Experimental Economics (2019) 22:980–1001
https://doi.org/10.1007/s10683-018-09599-3

ORIGINAL PAPER

Deception and reciprocity

Despoina Alempaki[1] · **Gönül Doğan**[2] · **Silvia Saccardo**[3]

Received: 26 September 2016 / Revised: 23 November 2018 / Accepted: 27 November 2018 /
Published online: 6 December 2018

Abstract

We experimentally investigate the relationship between (un)kind actions and subsequent deception in a two-player, two-stage game. The first stage involves a dictator game. In the second-stage, the recipient in the dictator game has the opportunity to lie to her counterpart. We study how the fairness of dictator-game outcomes affects subsequent lying decisions where lying hurts one's counterpart. In doing so, we examine whether the moral cost of lying varies when retaliating against unkind actions is financially beneficial for the self (selfish lies), as opposed to being costly (spiteful lies). We find evidence that individuals engage in deception to reciprocate unkind behavior: The smaller the payoff received in the first stage, the higher the lying rate. Intention-based reciprocity largely drives behavior, as individuals use deception to punish unkind behavior and truth-telling to reward kind behavior. For selfish lies, individuals have a moral cost of lying. However, for spiteful lies, we find no evidence for such costs. Taken together, our data show a moral cost of lying that is not fixed but instead context-dependent.

Keywords Deception · Lying costs · Reciprocity · Punishment · Laboratory experiment

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10683-018-09599-3) contains supplementary material, which is available to authorized users.

✉ Despoina Alempaki
despoina.alempaki@wbs.ac.uk

Gönül Doğan
dogan@wiso.uni-koeln.de; gonul2517@gmail.com

Silvia Saccardo
silviasaccardo@gmail.com

[1] Warwick Business School, University of Warwick, Coventry CV4 7AL, United Kingdom

[2] Faculty of Management, Economics and Social Sciences, University of Cologne, Universitätsstraße 22a, 50923 Cologne, Germany

[3] Department of Social and Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue, BP 208, Pittsburgh, PA 15213, USA

JEL Classification C90 · C92 · D0 · D82

1 Introduction

Work environment substantially affects employee satisfaction and behavior. Numerous examples demonstrate that low worker satisfaction decreases productivity. Consider labor disputes: Krueger and Mas (2004) showed that the 1994–1996 conflict between a labor union and the management of Bridgestone/Firestone was a major contributing factor to the increase in the number of defective tires produced by the company in that period. Similarly, labor conflicts at the construction equipment manufacturer Caterpillar resulted in a significant decrease in output quality (Mas 2008), and compensation disputes in police departments in New Jersey coincided with a decline in the number of arrests by police officers and an increase in crime rates (Mas 2006).

Changes in employee behavior as a reaction to the workplace environment might stem from different sources. Workers can alter not only the quality and quantity of effort provision, but also their engagement in unethical behavior. Unethical behavior is widespread in the corporate world and takes different forms, negatively affecting companies' productivity (see Treviño et al. 2014, for a review). Reciprocal motives are an important driver of employees' decision to behave unethically.[1] Unethical behavior, or the lack thereof, might serve as a reciprocity device any time a direct reaction (e.g., effort reduction) is not feasible, for example, due to lack of power or fear of retaliation. In many situations, workers might reciprocate negative encounters with deceptive behaviors that hurt the company, such as misreporting the quality of their work or faking sick days. Similarly, after positive experiences within the company, employees might abstain from carrying out unethical acts they would have otherwise engaged in; that is, they might reciprocate with honesty that benefits the company at a cost to themselves. Because deception, by its nature, is difficult to observe, empirical work offers limited guidance in investigating whether and to what extent (a lack of) deception is used as a (reward) punishment mechanism. The vast experimental work in economics on gift exchange has, thus far, mainly explored reciprocity within the domain of effort provision, looking at behaviors that do not involve deception (see e.g., the survey on reciprocal behavior by Fehr and Gächter 2000).

In this paper, we use a simple two-player, two-stage game to experimentally study whether deception serves as a reciprocity device. In the first stage, subjects play a variation of the dictator game. In the second stage, they play a deception game (Erat and Gneezy 2012), in which the recipient in the dictator game has the opportunity to lie to her counterpart by sending her an untruthful message. Compared to truth-telling, lying reduces the payoff of one's counterpart if the counterpart follows the

[1] Previous research has investigated other drivers of unethical conduct in the workplace, such as incentive schemes (Belot and Schröder 2013; Danilov et al. 2013), relative wages (John et al. 2014), and monitoring (Belot and Schröder 2013; Gino et al. 2013).

message. We use the strategy method (Selten 1967) and have the sender choose a message for each possible payoff received in the first stage.

Predictions concerning whether and to what extent deception could be used as a reciprocity device are conflicting. Reciprocity models (e.g., Dufwenberg and Kirchsteiger 2004; Rabin 1993) predict that previous encounters will affect subsequent behavior (see also Charness 2004; Charness and Haruvy 2002; Charness and Levine 2007; Offerman 2002). However, it is unclear whether previous encounters would affect subsequent behavior when reciprocity occurs via lying. A fast-growing experimental literature has documented that deception entails positive moral costs (Bucciol and Piovesan 2011; Fischbacher and Föllmi-Heusi 2013; Gneezy 2005; Mazar et al. 2008).[2] This literature predicts that, because lying is morally costly, less reciprocal behavior could occur when reciprocating requires individuals to lie than when it does not, regardless of the nature of a previous encounter. To test whether this prediction is true, we compare the results of our baseline game with the results of a game in which individuals can reciprocate without engaging in deception. That is, we modify the second stage of the baseline game such that an identical payoff allocation can be obtained through a *direct choice* rather than by telling a lie. If there were indeed a moral cost of lying, we would expect lying rates to be lower than direct choice rates.

This design allows us to test Hurkens and Kartik's (2009) conjecture that individuals have either a zero or infinite moral cost of lying—the former types always lie if they prefer the outcome from lying and the latter types never lie. In such a case, the conditional probability of lying (the ratio of lying to direct choice) would be constant regardless of the type of previous encounter.[3] On the other hand, moral costs might be malleable. The morality literature has suggested that moral costs decrease when one can blame the other party (Bandura et al. 1996; Bandura 1999, 2002, 2004; Cubitt et al. 2011; Ditto et al. 2009). In that case, we would expect that the more unfair the previous interaction, the lower the moral cost of lying. In other words, deceiving someone who was previously unkind would be easier than deceiving someone who was previously kind. Additionally, kind encounters could be rewarded with honesty from people who would otherwise lie. Thus, the conditional probability of lying would decrease with increasing kindness.

Further, moral costs of lying might be sensitive to whether individuals have monetary incentives to lie. There are conflicting predictions. On the one hand, the moral cost of lying may increase with one's payoff from lying. When reciprocating unkind behavior requires individuals to lie and lying leads to a financial benefit, people might worry about the signal they convey to themselves or to others about their own morality, as it can be unclear whether individuals lie to gain

[2] See also Battigalli et al. (2013), Charness and Dufwenberg (2006, 2010), Dreber and Johannesson (2008), Erat and Gneezy (2012), Greenberg et al. (2015), Houser et al. (2012), López-Pérez and Spiegelman (2013), Lundquist et al. (2009) and Sutter (2009).

