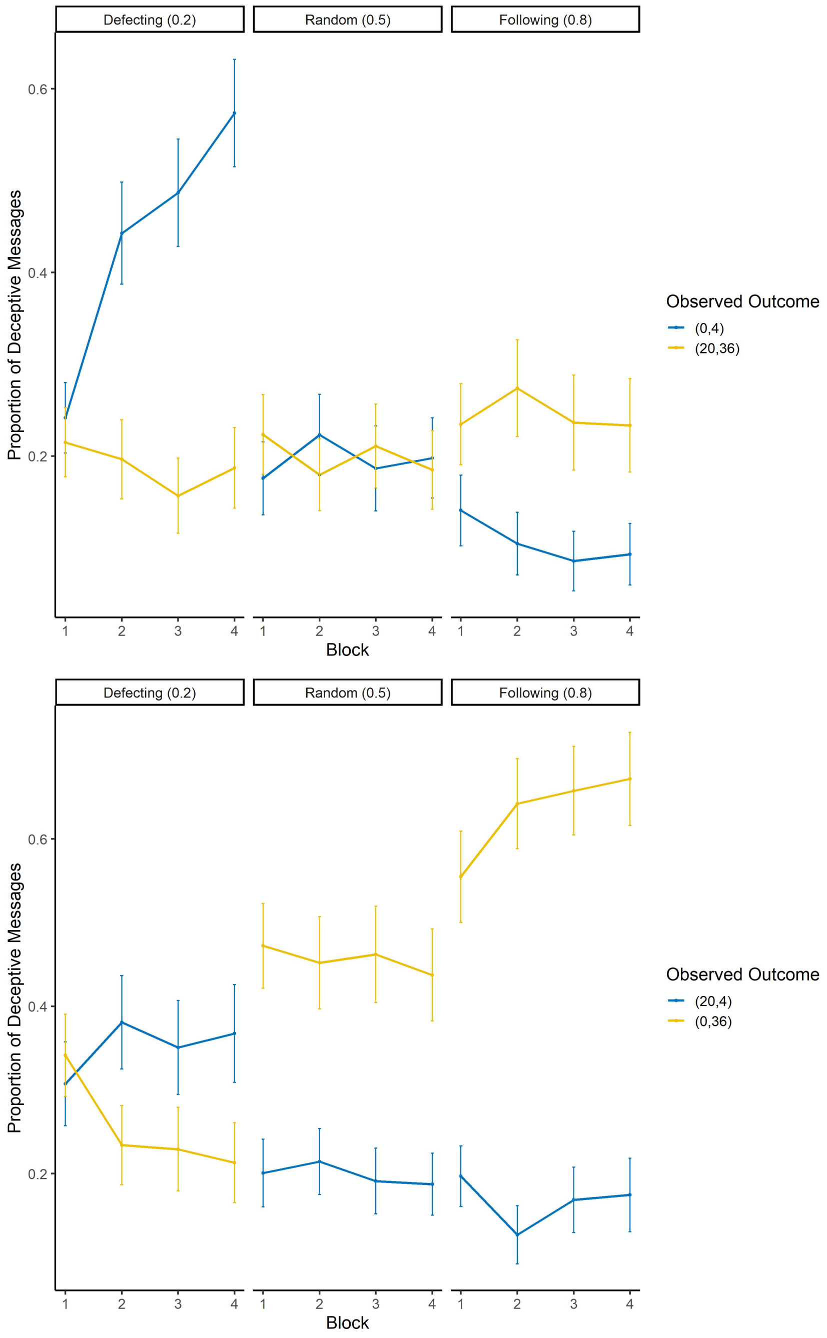


Figure 8. The proportion of lies send across three different type of bots by senders in both Cheap Truth (**top**) and Costly Truth (**bottom**) conditions. The error bars indicate standard error within each group.

The results of the ANOVA analysis indicated that in the Cheap Truth condition, there was a main effect of the block factor ($F(3,156) = 3.17, p = 0.026, \eta_p^2 = 0.058$), senders sent more deceptive messages in the fourth block compared to the first block, $t = 3.14, p = 0.01, d = 0.43$, a main effect of the type of bot interacted with ($F(2,104) = 7.62, p < 0.001, \eta_p^2 = 0.128$)—senders sent more deceptive information when interacted with the Defecting bot. The results also indicated an interaction between the block and the observed result ($F(2.20, 114.37) = 9.01, p < 0.001, \eta_p^2 = 0.15$), an interaction between the block and the bot-type ($F(4.08, 212.25) = 3.50, p = 0.008, \eta_p^2 = 0.06$), and an interaction between the bot-type and the observed result ($F(2,104) = 12.74, p < 0.001, \eta_p^2 = 0.197$). Additionally, there was a three-way interaction between the block, the observed outcome, and the type of bot ($F(4.42, 229.56) = 5.12, p < 0.001, \eta_p^2 = 0.09$). SVO was not significantly related to the proportion of deceptive messages sent by senders, $F(1, 52) = 1.63, p = 0.21$. However, it significantly interacted with the block and the observed outcome ($F(2.20, 114.37) = 3.33, p = 0.04, \eta_p^2 = 0.06$). Post hoc analysis in examining the three-way interaction indicated that the block interacted significantly with SVO when the observed outcome was (0,4), $F(2.27, 117.89) = 4.05, p = 0.02, \eta_p^2 = 0.072$, while such an interaction was not found when the observed outcome was (20,36), $F(2.57, 134.02) = 0.74, p = 0.51, \eta_p^2 = 0.014$.

To further interpret the three-way interaction among the block, observed outcome, and bot type, we performed repeated ANOVA for each type of bot. When the senders were instructed to interact with the Defecting bot, there was a main effect of the block ($F(2.39, 124.10) = 6.65, p < 0.001, \eta_p^2 = 0.113$), since the senders in the first block sent fewer deceptive messages compared to the second ($t = 3.24, p = 0.009, d = 0.44$), the third ($t = 3.32, p = 0.007, d = 0.45$), and the fourth block ($t = 5.40, p < 0.001, d = 0.74$), a main effect of the observed outcome ($F(2.39, 124.10) = 6.65, p < 0.001, \eta_p^2 = 0.113$), as the senders sent more deceptive messages when the observed outcome was (0,4) compared to the observed outcome (20,36), $t = 4.32, p < 0.001, d = 0.59$. In addition, there was a two-way interaction between the block and the observed outcomes: The senders sent more misleading information when the observed risk outcome was (0,4) compared to the other outcome (20,36) in the second ($t = 3.80, p = 0.008$), the third ($t = 5.09, p < 0.001$), and the fourth condition ($t = 5.97, p < 0.001$), while such a difference was not found in the first block. This indicated that the senders understood that their suggestions to persuade the senders to choose the safe option were not followed, and thus adjusted their strategies. When interacting with the Random bot, the results indicated that neither their main effect nor their interaction was significant. When interacting with the Following bot, there was only one main effect of the observed outcome ($F(1, 52) = 7.29, p = 0.009, \eta_p^2 = 0.123$), as senders sent more deceptive messages when the risky option (20,36) only brings the receiver more benefits, $t = 2.42, p = 0.02, d = 0.33$.

Costly Truth

Figure 8 (the bottom panel) describes the proportion of deceptive messages sent by the sender based on the different bot-types encountered in the Costly Truth condition. When interacted with the Following bot and the Random bot, senders lied about the risky outcome (0,36) where they were motivated. As expected, when the Defecting bot deviated from the received suggestions, the senders lied about the risky outcome (20,4) where they were not motivated. This suggests a similar behavior pattern in reacting to different bot types based on observed risky option payoffs as the Cheap Truth condition.

The ANOVA results indicated that there was a main effect of the observed outcome ($F(1, 52) = 13.72, p < 0.001, \eta_p^2 = 0.21$). The senders sent more misleading messages when the observed result (0,36) only gives more rewards to the receiver, $t = 4.46, p < 0.001, d = 0.61$. There was also a main effect of the bot-type ($F(1.73, 89.94) = 5.16, p = 0.01, \eta_p^2 = 0.09$). Senders sent more deceptive messages to the Following bot, compared to the Defecting bot ($t = 3.79, p < 0.001, d = 0.52$) and the Random bot ($t = 2.84, p = 0.02, d = 0.39$). There was only a two-way interaction between the type of bot that interacted and the observed result ($F(1.70, 88.50) = 17.43, p < 0.001, \eta_p^2 = 0.25$). The post hoc analysis of the interaction further indicated that the proportion of deceptive messages was significantly different when the two risky outcomes were observed when they interact with the Random bot ($t = 4.14, p < 0.001$) and Following bot ($t = 7.47, p < 0.001$), while this difference was not significant when they interact with the Defecting bot ($t = 1.56, p = 1.00$). Moreover, SVO was a significant negative predictor of the deceptive message, $t(322) = 3.10, p = 0.002$. Linear regression on senders' deceptive behavior against SVO indicated that for each unit increase in the sender's SVO (higher prosociality), there was a 0.19 decrease in the proportion of deceptive messages sent.