[3] Under the assumption that anyone who would lie for personal benefit would also choose the selfish option when it is directly available, a decrease in the conditional probability of lying indicates that the cost of lying increases for at least some participants. We discuss our findings in detail in the results section.

a personal material benefit or to punish preceding unkind behavior. This concern may be present when lying is beneficial, but it cannot be present when lying is costly in terms of payoffs. Indeed, work on perceptions of generosity in psychology has found that people morally discredit prosocial actions if those actions result in monetary benefits for the self but not if they result in a monetary loss or no benefit for the self (e.g., Carlson and Zaki 2018; Lin-Healy and Small 2012; Newman and Cain 2014). If the moral cost of lying increases with one's own payoff from lying—when lying is used as a tool to reciprocate unkind behavior—the conditional probability of lying would be smaller if lying were beneficial to the self than if it were costly. On the other hand, work on excuses to behave selfishly (e.g., Babcock et al. 1995; Dana et al. 2007; Exley 2015; Gneezy et al. 2017; Konow 2000) suggests that when lying is beneficial, self-serving biases might decrease one's moral cost of lying. When lying is costly for the self, however, a self-serving bias cannot affect the size of moral costs. Hence, this mechanism would predict that the conditional probability of lying is higher when lying is beneficial to the self than when it is costly.

To investigate whether lying costs are malleable in the presence of reciprocal motives as well as different incentives for lying, we study two domains of lies: In the *Selfish* domain, lying increases one's own payoff at a cost to the receiver, and in the *Spiteful* domain, lying is costly for the self but imposes a larger cost on the receiver (see the taxonomy of lies by Erat and Gneezy 2012). In organizations, selfish lies would be akin to over-reporting the quality of one's work or taking credit for someone else's work. An illustration of spiteful lies is misrepresenting one's private information in order to sabotage a project that has a small value to oneself and a high value to one's counterpart.

We also measure the extent to which lying rates change in response to increasingly kinder previous interactions, and whether the effect is driven by intentions rather than initial payoff allocations. The issue is quite elusive, because even in settings where reciprocity doesn't involve deception, evidence is mixed. Some studies (Blount 1995; Falk et al. 2008) point to intentions, while others (Bolton et al. 1998; Charness 2004) highlight distributional concerns as the main driver of reciprocity. In our baseline game, a response to both intentions and initial payoff allocations—including inequity aversion and an income effect—would predict lying and direct-choice rates of the same option to decrease with increasing payoffs from the first stage. To identify whether intentions drive the reciprocal response, we compare our baseline treatment with a treatment in which the experimenter rather than one's counterpart implements the first-stage outcome.

We find that individuals use lies in order to reciprocate (un)kind behavior in both the *Selfish* and *Spiteful* domains: lying rates are highest after an unkind encounter and decrease with increasing kindness of the encounter. For selfish lies, the proportion of individuals who lie is always lower than the proportion of individuals who directly choose the selfish payoff allocation, and the conditional probability of lying decreases as a function of the counterpart's kindness. This finding is in line with the predictions of a positive moral cost of lying that increases with increasing kindness, refuting the hypothesis of Hurkens and Kartik (2009). Surprisingly, for spiteful lies, lying rates are similar to the percentage of direct choices. This result provides

evidence against a moral cost of lying when lying is costly for the self and is used as a punishment device, demonstrating the malleability of the moral cost of lying.

For selfish lies, both intentions and initial payoff allocations affect lying rates, but the latter effect is larger. We find evidence for positive reciprocity, whereas the evidence for negative reciprocity is only directional and not statistically significant. For spiteful lies, only intentions matter. When the experimenter implements the first-stage payoffs, lying rates are close to zero in all cases; however, individuals punish unkind intentions by lying. As expected, lying rates are close to zero after kind encounters.

Our paper adds to the growing literature on deception by investigating the drivers of deception within an *interacting pair* after an initial encounter.[4] Previous work by Ellingsen et al. (2009) showed that honest revelation of private information in bilateral bargaining is affected by whether individuals previously cooperated or defected in a prisoner's dilemma game. Several features distinguish our experiment from theirs. First, we systematically vary the levels of kindness of the initial encounter to show whether the moral cost of lying affects retaliation via deception. We provide the first evidence of reciprocal deception that can be attributed to intentions. More importantly, we compare whether reciprocating through lying is conceptually different from reciprocity that doesn't involve deception, and whether the presence (absence) of financial costs from deception affects the retaliation behavior of the deceiver. Finally, we contribute to the recent stream of papers that investigate the structure of lying costs (e.g., Abeler et al. forthcoming; Dufwenberg and Dufwenberg 2018; Gneezy et al. 2018; Khalmetski and Sliwka 2017). Our results show that there is an interaction between reciprocity concerns, lying costs, and incentives from lying, and highlight the need to combine insights from both the reciprocity and the lying literature in order to fully understand the drivers of deception.

2 Experimental design and procedures

2.1 Experimental design

In the experiment, two subjects (Player A and Player B) are matched for two consecutive stages. This information is common knowledge. In the first stage, Player A (dictator) is given an endowment of 10 tokens and has to decide how much to send to Player B (receiver). Player B does not make any decision in the first stage. Player A can choose between sending 0, 2, 4, 6, 8, or 10 tokens to Player B. Player A keeps the amount not sent to Player B. The amount sent is tripled. We chose to triple the

[4] Houser et al. (2012) investigated cheating behavior toward the experimenter following unfair treatment by another participant. In a similar vein, Kajackaite and Gneezy (2017) investigated cheating behavior toward the experimenter following unfair treatment by the experimenter. Whereas Houser et al. (2012) found cheating increased after an unkind encounter, Kajackaite and Gneezy (2017) found no effect. A separate strand of papers has investigated deception when the interacting partner is a potential accomplice in deception rather than a partner exposed to kind/unkind behaviour (Barr and Michailidou 2017; Behnk et al. 2017; Kocher et al. 2017; Weisel and Shalvi 2015).

Table 1 Payoffs in the *Selfish* and *Spiteful* domains

	Selfish	*Spiteful*
Option 1	10, 10	10, 10
Option 2	4, 12	4, 9

amount sent because doing so makes being kind to Player B easier for Player A, compared to a conventional dictator game. In this setup, we consider an interaction as being increasingly kind the higher the amounts sent in the first stage. Therefore, sending nothing is clearly unkind, and the kindness of the interaction increases with the amount sent.

In the second stage of the experiment, participants play the deception game (Gneezy 2005; Erat and Gneezy 2012). The experimenter rolls a six-sided die in front of each Player B. Afterwards, Player B is asked to send one of six possible messages to Player A. The six messages are "The outcome of the roll of die was x," where x can be any number from the set {1, 2, 3, 4, 5, 6}. After receiving Player B's message, Player A declares what she believes to be the outcome of the die roll. Player A's choice and the actual die outcome determine the payoffs for both players. There are two payment options. Option 1 yields equal payments to both participants and is the same across domains, whereas Option 2 varies between the *Selfish* and *Spiteful* domains as described below. Both players are informed that if Player A's choice coincides with the real outcome, Option 1 is implemented. Otherwise, Option 2 is implemented. Player A is not informed of the possible payoffs associated with the two options; she learns her own earnings from the task only at the end of the experiment. Player B knows the payoffs associated with Options 1 and 2 for both players and also knows that Player A does not and will not know the payoffs, aside from her own final earnings.

If Player A chooses the number that is the actual outcome of the die roll, Option 1 is implemented and both players earn 10 tokens. Otherwise, Option 2 is implemented. We vary the payoff from Option 2 for Player B while keeping it constant for Player A as follows. In the *Selfish* domain, deception is beneficial for Player B; Player B earns 12 tokens and Player A earns 4. In the *Spiteful* domain, deception is costly for Player B; she earns 9 tokens and Player A earns 4.[5] Thus, in both cases, Option 2 is worse for Player A. Table 1 summarizes the payoffs in the two domains.