4.2.2. Human Receiver

Cheap Truth

Figure 9 (top panel) describes the proportion of risky choices chosen by receivers based on the bot type in the Cheap Truth condition. In general, receivers selected the risky option when informed that the risky option (20,36) is more profitable. There was also a weak decrease trend for choosing the risky option, when receivers were informed by the Truthful bot that the risky option was less profitable (more trust) and by the Deceptive bot that the risky option was more profitable (less trust). This suggests that receivers are able to adjust their strategies based on the inferred truth status of the received messages. When comparing interactions between different bot types, the difference was not distinct in risky choices after receiving two different types of message.

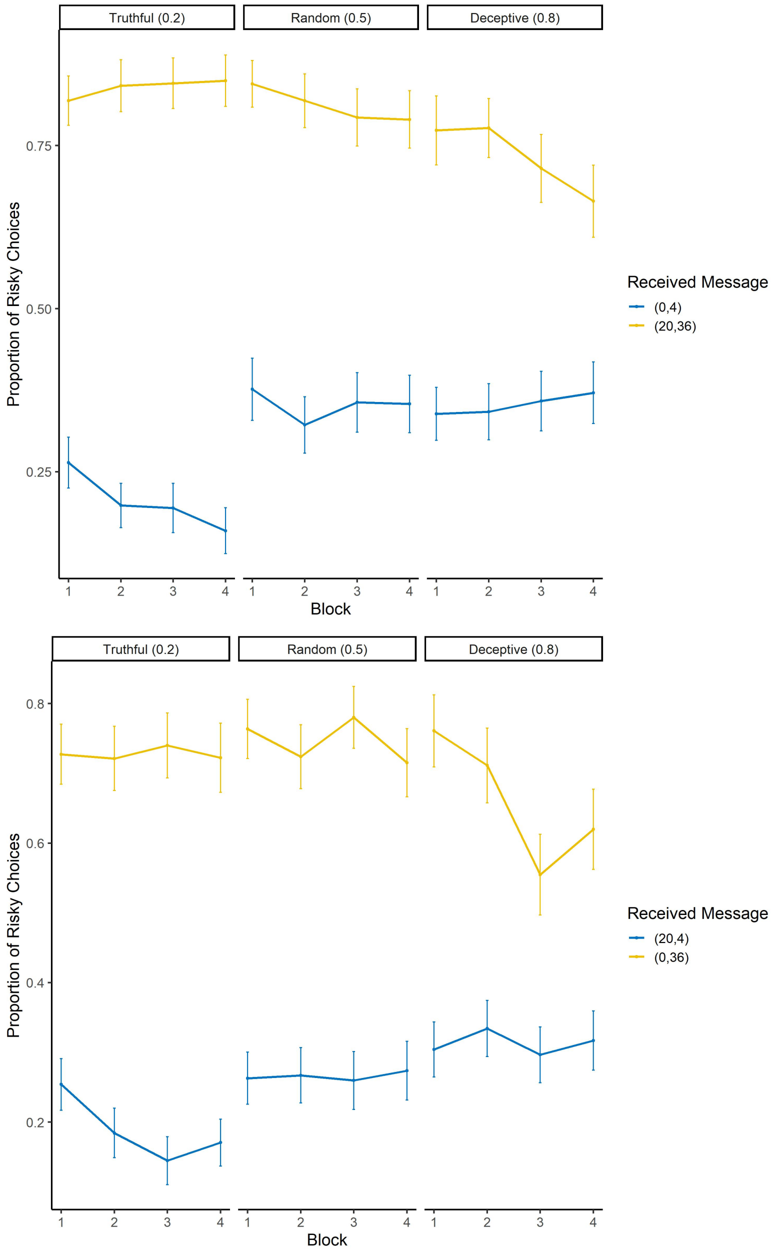


Figure 9. The proportion of choosing risky card across three different types of bots by receivers in both Cheap Truth (**top**) and Costly Truth (**bottom**) conditions. The error bars indicate standard error within each group.

The ANOVA results (similar to our analysis in Experiment 1, we also exclude participants who only received one type of information. This leaves 40 participants for the Cheap Truth condition and 42 participants for the Costly Truth condition in the repeated ANOVA analysis) indicated that there was a main effect of the block ($F(3, 114) = 5.07, p = 0.002, \eta_p^2 = 0.12$), receivers in the first block selected more risky choices compared to the fourth block, $t = 3.23, p = 0.007, d = 0.53$; a main effect of the received message ($F(1, 38) = 34.00, p = 0.001, \eta_p^2 = 0.47$), receivers selected a more risky option when told that it is more profitable, $t = 11.46, p < 0.001, d = 1.81$. There was also a two-way interaction between the message received and the type of bots that interacted, $F(2, 76) = 3.33, p = 0.04, \eta_p^2 = 0.08$. Post hoc analysis indicated that when informed that the risky option (0,4) was less profitable, receivers selected the less risky option when interacting with the Truthful bot compared to its interaction with the Deceptive bot ($t = 3.91, p = 0.002, d = 0.40$), while there was no significant difference in receiving the other message (20,36) between different types of bots.

Costly Truth

Figure 9 (bottom panel) describes the proportion of risky choices chosen by receivers based on the type of bots that interact in the Costly Truth condition. Similarly to the Cheap Truth condition, there was only a weak trend that, as interacting with the Deceptive bot, receivers chose the less risky option when being informed by the Deceptive bot that risky card was more profitable; receivers chose the less risky option when being informed by the Truthful bot that risky card was less profitable. The results of the ANOVA indicated that there was only a main effect of the received message ($F(1, 40) = 21.48, p < 0.001, \eta_p^2 = 0.35$), the receivers selected the more risky option when informed that it was more profitable (0,36), $t = 9.85, d = 1.52, p < 0.001$. Furthermore, there was an interaction between the message received and the type of bots that interacted, $F(1.74, 69.57) = 3.40, p = 0.045, \eta_p^2 = 0.08$. Post hoc analysis again indicated that when informed that the risky option (20,4) was less profitable, the receivers selected a less risky option when they interacted with the Truthful bot compared to its interaction with the Deceptive bot ($t = 3.00, p = 0.047, d = 0.33$), while there were no significant differences in the choices when informed that the risky option (0,36) was more profitable in different types of bots.

4.3. Discussion

Experiment 2 extends the findings of Experiment 1 by instructing players to interact with different computer algorithms that performed predefined probabilistic actions. As expected, we found evidence that the senders acted according to their expectations of the receiver behavior. The deceptive behavior of the senders depended on whether the message would be followed by the receivers. The deceptive motivation of the senders was also influenced by their knowledge of the incentive context (e.g., [20]). Importantly, in the Cheap Truth condition, there was a learning effect, as senders sent more deceptive information on the risky outcome (0,4) when they interacted with the Defecting bot. Furthermore, in the Cheap Truth condition, senders sent more deceptive messages to the Defecting bot where lies were not expected and to the Following bot where lies were expected; In the Costly Truth condition, senders sent considerably more deceptive information to the Following bot and Random bot where lies were expected.

When participants in the receiver role interacted with different sender bots, our results indicated that receivers were willing to accept the suggestion. They selected the risky option more often when informed by the sender bot that the outcome of the risky choice was (20,36) or (0,36) and less often when the risky choice was (0,4) or (20,4). In both the Cheap Truth and Costly Truth conditions, the frequency of deceptive messages affects the receiver's choice: They chose the safe option more often (thus, benefitting the sender) when they were informed of a less profitable risky outcome by the Truthful bot compared to the Deceptive bot. Interestingly, receivers in the first block also selected more risky options compared to the fourth block in the Cheap Truth condition, which needs future research explorations.

SVO was a reliable predictor of the frequency with which a sender sent deceptive messages (explicitly in the Costly Truth condition), but not of the frequency with which a receiver chose the risky option. In Experiment 1, we raised two possibilities for prosocial receivers choosing safe options more often: (1) risk aversion; (2) try to benefit the other player. Our results in Experiment 2 ruled out the first possibility, as their risk preference was not related to their social preference. Furthermore, this also suggests that prosocial receivers were less likely to benefit computer bots compared to their interaction with other online players.

5. General Discussion

In the current research, we explore players' strategies in a repeated sender–receiver game that allows us to explore the effects of economic incentives and social preferences related to deceptive behavior. In Experiment 1, we employ two different misaligned incentive conditions for paired online players. Consistent with previous literature, we find that senders are willing to send a deceptive message when telling the truth is costly and receivers are willing to accept the suggestions received. Our results also indicate that receivers take advantage of their knowledge about economic incentives to infer the senders' intention, they deviate from the senders' suggestion when they know that the truth is costly for senders. Importantly, in Experiment 1, we find that social preferences are related to both players' behaviors. Prosocial senders send less deceptive messages, whereas prosocial receivers choose less risky choices.

Experiment 2 extends these findings by further examining how the players' choices react to the other players' strategies. Instructing players to play with a computer algorithm that displays consistent behavior patterns, we find that the players' interactive strategies are affected by both the paired partners' behavior and the incentive conditions. When truth telling is costly in Costly Truth, senders send economically expected deceptive messages when the observed risky outcome is (0,36) with the prediction that their messages will be followed. In the Cheap Truth condition, senders also send deceptive messages with the prediction that their messages will not be followed where the truthful messages are expected. When players are in the role of receivers, under both incentive conditions, they "trust" the Truthful bot by choosing the less risky option when informed that the risky option is less profitable compared to their less trusted interaction with the Deceptive bot. Furthermore, social preferences are not related to the receiver's choices of risky options.