The richness of the message space in the deception game eliminates strategic truth-telling (Sutter 2009), because with such a rich space it is unlikely that participants will choose to tell the truth in order to obtain the unequitable payoff (see also the discussion in Erat and Gneezy 2012). Further, we assume Player B generally

[5] In Appendix D in ESM, we report the data of a variation of the treatments conducted in the *Spiteful* domain in which the cost of punishment is higher, with Player A earning eight instead of nine tokens, and Player B earning four tokens when choosing Option 2. The results of these treatments are in line with the results we report in this paper, but the lying rates were too low to make conclusive claims. We thank the editor and an anonymous referee for suggesting us to re-run the treatments in the *Spiteful* domain with a lower punishment cost.

expects her recommendations to be followed. This assumption rests on empirical evidence showing that telling the truth is unlikely to result in Option 2, even if senders expect receivers not to follow the message (van de Ven and Villeval 2015). As we will show in detail in Sect. 4.3, in our setup there is also a very small probability that telling the truth implements a lie. Therefore, we assume Player B would send an untruthful message in order to either increase her own payoff (*Selfish* domain) or to decrease Player A's payoff (both *Selfish* and *Spiteful* domains). In this paper, we classify all messages that contain a number different from the real outcome of the die roll as a lie.

Using a 2×3 design, we conducted six treatments, which are presented in Table 2. In all of our treatments, we used the strategy method (Selten 1967) for the second-stage decisions, and therefore Player B had to choose her action conditional on all possible first-stage outcomes. In a separate pilot study, we conducted the *Selfish-Intentions* treatment both using a between-subjects design and the strategy method. The study showed similar results, suggesting that in our setting using the strategy method leads to the same qualitative results obtained by the standard direct-response method. A detailed discussion of these results can be found in "Appendix C" in ESM.

In the *Intentions* treatments, Player B is asked (in Stage 2) to decide which message to send to Player A, for each feasible action of Player A in Stage 1. The *Intentions* treatments allow us to detect whether a relationship exists between the amount sent by Player A in Stage 1 and Player B's lying rates in Stage 2. In the *Direct-Choice* treatments, Player B's second-stage decision is a dictator decision, and therefore Player B directly decides whether Option 1 or Option 2 will be implemented without having to lie. The comparison between the lying rates in the *Intentions* treatments and the direct implementation rates in the *Direct-Choice* treatments informs us on whether reciprocity via deception is observed less often than reciprocity without deception. In the *No-Intentions* treatments, the experimenters determine the amount sent in the first stage; this information is common knowledge. By comparing the lying rates in the *Intentions* treatments with the lying rates in the *No-Intentions* treatments, we can isolate the effect of intentions from distributional concerns.

In the *No-Intentions* treatments, we build on previous literature (e.g., Blount 1995; Bolton et al. 1998; Charness 2004; Falk et al. 2008) and determine the payoff outcome based on the distribution of the first-stage outcomes of previous sessions of the other treatments.[6] In this way, we ensure Player B has the same beliefs about Player A's choice in both the *No-Intentions* and *Intentions* treatments. This distribution is made known to Player B without giving her information on how it is derived. Note that the first-stage instructions are the same in the *Intentions* and *Direct-Choice* treatments. Therefore, we pooled Stage 1 choices from these two treatments to determine the distribution of Stage 1 payoffs in the *No-Intentions* treatments. Player B is made aware of this distribution of the first-stage outcomes before she has to make a

[6] We presented subjects with pooled data from eight previous sessions. The distribution of the first-stage outcome between the subset and the full sessions is similar (p value = 0.672, two-sided Mann–Whitney test).

Table 2 Overview of the treatments in the *Selfish* and *Spiteful* domains

		Intentions	*Direct Choice*	*No-Intentions*
Selfish domain (Option 2: 4, 12)	Stage 1	*Dictator game* Player A chooses how much to give to Player B	*Dictator game* Player A chooses how much to give to Player B	*Experimenter* The experimenter informs Player B of her and Player A's initial allocation
	Stage 2	*Deception game* Player B is the sender	*Dictator game* Player B is the dictator	*Deception game* Player B is the sender
Spiteful domain (Option 2: 4, 9)	Stage 1	*Dictator game* Player A chooses how much to give to Player B	*Dictator game* Player A chooses how much to give to Player B	*Experimenter* The experimenter informs Player B of her and Player A's initial allocation
	Stage 2	*Deception game* Player B is the sender	*Dictator game* Player B is the dictator	*Deception game* Player B is the sender

decision. That is, after Player B has decided on her strategy, the experimenter asks her to draw a chip out of a bag containing 100 numbered chips. The selected chip determines the starting earnings that will be implemented. "Appendix B" in ESM shows the distribution of the possible earnings. This way, it is completely transparent to each Player B that the initial payoff allocation between her and Player A is not determined by Player's A intentions. After Player B draws the first-stage payoffs, Player A is informed about her own and her matched partner's payoffs. Player A then decides whether to follow Player B's message. This procedure ensures the information each player has in each decision-making stage in the *No-Intentions* treatments is comparable to the information received in the *Intentions* treatments.

2.2 Procedures

We conducted the experiment in the CeDEx laboratory at the University of Nottingham using students from a wide range of disciplines recruited through ORSEE (Greiner 2015). The experiment was carried out in a pen-and-paper format. We conducted 18 sessions (three per treatment) with a total of 560 subjects (363 of them female). Each session had an even number of participants ranging between 28 and 32. In each experimental session, every participant was randomly assigned to one of the two roles in the experiment (Player A or Player B) and was matched with another participant playing the other role. At the beginning of the experiment, subjects were informed about their role (A or B) and told that the experiment had two stages with the same matching of participants. They were also informed about the payment procedure that would follow at the end of the experiment and that the sum of payoffs obtained in both stages would determine the final payoff. Payoffs were in tokens, and each token was exchanged at a rate of 0.5 lb per token. Every session lasted approximately 45 min, and average earnings were 8.4 lb. Full instructions of the experiment are reported in "Appendix A" in ESM.

3 Theoretical predictions

In this section, we discuss the theoretical predictions across treatments for the behavior of Player B when taking into account purely selfish preferences, outcome-based fairness models, intention-based fairness models, and lying costs. The predictions are summarized in Table 3 for both the *Selfish* and *Spiteful* domains. In the table, b denotes the first-stage payoff of Player B, and *Option2* (*O2* hereafter) denotes the percentage of Option 2 choices by Player B in the second stage, regardless of whether the choice is implemented by direct choice or lying. Further, for ease of exposition, Table 3 only depicts predictions within each domain.

Standard models with selfish payoff-maximizing agents predict Player B always chooses the option with the higher payoff for herself, that is, Option 2 in the *Selfish* domain and Option 1 in the *Spiteful* one. This choice does not depend on whether a particular option is implemented through a lie or through direct choice.