The observations of Experiments 1 and 2 on social preferences together suggest that prosocial receivers appear to benefit other players when interacting with human players (Experiment 1) but not when interacting with computer bots (Experiment 2). Given that in Experiment 2, computer algorithms display predefined consistent behaviors, we can thus infer that in Experiment 1, senders who play the sender–receiver online are willing to believe that their messages would be followed [3] and receivers tend to believe that they interact with truthful senders (especially in the Cheap Truth condition).

By introducing repeated measures, we are able to test the behavior change over a series of interactive plays. Our results indicate that players display consistent behavioral patterns in the online sender–receiver game. Additionally, senders are not less likely to lie after their deceptive messages are detected. While this may suggest the inability to infer others' behaviors and to act against it, our Experiment 2 suggests that players do interpret partners' behavior as they interact with computer bots with different strategies. This contributes to the extension of previous research on sophisticated truth-telling behavior [7], which is also consistent with the theory of mind reported in many economic games (e.g., [36]). In Experiment 2, we only observed the behavior change (the block effect) in the Cheap Truth condition. This implies that players adopt different strategies in playing the sender–receiver games. As there is emerging evidence suggesting that players engage a limited ability in playing games that need recursive thinking (e.g., [37,38]), future research can further test whether players' ToM ability is influenced by incentive conditions and for those who have better knowledge of using ToM ability, they would perform better in the sender–receiver game.

Furthermore, our current study establishes the relationship between social preference and behaviors in the sender–receiver game by implementing the measure of prosociality. Our findings suggest that prosociality of players rather than monetary rewards drive the senders' deceptive behavior (e.g., [10]). That is, the non-prosocial sender would decide to lie regardless of monetary incentives. We suspect that this observed phenomenon is also related to the on-line platform employed in playing the sender–receiver game, as senders do not decrease their lying propensity after being detected by receivers even though they are able to infer receivers' choices. Future research can further explore whether there is a difference in senders' deceptive behavior between the online and offline sender–receiver game. In addition, our results also indicate that prosocial receivers take into account the reward of the sender when making choices, choosing the option that will bring more value to the sender when interacting with human players. The difference in their performance implies that players behave differently in the human–human interactions (Experiment 1) compared to the human–bot interactions (Experiment 2). In Experiment 2, we predefine the behavior of computer algorithms to examine the behavior of human players in the deceptive interaction. In the future, we hope to use Instance-Based Learning Theory [39,40] to construct bots that are more dynamic and human-like. These bots can also rely on memories about partners' actions to make more accurate predictions about the player's ToM [41].

To conclude, our current research explores human behavior in a repeated sender–receiver deceptive game when paired with other online players and against computer algorithms that display probabilistic behaviors. We find that players display consistent behavior in reacting to misaligned incentives when playing with other online players. In addition, when truth-telling is costly, senders send more deceptive messages, and receivers deviate more from received suggestions. The interaction with computer algorithms indicates that players' behavior is also influenced by their paired partners' strategies. Senders adaptively send deceptive or truthful messages based on their expectation of whether their messages would be followed, and receivers selectively follow the suggestions by choosing the option that benefits the senders when interacting with the truthful senders. In addition, our work also establishes the association with behavior in the game: Prosocial senders send less deceptive messages, and prosocial receivers choose options that benefit senders more when interacting with human sender players.

Author Contributions: Conceptualization, F.M., C.G., P.A. and H.Z.; methodology, F.M., C.G., H.Z.; software, F.M. and P.A.; validation, F.M., C.G., P.A. and H.Z.; formal analysis, H.Z.; investigation, F.M. and C.G.; resources, C.G.; data curation, F.M. and H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, F.M., C.G., P.A. and H.Z.; visualization, H.Z.; supervision, C.G.; project administration, C.G.; funding acquisition, C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was sponsored by the Army Research Office and accomplished under Grant Number W911NF-17-1-0370 to Cleotilde Gonzalez.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of Carnegie Mellon University.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data and more details about the experiments are available at https://osf.io/fmbv7/ (accessed on July 27, 2022).

Acknowledgments: The authors thank Dr. Michael Yu for some initial analyses performed on this data. We also acknowledge the help of research assistants at the Dynamic Decision Making Laboratory in programming the deception game and supporting data collection for this study.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Appendix A.1. Theoretical Analysis

Let us construct the payoff matrix associated with the deception game in Table A1 given every strategy profile and the exogenous probability p (note that every occurrence of $px_r + (1-p)y_r$ was replaced with z_r for convenience, consistent with the original assumption).

Table A1. Payoff matrix.