Table 3 Predictions for the rate of choosing Option 2 or lying in the second stage

		Intentions	*No-Intentions*	*Direct-Choice*
Models with social preferences				
Purely selfish	*Selfish*	$O2=1, \forall b$	Same as in *Intentions*	Same as in *Intentions*
	Spiteful	$O2=0, \forall b$	Same as in *Intentions*	Same as in *Intentions*
Outcome-based fairness	*Selfish*	$O2$ decreases with b	Same as in *Intentions*	Same as in *Intentions*
	Spiteful	$O2$ decreases with b	Same as in *Intentions*	Same as in *Intentions*
Income effect	*Selfish*	$O2$ decreases with b	Same as in *Intentions*	Same as in *Intentions*
	Spiteful	$O2=0, \forall b$	Same as in *Intentions*	Same as in *Intentions*
Intention-based fairness	*Selfish*	$O2$ decreases with b	$O2$ constant, $\forall b$ $O2$ lower than in Intentions for low b, and higher for high b	Same as in *Intentions*
	Spiteful	$O2$ decreases with b, $O2=0$ for high b	$O2=0, \forall b$	Same as in *Intentions*
Models with moral cost of lying				
Fixed moral cost of lying and Hurkens and Kartik (2009)	*Selfish*	$O2$ constant, $\forall b$	Same as in *Intentions*	$O2$ higher than in *Intentions*, $\forall b$ $\frac{O2^{\text{Intentions}}}{O2^{\text{Direct-Choice}}}$ constant, $\forall b$
	Spiteful	$O2=0, \forall b$	Same as in *Intentions*	Same as in *Intentions*
Moral costs decrease with unfairness of the other	*Selfish*	$O2$ decreases with b	$O2$ constant, $\forall b$ $O2$ lower than in *Intentions* for low b, and higher for high b	$O2$ higher than in *Intentions*, $\forall b$ $\frac{O2^{\text{Intentions}}}{O2^{\text{Direct-Choice}}}$ decreases with b
	Spiteful	$O2$ decreases with b, $O2=0$ for high b	$O2=0, \forall b$	$O2=0$ for high b $O2$ higher than in *Intentions* and $\frac{O2^{\text{Intentions}}}{O2^{\text{Direct-Choice}}}$ decreases with b for all other b

Purely outcome-based theories of social preferences (e.g., Bolton and Ockenfels 2000; Fehr and Schmidt 1999) argue that people derive disutility from inequity in monetary payoffs. Such models would predict the percentage of Players B who choose the equitable option, namely Option 1, to be an increasing function of their first-stage payoff, because the higher Player B's first-stage payoff, the lower Player A's payoff. Player B's first-stage payoff is higher than Player A's if Player A sends four tokens or more, which would result in Player B choosing the equitable option. Note that Player B dislikes advantageous inequality, i.e., earning more than the other. Given Player B's first-stage payoffs ranging from 0, 6,… 30, the switching point from Option 2 to Option 1 depends on her disutility parameters from inequity. In general, the lower the disutility from being ahead or from having a higher than equitable share, the higher the switching point.[7] Because outcome-based social preference models do not take intentions or moral costs into account, they predict no treatment differences within a given domain. The result of the comparison of the outcome-based social preferences regarding the *Selfish* and *Spiteful* domains depends on the distribution of estimated disutility parameters.[8]

An income effect predicts a treatment pattern similar to the purely outcome-based social preferences in the *Selfish* domain: If Player B is more likely to choose the generous option (i.e., the equitable option) when she has more income, then in all the treatments in the *Selfish* domain, the percentage of Players B who choose the equitable option, namely Option 1, would be an increasing function of their first-stage payoff. An income effect predicts no treatment differences across the three treatments.

[7] For example, Fehr and Schmidt's (1999) model (hereafter, FS) distinguishes between advantageous and disadvantageous inequality, assuming the disutility parameter for the former is smaller than the latter. Their model only takes into account the difference between the final payoffs of the players. In the *Selfish* domain, the final-payoff vectors of Players A and B when Player B chooses the equitable option in the second stage are (20, 10), (18, 16), (16, 22), (14, 28), (12, 34), and (10, 40) if the first-stage payoff of Player A is 10, 8, 6, 4, 2, and 0, respectively. The final-payoff vectors when choosing the selfish option are (14, 12), (12, 18), (10, 24), (8, 30), (6, 36), and (4, 42) in the same order. According to previous parameter estimations of this model (Blanco et al. 2011; Beranek et al. 2015; Fehr and Schmidt 2006), most Players B would prefer the selfish option when the first-stage payoff of Player A is 10 and 8. If Player A's payoff is 6 or less, Player B's decision is the same regardless of the initial payoff allocation. This is due to the fact that FS compares the payoff of Player B earning 2 units extra, which reduces Player A's payoff by 6 units (because of the tripling), when in both payoff allocations Player B is ahead. In the *Spiteful* domain, FS would predict a majority of Players B will opt for the inequitable outcome in the second stage if the first-stage payoff of Player A is 10, a minority to do the same if the first-stage payoff of Player A is 8, and no Player B to choose the inequitable option if Player A's first-stage payoff is 6 or less. More concretely, based on the parameter estimates of the aggregate data of Beranek et al. (2015), we would expect Option 2 to be chosen at a rate of 100%, 52%, 20%, 20%, 20%, and 20% in the *Selfish* domain if Player A's first-stage payoff is 10, 8, 6, 4, 2, and 0, respectively. The corresponding rates in the *Spiteful* domain are 59%, 28%, 0%, 0%, 0%, and 0%, respectively.

[8] Comparing the *Selfish* and *Spiteful* domains requires parameter values of the FS model. Based on the parameter estimates of the aggregate data of Beranek et al. (2015), given the same first-stage payoff outcome, lying or direct-choice rates would always be higher in the *Selfish* than in the *Spiteful* domain. Because FS predicts no treatment differences within a domain, given the same first-stage outcome, we would expect to see all treatments in the *Selfish* domain to have higher inequitable choice rates than all treatments in the *Spiteful* domain.

In the *Spiteful* domain, an income effect predicts that subjects always choose the equitable option.

Purely intention-based models of reciprocity (e.g., Dufwenberg and Kirchsteiger 2004; Rabin 1993) posit that people respond to intentions by rewarding kind actions and punishing unkind ones. Thus, similar to the outcome-based social preference models, intention-based models would predict that the higher the first-stage payoff, the higher the likelihood that Player B chooses Option 1. However, the rationale is different: The higher first-stage payoff means Player A is kinder, and therefore Player B would reciprocate with a kinder choice. In the *Selfish* domain, the percentage of liars in the *No-Intentions* treatment would be the same for all first-stage payoffs. If the previous encounter is unkind, more lying would occur in the *Intentions* than in the *No-Intentions* treatment, and vice versa for kind encounters. Finally, no difference would exist between the *Direct-Choice* and *Intentions* treatments. In the *Spiteful No-Intentions* treatment, Player B should always choose according to her self-interest, namely Option 2. If the previous encounter is unkind, punishment would take place in both the *Intentions* and *Direct-Choice* treatments and at the same rate. Note that a model that is a mixture of both outcome- and intention-based preferences, such as the one proposed by Falk and Fischbacher (2006), only changes the predictions for the *No-Intentions* treatments: The prediction would be to observe a decrease in the percentage of Option 2 choices with increasing first-stage payoffs.

Finally, intention-based models of reciprocity would predict that, given an unkind first-stage choice of Player A, the percentage of Players B who choose the inequitable option is smaller in the *Spiteful* domain than in the *Selfish* domain because punishing an unkind offer in the *Spiteful* domain is costlier. Given a kind first-stage choice of Player A, we would expect some players to opt for the selfish option in the *Selfish* domain but no one to choose the inequitable option in the *Spiteful* domain. Therefore, given the same first-stage interaction, the percentage of players who choose the inequitable option in the *Intentions* and *Direct-Choice* treatments would be higher in the *Selfish* domain than in the *Spiteful* one.