	$(\boldsymbol{A^{M1}},\boldsymbol{A^{M2}})$	$(\boldsymbol{A^{M1}},\boldsymbol{B^{M2}})$	$(\boldsymbol{B^{M1}},\boldsymbol{A^{M2}})$	$(\boldsymbol{B^{M1}},\boldsymbol{B^{M2}})$
$(M1^{S1},M1^{S2})$	$px_s + (1-p)y_s$ z_r	$px_s + (1-p)y_s$ z_r	z_s z_r	z_s z_r
$(M1^{S1},M2^{S2})$	$px_s + (1-p)y_s$ z_r	$px_s + (1-p)z_s$ $px_r + (1-p)z_r$	$pz_s + (1-p)y_s$ $pz_r + (1-p)y_r$	z_s z_r
$(M2^{S1},M1^{S2})$	$px_s + (1-p)y_s$ z_r	$pz_s + (1-p)y_s$ $pz_r + (1-p)y_r$	$px_s + (1-p)z_s$ $px_r + (1-p)z_r$	z_s z_r
$(M2^{S1},M2^{S2})$	$px_s + (1-p)y_s$ z_r	z_s z_r	$px_s + (1-p)y_s$ z_r	z_s z_r

Given the initial assumptions $y_r < z_r < x_r$ and $x_s, y_s < z_s$, it follows directly from Table A1 that the only Nash equilibria in pure strategies are uninformative and correspond to the following strategy profiles: $(M1^{S1},M1^{S2};A^{M1},A^{M2})$, $(M1^{S1},M1^{S2};B^{M1},A^{M2})$, $(M1^{S1},M1^{S2};B^{M1},B^{M2})$, $(M2^{S1},M2^{S2};A^{M1},A^{M2})$, $(M2^{S1},M2^{S2};A^{M1},B^{M2})$, $(M2^{S1},M2^{S2};B^{M1},B^{M2})$

Appendix A.2. SVO Consistency

The consistency of the preferences revealed across the 6 primary SVO sliders used in the experiments is verified using the method proposed by [30]. More precisely, we check for transitivity in a subject's preferences by first categorizing the selected outcomes as altruistic, prosocial, individualist, or competitive for each of the 6 items of the measure. We do so by computing, for a particular item, the similarity (Euclidean distance) between the subject's selected answer and each of the two most extreme outcomes, which are associated with different types. The shortest distance will then determine the subject's preference for one type of outcome over another. For example, in the first item (as defined in [30]), (85, 85) is considered a prosocial outcome, while (85, 15) is considered a competitive outcome. If the subject's choice is more similar to (85, 85) than (85, 15), then it implies that she prefers prosocial outcomes to competitive ones. The specific outcome types associated with each 6 primary items are as follows:

- Item 1: prosocial vs. competitive
- Item 2: competitive vs. individualist
- Item 3: altruist vs. prosocial
- Item 4: altruist vs. competitive
- Item 5: individualist vs. altruist
- Item 6: individualist vs. prosocial

Once the subject's types of preference are identified for each choice, we simply check for any breakdown of transitivity in preferences. For example, if the subject prefers an altruistic outcome to a competitive one and an individualistic outcome to an altruistic one, then she cannot prefer a competitive outcome to an individualistic one. If any such inconsistency is detected, the data associated with the corresponding pair (not only the subject) are excluded from our analyses.

Appendix A.3. Distribution of Participants' Social Preference

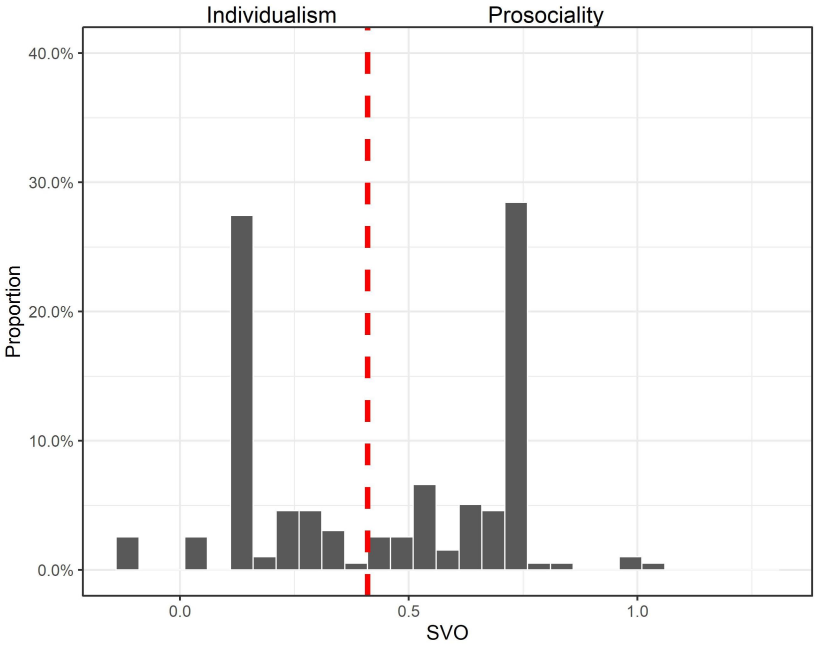


Figure A1. Distribution of participants' SVO in Experiment 1.