With a fixed moral cost of lying, Player B's decision to lie depends on the lie's benefits and costs. That is, people are selfish but also incur a cost from lying, and therefore the rate of lying does not depend on the first-stage outcome. In the *Selfish* domain, the percentage of lies in the *Intentions* treatment would be lower than the percentage of direct choices in the *Direct-Choice* treatment. Because intentions do not matter, lying rates would be the same in the *Intentions* and *No-Intentions* treatments. In the *Spiteful* domain, no lying would occur.

Three additional mechanisms also assume a moral cost of lying. First, Hurkens and Kartik (2009) conjectured that individuals have either a zero or infinite moral cost of lying. Absent lying costs, people always lie if they prefer the outcome from lying to the outcome from not lying. People with an infinite cost of lying never lie. Then the conditional probability of lying (i.e., the ratio of lying to direct choice) would be constant regardless of the type of previous encounter, because a constant proportion of people who directly choose an outcome have an infinite moral cost of lying. This argument assumes no relationship between preferences toward fairness and one's moral cost type. A second mechanism is based on the premise that moral costs are malleable and decrease when one can blame the other party (Bandura et al.

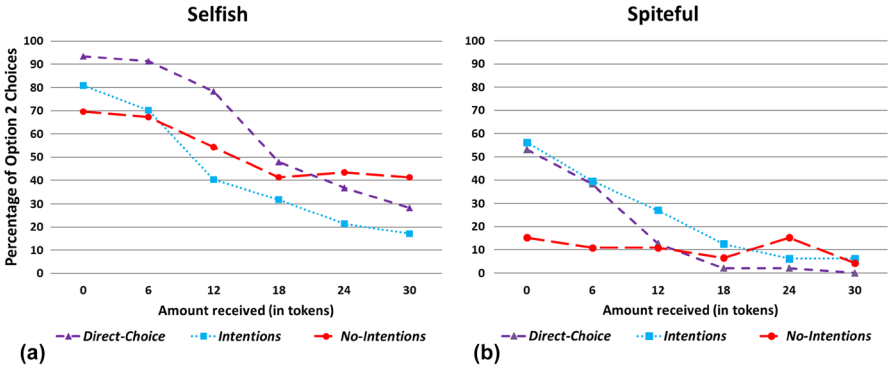

Fig. 1 Percentage of Option 2 choices per amount received across treatments in the *Selfish* and *Spiteful* domains

1996; Bandura 1999, 2002, 2004; Ditto et al. 2009). Consequently, the unkinder the previous interaction, the lower the moral cost of lying. Thus, the conditional probability of lying would decrease with increasing kindness. Whether subjects would also lie in the *Spiteful* domain depends on whether they are willing to forgo payoffs to punish unkind behaviors, and if so, the prediction on the conditional probability of lying would be the same as in the *Selfish* domain. Finally, the moral cost of lying may also depend on whether lying results in a benefit or a loss for the deceiver. Two competing hypotheses arise from prior literature. If the moral cost is higher when the lie is beneficial due to being perceived as selfish (e.g., Carlson and Zaki 2018; Lin-Healy and Small 2012; Newman and Cain 2014), the conditional probability of lying would be smaller in the *Selfish* domain than in the *Spiteful* domain. Instead, if the moral cost is smaller when the lie is beneficial due to a self-serving bias (e.g., Babcock et al. 1995; Dana et al. 2007; Exley 2015; Gneezy et al. 2017; Konow 2000), the conditional probability of lying would be larger in the *Selfish* domain than in the *Spiteful* domain.

4 Results

In this section, we focus on our main research question, which concerns the behavior of Players B. Player A's dictator-game decisions are reported in "Appendix B" in ESM. Figure 1 depicts the percentage of Players B who chose Option 2 either via lying or direct choice in the *Selfish* (Fig. 1a) and *Spiteful* (Fig. 1b) domains, respectively. In each figure, we report the percentage of Players B who chose Option 2 in the second stage for each possible amount of tokens they received from the first stage. For example, in Fig. 1a, the percentage of Players B who chose the selfish option directly when Players A sent them zero tokens was 93.5%, whereas 80.9% of Players B lied to implement the same outcome. When the experimenters implemented identical first-stage payoffs, 69.6% of Players B lied to implement the same outcome.

Looking at Fig. 1, we observe that the percentage of Players B who choose Option 2 decreased with the amount sent by Player A in the first stage in both the *Selfish* and *Spiteful* domains. Such a decrease cannot be explained by purely selfish preferences. We further observe a difference in the percentage of choosing Option 2 between the *Intentions* and *No-Intentions* treatments in both domains, a result that rules out models that focus exclusively on distributional preferences. We further investigate the role of intentions and payoff outcomes in Sect. 4.2. Finally, the percentage of Option 2 choices markedly differs between *Direct-Choice* and *Intentions* treatments for all first-stage payoffs only in the *Selfish* domain, pointing to a moral cost of lying that is not fixed, but highly context-dependent. Next, we describe our results regarding the moral cost of lying in detail.

4.1 Moral cost of lying

Selfish lies We first focus on the results concerning the moral cost of lying. Figure 1a depicts the percentage of Players B who choose Option 2 either directly via a dictator-game choice or by engaging in deception after receiving any of the six possible amounts in the *Selfish* domain. Option 2 is beneficial for Player B but is costly for Player A. Our results reveal that Option 2 was chosen less often when doing so required Players B to lie, suggesting lying entails a moral cost. Indeed, lying rates in the *Selfish-Intentions* treatment were smaller than the rates of Option 2 choices in the *Selfish-Direct-Choice* treatment for *any* amount received in the first stage.[9]

The results of a linear regression corroborate this finding. In Table 4, we report the results of a linear probability model in which we regress the choice of Option 2 on the amount received by Player B, treatment dummies, and the interaction terms.[10] Column 1 reports the results for the *Selfish* domain. The results reveal that, when receiving no tokens from their counterpart, Players B chose Option 2 more often in the *Direct-Choice* than in the *Intentions* treatment ($b = 0.22$, $p = 0.003$). With each additional token sent from Player A, the probability of choosing Option 2 fell by approximately 7% points. The results are the same when pooling the data from both domains (Column 3).

A comparison of the rates of Option 2 choices after receiving nothing and after receiving everything further shows the effect of first-stage payoffs. In the *Selfish-Direct-Choice* treatment, 93.5% of Players B chose Option 2 if faced with receiving zero tokens, while this share dropped to 28.3% if they were to receive 30 tokens. In the *Selfish-Intentions* treatment, the corresponding values were 80.9% and 17%, respectively. Individuals did not respond differently to a given amount if the payoff

[9] One possible concern is that Player B's behavior would differ in the two domains not only because of the absence of a moral cost of lying in the *Direct-Choice* treatment, but also because in that treatment, her preferred outcome can be implemented with certainty. However, such an effect is negligible, because if Player B lies, the probability of Player A guessing the correct number is very low, and this information is common knowledge. Indeed, we observe that if Player B lied, Option 2 was implemented 98% of the time in the *Intentions* treatments (43/44).

[10] In Appendix B in ESM, we report the results from the linear probability model while controlling for gender. We find some—but not systematic—evidence that females opt more for Option 2.