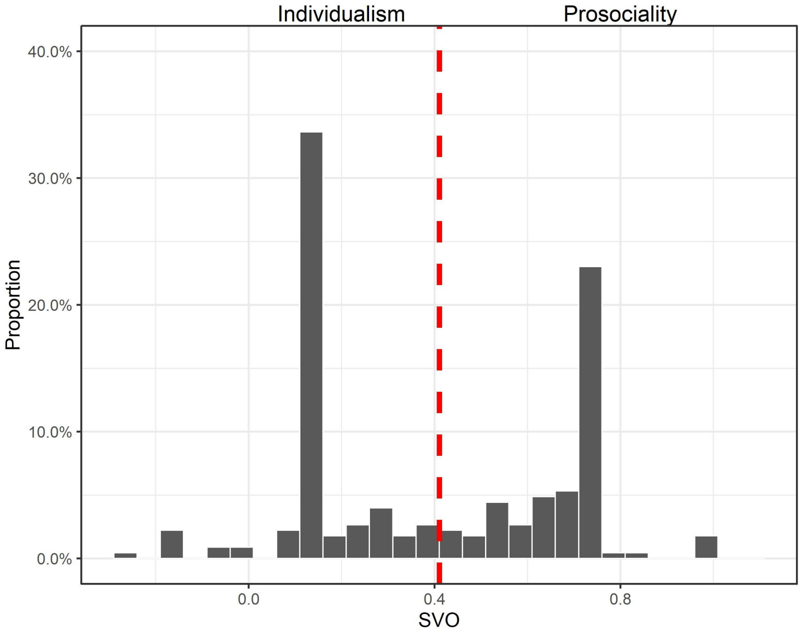


Figure A2. Distribution of participants' SVO in Experiment 2.

References

1. Rowe, N.C.; Rrushi, J. *Introduction to Cyberdeception*; Springer: Berlin/Heidelberg, Germany, 2016.
2. Gerlach, P.; Teodorescu, K.; Hertwig, R. The truth about lies: A meta-analysis on dishonest behavior. *Psychol. Bull.* **2019**, *145*, 1–44.
3. Gneezy, U. Deception: The role of consequences. *Am. Econ. Rev.* **2005**, *95*, 384–394.
4. Mazar, N.; Amir, O.; Ariely, D. The dishonesty of honest people: A theory of self-concept maintenance. *J. Mark. Res.* **2008**, *45*, 633–644.
5. Fischbacher, U.; Föllmi-Heusi, F. Lies in disguise—An experimental study on cheating. *J. Eur. Econ. Assoc.* **2013**, *11*, 525–547.
6. Kajackaite, A.; Gneezy, U. Incentives and cheating. *Games Econ. Behav.* **2017**, *102*, 433–444.
7. Sutter, M. Deception through telling the truth?! Experimental evidence from individuals and teams. *Econ. J.* **2009**, *119*, 47–60.

8. Gonzalez, C. Learning and dynamic decision making. *Top. Cogn. Sci.* **2022**, *14*, 14–30.
9. Gino, F.; Ayal, S.; Ariely, D. Self-serving altruism? The lure of unethical actions that benefit others. *J. Econ. Behav. Organ.* **2013**, *93*, 285–292.
10. Erat, S.; Gneezy, U. White lies. *Manag. Sci.* **2012**, *58*, 723–733.
11. Cappelen, A.W.; Sørensen, E.Ø.; Tungodden, B. When do we lie? *J. Econ. Behav. Organ.* **2013**, *93*, 258–265.
12. Biziou-van Pol, L.; Haenen, J.; Novaro, A.; Occhipinti Liberman, A.; Capraro, V. Does telling white lies signal pro-social preferences? *Judgm. Decis. Mak.* **2015**, *10*, 538–548.
13. Utikal, V.; Fischbacher, U. Disadvantageous lies in individual decisions. *J. Econ. Behav. Organ.* **2013**, *85*, 108–111.
14. Abeler, J.; Becker, A.; Falk, A. Representative evidence on lying costs. *J. Public Econ.* **2014**, *113*, 96–104.
15. Gneezy, U.; Kajackaite, A.; Sobel, J. Lying aversion and the size of the lie. *Am. Econ. Rev.* **2018**, *108*, 419–53.
16. Hurkens, S.; Kartik, N. Would I lie to you? On social preferences and lying aversion. *Exp. Econ.* **2009**, *12*, 180–192.
17. Kerschbamer, R.; Neururer, D.; Gruber, A. Do altruists lie less? *J. Econ. Behav. Organ.* **2019**, *157*, 560–579.
18. Maggian, V.; Villeval, M.C. Social preferences and lying aversion in children. *Exp. Econ.* **2016**, *19*, 663–685.
19. Sheremeta, R.M.; Shields, T.W. Do liars believe? Beliefs and other-regarding preferences in sender–receiver games. *J. Econ. Behav. Organ.* **2013**, *94*, 268–277.
20. Rode, J. Truth and trust in communication: Experiments on the effect of a competitive context. *Games Econ. Behav.* **2010**, *68*, 325–338.
21. Gneezy, U.; Rockenbach, B.; Serra-Garcia, M. Measuring lying aversion. *J. Econ. Behav. Organ.* **2013**, *93*, 293–300.
22. Brandts, J.; Charness, G. Truth or consequences: An experiment. *Manag. Sci.* **2003**, *49*, 116–130.
23. Sánchez-Pagés, S.; Vorsatz, M. An experimental study of truth-telling in a sender–receiver game. *Games Econ. Behav.* **2007**, *61*, 86–112.
24. Sánchez-Pagés, S.; Vorsatz, M. Enjoy the silence: an experiment on truth-telling. *Exp. Econ.* **2009**, *12*, 220–241.
25. Murphy, R.O.; Ackermann, K.A. Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personal. Soc. Psychol. Rev.* **2014**, *18*, 13–41.
26. Moisan, F.; ten Brincke, R.; Murphy, R.O.; Gonzalez, C. Not all Prisoners' Dilemma games are equal: Incentives, social preferences, and cooperation. *Decision* **2018**, *5*, 306–322.
27. Ponzano, F.; Ottone, S. Prosociality and fiscal honesty: Tax evasion in Italy, United Kingdom, and Sweden. In *Dishonesty in Behavioral Economics*; Bucciol, A., Montinari, N., Eds.; Academic Press: Cambridge, MA, USA, 2019.
28. Fleiß, J.; Ackermann, K.A.; Fleiß, E.; Murphy, R.O.; Posch, A. Social and environmental preferences: measuring how people make tradeoffs among themselves, others, and collective goods. *Cent. Eur. J. Oper. Res.* **2020**, *28*, 1049–1067.
29. Murphy, R.O.; Ackermann, K.A. Social preferences, positive expectations, and trust based cooperation. *J. Math. Psychol.* **2015**, *67*, 45–50.
30. Murphy, R.O.; Ackermann, K.A.; Handgraaf, M. Measuring social value orientation. *Judgm. Decis. Mak.* **2011**, *6*, 771–781.
31. Hedden, T.; Zhang, J. What do you think I think you think?: Strategic reasoning in matrix games. *Cognition* **2002**, *85*, 1–36.
32. Van Lange, P.A. The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *J. Personal. Soc. Psychol.* **1999**, *77*, 337–349.
33. JASP Team. JASP (Version 0.14.1)[Computer Software]. 2020. https://jasp-stats.org/
34. Williams, P.; Heathcote, A.; Nesbitt, K.; Eidels, A. Post-error recklessness and the hot hand. *Judgm. Decis. Mak.* **2016**, *11*, 174–184.
35. Charness, G.; Gneezy, U.; Imas, A. Experimental methods: Eliciting risk preferences. *J. Econ. Behav. Organ.* **2013**, *87*, 43–51.
36. Camerer, C.F.; Ho, T.H.; Chong, J.K. A cognitive hierarchy model of games. *Q. J. Econ.* **2004**, *119*, 861–898.
37. Zhang, H.; Moisan, F.; Gonzalez, C. Rock-Paper-Scissors Play: Beyond the Win-Stay/Lose-Change Strategy. *Games* **2021**, *12*, 52.
38. Brockbank, E.; Vul, E. Recursive Adversarial Reasoning in the Rock, Paper, Scissors Game. In Proceedings of the Proceedings of the Annual Meeting of the Cognitive Science Society, 2020. Online Virtual Meeting. https://cognitivesciencesociety.org/cogsci20/
39. Gonzalez, C.; Lerch, J.F.; Lebiere, C. Instance-based learning in dynamic decision making. *Cogn. Sci.* **2003**, *27*, 591–635.
40. Gonzalez, C.; Dutt, V. Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychol. Rev.* **2011**, *118*, 523–551.
41. Nguyen, T.N.; Gonzalez, C. Theory of mind from observation in cognitive models and humans. *Top. Cogn. Sci.* **2021**, https://doi.org/10.1111/tops.12553.