Table 4 Regression analysis of choosing Option 2

	DV: Option 2		
	(1) Selfish	(2) Spiteful	(3) Pooled
AmountSent	−0.068*** (0.008)	−0.052*** (0.008)	−0.068*** (0.008)
Direct-Choice	0.223*** (0.072)	−0.051 (0.096)	0.223*** (0.072)
Direct-Choice × AmountSent	−0.006 (0.012)	−0.003 (0.011)	−0.006 (0.012)
NoIntentions	−0.085 (0.094)	−0.369*** (0.083)	−0.085 (0.094)
NoIntentions × AmountSent	0.035*** (0.012)	0.046*** (0.010)	0.035*** (0.012)
Spiteful			−0.268*** (0.092)
Spiteful × AmountSent			0.016 (0.011)
Spiteful × Direct-Choice			−0.274** (0.120)
Spiteful × Direct-Choice × AmountSent			0.004 (0.017)
Spiteful × NoIntentions			−0.285** (0.125)
Spiteful × NoIntentions × AmountSent			0.010 (0.016)
Constant	0.775*** (0.061)	0.507*** (0.069)	0.775*** (0.061)
Observations	834	846	1680
Clusters	139	141	280
R^2	0.199	0.178	0.300

This table reports coefficient estimates from a linear probability model [we use OLS instead of probit in order to analyze and interpret the interaction terms, because the significance of the multiplicative term in a non-linear model is not a proper indicator for the significance of the interaction (see Ai and Norton 2003)], with standard errors clustered at the individual level reported in parentheses. The dependent variable is the choice of Option 2. Direct-Choice is a dummy for the treatment *Direct-Choice,* NoIntentions is a dummy for the *No-Intentions* treatment, and Spiteful is a dummy for the *Spiteful* domain. AmountSent refers to the amount sent by Player A in Stage 1

***, **, and * represent significance at the 1%, 5%, and 10% levels, respectively

allocation was determined via a direct choice or via a lie (AmountSent × Direct-Choice b = −0.006, $p = 0.593$). In other words, Option 2 choices decreased with increasing amounts sent at the same rate in the *Intentions* and *Direct-Choice* treatments.

A second observation relates to the *relative* moral cost of lying, that is, the conditional probability of lying when given an incentive to do so. If either the fixed moral cost of lying or the Hurkens and Kartik's (2009) conjecture is correct, the ratio of lying in the *Intentions* treatment to the ratio of Option 2 choices in the *Direct-Choice* treatment should be constant for all first-stage payoffs of Players B.[11] On the other hand, if the relative moral cost of lying increases with the kindness of the previous interaction, this ratio would decrease. Our results support the latter: The conditional probability of lying was 86.5% (80.9/93.5) when Player A sent zero tokens, whereas

[11] Following Hurkens and Kartik's (2009) argumentation, we assume the subjects for both the *Selfish-Intentions* and the *Selfish-Direct-Choice* treatments are drawn randomly from the same population distribution and that those who choose the equitable option (Option 1 gives 10 for both) directly would never lie to implement Option 2.

it was only 60.2% (17.0/28.3) when Player A sent 30 tokens. The difference between the two conditional probabilities is significant ($p = 0.031$).[12] This finding suggests that lying to a person who has been kind is morally costlier.[13]

Spiteful Lies The results for costly deception are depicted in Fig. 1b. Surprisingly, in contrast with the pattern observed in the *Selfish* domain, the moral cost of lying disappears when deception entails a monetary cost for Player B. Option 2 rates were similar between the *Intentions* and *Direct-Choice* treatments, indicating the lack of a moral cost of lying. The regression results reported in Table 4 (column 2) confirm these findings. The results are confirmed when pooling the data from the *Selfish* and *Spiteful* domains (Column 3). When receiving no tokens from Player A, Players B were equally likely to punish their counterpart by choosing Option 2 in the *Direct-Choice* treatment and in the *Intentions* treatment ($b = -0.052$, $p = 0.596$). With each additional token sent by Player A in Stage 1, the rate of Option 2 choices dropped by approximately 5 percentage points, with no difference across the two treatments (AmountSent × Direct-Choice $b = -0.055$, $p = 0.799$). This rate was close to zero for kind encounters.

That Option 2 rates were similar between the *Intentions* and *Direct-Choice* treatments in the *Spiteful* domain but not in the *Selfish* domain proves that the moral cost of lying interacts with one's monetary benefit or cost from lying. If anything, individuals were more likely to punish via lying than via direct choice when lying is costly, a result in the opposite direction to what the presence of a moral cost of lying predicts (see also Fig. 1b). This result is in line with the idea that the morality of an action is judged differently if the action is beneficial for the self (e.g., Carlson and Zaki 2018; Lin-Healy and Small 2012; Newman and Cain 2014).

4.2 The role of intentions

Next, we explore the extent to which outcome-based preferences and intentions play a role in second stage choices by Players B. The change in Option 2 choices with increasing first-stage payoffs in the *No-Intentions* treatment shows whether and how

[12] We use the normal approximation to the log odds ratio, which gives us two normal distributions. The difference between two log odds ratios is therefore also normally distributed.

[13] Notice that a decreasing conditional probability of lying also refutes a model in which individuals have a fixed moral cost of lying that is continuously distributed between zero and a finite value, and social preferences are independent of lying costs. In such a case, the proportion of people who do not lie when they would prefer the same option via direct choice should also be constant, because within every group of people who prefer one outcome to another, their lying costs would be randomly drawn from the population distribution. On the other hand, a decreasing conditional probability of lying is in line with a model in which individuals have a fixed moral cost of lying, but social preferences are correlated with lying costs. Such a model would need to assume punishers or people who dislike disadvantageous inequity have on average a lower moral cost of lying than selfish people, a rather implausible assumption, as the conditional probability of lying is smaller after the unkind encounter than after the kindest encounter. Whereas both fairness-constrained and selfish people would choose the selfish option after an unkind encounter, only selfish people would choose the selfish option after receiving everything from the other player in the first stage; thus, selfish people should have a higher moral cost of lying for the conditional probability of lying to be smaller.

much payoff outcomes matter for Players B. The comparison of the *Intentions* treatments, in which Players A determine the first-stage outcome, with the *No-Intentions* treatments, in which the experimenter implements the first-stage outcome, allows us to isolate the role of intentions from other drivers of behavior, namely payoff inequality concerns and income effects.[14]

Selfish Lies A first observation is that lying decreased significantly with increasing first-stage payoffs in the *No-Intentions* treatment ($p = 0.001$), which was also the case in the *Intentions* treatment. However, intentions matter more than initial payoff allocations: Column 1 of Table 4 shows that each token sent by Player A in Stage 1 decreased the probability of lying by 3.3 percentage points in the *No-Intentions* treatment versus 6.8 percentage points in the *Intentions* treatment. The difference between the two treatments was statistically significant ($p < 0.001$). We can therefore conclude that both intentions and initial payoff allocations affect the decision to lie, with the former effect being significantly larger when lying is beneficial to oneself.

When comparing the effect of unkind and kind interactions at the extremes, our data showed that the effect of intentions was larger for kind interactions. Players B rewarded kindness by lying less (i.e., they behaved more honestly): After receiving 30 tokens, 41.3% of Players B lied in the *Selfish-No-Intentions* treatment versus 17% in the *Selfish-Intentions* treatment ($p = 0.012$, Fisher's two-sided exact test). Players B punished unkind interactions by lying more. However, this effect was smaller and not statistically significant: After receiving zero tokens, 69.6% of Players B lied in the *Selfish-No-Intentions* treatment versus 80.9% in the *Selfish-Intentions* ($p = 0.237$, Fisher's two-sided exact test).

That the effect of kindness is larger than the effect of unkindness in the *Selfish* domain is in line with the findings of Rand et al. (2009), who showed that reward is just as effective as punishment in sustaining cooperation, but is in contrast with previous work on reciprocity showing that negative reciprocity is stronger than positive reciprocity (e.g., Charness 2004; Nosenzo et al. 2014; Offerman 2002). The discrepancy could be due to the fact that previous studies on reciprocity utilized direct choice rather than deception. Our results point to a more complex picture regarding lying and truth-telling as reciprocity devices: The moral cost of lying interacts with the kindness of a prior interaction, and hence we see honesty being used as a reward. Notice also that the weak evidence for punishment in our setup might also be due to the very high lying rate in the *No-Intentions* treatment. If, as some previous studies suggested (e.g., Erat and Gneezy 2012; Gneezy 2005; Hurkens and Kartik 2009), a minority of individuals never lie regardless of the consequences, the upper bound

[14] Two factors might be driving deceptive behavior in the *No-Intentions* treatments. The first is simply an income effect. The second is a behavioral response to reduce inequality between the interacting pair. Previous evidence suggests income effects do not drive behavior in reciprocity (Johnson et al. 2006). Additionally, a pilot experiment we conducted, which is reported in Appendix C in ESM, also suggests income effects do not affect lying behavior. Because our focus is on isolating the effect of intentions, teasing out income effects and distributional concerns from the effect of initial payoff allocations is beyond our scope.

on the percentage of lying would be smaller than 100%.[15] Given that we observed 69.6% of people lying in the *Selfish-No-Intentions* treatment after receiving zero tokens in the first stage, detecting an upward effect is unlikely. To sum up, in the *Selfish* domain, lying after an initial encounter depends on intentions and initial payoff allocations, but the effect of intentions is larger than the effect of initial payoff allocations.

Spiteful Lies As depicted in Fig. 1b, in the *No-Intentions* treatment, lying rates were close to zero irrespective of the nature of the previous encounter. The percentage of Players B who lied did not significantly change with their own payoff in the first stage. Column 2 of Table 4 shows that whereas in the *Intentions* treatment each token sent by Player A in Stage 1 decreased the probability of lying by 5.2 percentage points ($p < 0.001$), in the *No-Intentions* treatment this effect dropped to 0.6 percentage points ($p = 0.183$). Thus, when lying is costly for the self, only (unkind) intentions lead individuals to lie. We find strong evidence for punishment in the *Spiteful* domain. After Player A sent zero tokens, 56.3% of Players B lied in the *Intentions* treatment versus only 15.2% in the *No-Intentions* treatment ($p = 0.003$, Fisher's two-sided exact test). As for rewarding kind intentions, because lying rates were close to zero irrespective of the first-stage payoffs in the *No-Intentions* treatment, no room remained for detecting rewards via truth-telling in the *Intentions* treatment of the *Spiteful* domain.

To conclude, when lying benefits the deceiver, both intentions and initial payoff allocations affect the decision to lie, whereas when lying is costly for the deceiver, only intentions affect lying behavior.

4.3 Following rates and discussion

Following rates in the selfish and spiteful domains We report Player A's behavior in the *Intentions* and *No-Intentions* treatments after she receives a message from Player B. The first-stage choices of Players A are reported in "Appendix B" in ESM. Overall, the percentage of Players A who followed the message was 72.2%. This follow rate is in line with reported follow rates from the literature (Gneezy 2005; Hurkens and Kartik 2009; van de Ven and Villeval 2015). In addition, in the *Intentions* treatments, Player A's following behavior was correlated with the kindness of her first-stage action: Player A was significantly less likely to follow Player B's message when she was unkind and sent zero tokens than when she was kind. Therefore, some Players A anticipated that Player B's message might be untruthful. However, note that in a separate study we conducted prior to the current work, we observed no correlation between the kindness of Players A and their subsequent following behavior (see "Appendix C" in ESM for details).

[15] In our lying treatments in the *Selfish* domain, 48.4% of Players B switched from lying to telling the truth, 20.4% always lied, 21.5% never lied, and 9.7% switched non-monotonically with the increasing kindness of Player A. In the lying treatments in the *Spiteful* domain, 52.2% of Players B responded to unkindness, whereas 42.4% never lied and the remaining 5.4% switched non-monotonically.

The correlation between the kindness of Player A's choice and the subsequent follow rate opens up the possibility that some Players B anticipated that Players A would not follow their message (second-order beliefs) and therefore chose to lie in order to increase the chance of implementing the equal payoff outcome. We call this behavior "truthful lying", following Sutter (2009). If Players B have correct beliefs about which option will be implemented as a result of their message, they should only "truthfully lie" if doing so is effective. However, in our experiment, the best way to implement the equal-payoff outcome was by sending an honest message. In the *Intentions* treatments, a lie implemented the unequal outcome (Option 2) 98% of the time, and honesty implemented the equal-payoff outcome (Option 1) 82% of the time. Thus, if Player B wanted to implement the equal outcome, she would have sent an honest message.

5 Conclusion

In this paper, we investigate whether (a lack of) deception serves as a reciprocity device and the moral cost it entails. Whereas honesty can be used to reward kind behavior, lying can be used to punish unkind behavior. We study situations in which lying creates a monetary advantage at the receiver's expense and situations in which lying is costly for both the sender and the receiver. We further explore whether differences in deception rates in both circumstances are triggered by initial payoff allocations or unkind intentions.

We find evidence of a moral cost of lying when punishment via deception benefits the deceiver (*Selfish* domain), but not when it is costly (*Spiteful* domain). In both domains, reciprocity plays a crucial role in lying decisions: With selfish lies, individuals reward kind intentions with honesty, and we only find directional evidence for punishment of unkind intentions via lying. With spiteful lies, individuals deceive in order to punish unkind intentions. Overall, our findings suggest that the moral cost of lying is malleable and depends on whether the deceiver benefits from or pays a cost for lying. The presence of moral costs only in the *Selfish* domain supports the conjecture of reciprocity motives being tainted when self-interested motives are also present: Individuals with reputational concerns might refrain from lying as it is unclear whether lying is driven by the desire to gain a material benefit or the desire to punish unkind behavior. Future work can further explore how the presence of different motives affects the moral cost of lying.

This study contributes to our understanding of the prevalence of unethical behavior and the motives for it. Our results demonstrate the malleability of lying costs and thereby indicate that models of deception need to take into account that, in repeated interactions between individuals, the decision to lie depends both on one's payoff from lying and reciprocity concerns. Our study is a first step in that direction. Taken to an organizational setting, our results suggest that employees' perception of how fairly they are treated in the workplace matters for preventing unethical behaviors, because kind acts are likely to be reciprocated with honest behavior. Further, some individuals might be willing to punish unfair behavior by using deception that is costly to themselves. Especially in situations in which other punishment alternatives

are absent, the moral costs of lying might be too small to prevent individuals from using deception as a form of (costly) retaliation. Organizations could design interventions that encourage positive reciprocity to enhance the saliency of the moral costs associated with behaving unethically and contribute to fostering a culture of honesty. Setting up channels through which workers can address unfair treatment could help prevent deception from being used as a punishment device.

Acknowledgements We gratefully acknowledge the financial support by the Centre for Decision Research and Experimental Economics (CeDEx) of the University of Nottingham, the Rady School of Management of the University of California San Diego, the Research Priority Area of Behavioral Economics of the University of Amsterdam (201501270301), and the Jürgen Meyer Stiftung. We are grateful to Vasiliki Karasi for excellent research assistance. We would like to thank Simon Gächter, Uri Gneezy, Martin Kocher, Rudy Ligtvoet, Marina Schröder, Chris Starmer, Fabio Tufano, Jeroen van de Ven, and Roel van Veldhuizen, as well as seminar participants at the University of Nottingham, NIBS Brown-Bag Seminar, ESA conference in Heidelberg, and Birmingham Ph.D. Decision Making Workshop for helpful comments. Despoina Alempaki acknowledges support from the ESRC funded Network for Integrated Behavioral Science (ES/K002201/1, ES/P008976/1) and the Leverhulme 'Value' Programme (RP2012-V-022).

References

Abeler, J., Nosenzo, D., & Raymond, C. (forthcoming). Preferences for truth-telling. *Econometrica* **(in press)**.

Ai, C., & Norton, E. C. (2003). Interaction terms in logit and probit models. *Economics Letters, 80*(1), 123–129.

Babcock, L., Loewenstein, G., Issacharoff, S., & Camerer, C. (1995). Biased judgments of fairness in bargaining. *The American Economic Review, 85*(5), 1337–1343.

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review, 3*(3), 193–209.

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education, 31*(2), 101–119.

Bandura, A. (2004). The role of selective moral disengagement in terrorism and counterterrorism. In F. M. Moghaddam & A. J. Marsella (Eds.), *Understanding terrorism: Psychological roots, consequences and interventions* (pp. 121–150). Washington: American Psychological Association.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Mechanisms of moral disengagement in the exercise of moral agency. *Journal of Personality and Social Psychology, 71*(2), 364.

Barr, A., & Michailidou, G. (2017). Complicity without connection or communication. *Journal of Economic Behavior & Organization, 142,* 1–10.

Battigalli, P., Charness, G., & Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization, 93,* 227–232.

Behnk, S., Hao, L., & Reuben, E. (2017). Partners in crime: Diffusion of responsibility in antisocial behaviors. IZA Discussion Paper No. 11031. http://ftp.iza.org/dp11031.pdf. Accessed 23 Nov 2018.

Belot, M., & Schröder, M. (2013). Sloppy work, lies and theft: A novel experimental design to study counterproductive behavior. *Journal of Economic Behavior & Organization, 93,* 233–238.

Beranek, B., Cubitt, R., & Gächter, S. (2015). Stated and revealed inequality aversion in three subject pools. *Journal of the Economic Science Association, 1*(1), 43–58.

Blanco, M., Engelmann, D., & Normann, H. T. (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior, 72*(2), 321–338.

Blount, S. (1995). When social outcomes aren' t fair: The effect of causal attributions on preferences. *Organizational Behavior and Human Decision Processes, 63*(2), 131–144.
Bolton, G. E., Brandts, J., & Ockenfels, A. (1998). Measuring motivations for the reciprocal responses observed in a simple dilemma game. *Experimental Economics, 1*(3), 207–219.
Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review, 90*(1), 166–193.
Bucciol, A., & Piovesan, M. (2011). Luck or cheating? A field experiment on honesty with children. *Journal of Economic Psychology, 32*(1), 73–78.
Carlson, R. W., & Zaki, J. (2018). Good deeds gone bad: Lay theories of altruism and selfishness. *Journal of Experimental Social Psychology, 75,* 36–40.
Charness, G. (2004). Attribution and reciprocity in an experimental labor market. *Journal of Labor Economics, 22*(3), 665–688.
Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica, 74*(6), 1579–1601.
Charness, G., & Dufwenberg, M. (2010). Bare promises: An experiment. *Economics Letters, 107*(2), 281–283.
Charness, G., & Haruvy, E. (2002). Altruism, equity, and reciprocity in a gift-exchange experiment: An encompassing approach. *Games and Economic Behavior, 40*(2), 203–231.
Charness, G., & Levine, D. I. (2007). Intention and stochastic outcomes: An experimental study. *The Economic Journal, 117*(522), 1051–1072.
Cubitt, R. P., Drouvelis, M., Gächter, S., & Kabalin, R. (2011). Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics, 95*(3), 253–264.
Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory, 33*(1), 67–80.
Danilov, A., Biemann, T., Kring, T., & Sliwka, D. (2013). The dark side of team incentives: Experimental evidence on advice quality from financial service professionals. *Journal of Economic Behavior & Organization, 93,* 266–272.
Ditto, P. H., Pizarro, D. A., & Tannenbaum, D. (2009). Motivated moral reasoning. *Psychology of Learning and Motivation, 50,* 307–338.
Dreber, A., & Johannesson, M. (2008). Gender differences in deception. *Economics Letters, 99*(1), 197–199.
Dufwenberg, M., & Dufwenberg, M. A. (2018). Lies in disguise—A theoretical analysis of cheating. *Journal of Economic Theory, 175,* 248–264.
Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior, 47*(2), 268–298.
Ellingsen, T., Johannesson, M., Lilja, J., & Zetterqvist, H. (2009). Trust and truth. *The Economic Journal, 119*(534), 252–276.
Erat, S., & Gneezy, U. (2012). White lies. *Management Science, 58*(4), 723–733.
Exley, C. L. (2015). Excusing selfishness in charitable giving: The role of risk. *The Review of Economic Studies, 83*(2), 587–628.
Falk, A., Fehr, E., & Fischbacher, U. (2008). Testing theories of fairness—Intentions matter. *Games and Economic Behavior, 62*(1), 287–303.
Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior, 54*(2), 293–315.
Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review, 90,* 980–994.
Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114*(3), 817–868.
Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism—Experimental evidence and new theories. In S. C. Kolm & J. M. Ythier (Eds.), *Handbook of the economics of giving, altruism and reciprocity* (Vol. 1, pp. 615–691). Amsterdam: Elsevier.
Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—An experimental study on cheating. *Journal of the European Economic Association, 11*(3), 525–547.
Gino, F., Krupka, E. L., & Weber, R. A. (2013). License to cheat: Voluntary regulation and ethical behavior. *Management Science, 59*(10), 2187–2203.
Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95,* 384–394.
Gneezy, U., Kajackaite, A., & Sobel, J. (2018). Lying aversion and the size of the lie. *American Economic Review, 108*(2), 419–453.
Gneezy, U., Saccardo, S., Serra-Garcia, M., & van Veldhuizen, R. (2017). Bribing the Self. *mimeo*.

Greenberg, A. E., Smeets, P., & Zhurakhovska, L. (2015). Promoting truthful communication through ex-post disclosure. https://doi.org/10.2139/ssrn.2544349.
Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association, 1*(1), 114–125.
Houser, D., Vetter, S., & Winter, J. (2012). Fairness and cheating. *European Economic Review, 56*(8), 1645–1655.
Hurkens, S., & Kartik, N. (2009). Would I lie to you? On social preferences and lying aversion. *Experimental Economics, 12*(2), 180–192.
John, L. K., Loewenstein, G., & Rick, S. I. (2014). Cheating more for less: Upward social comparisons motivate the poorly compensated to cheat. *Organizational Behavior and Human Decision Processes, 123*(2), 101–109.
Johnson, L. T., Rutström, E. E., & George, J. G. (2006). Income distribution preferences and regulatory change in social dilemmas. *Journal of Economic Behavior & Organization, 61*(2), 181–198.
Kajackaite, A., & Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior, 102,* 433–444.
Khalmetski, K., & Sliwka, D. (2017) Disguising lies—Image concerns and partial lying in cheating games. *CESifo Working Paper Series No. 6347*. https://ssrn.com/abstract=2933434. Accessed 23 Nov 2018.
Kocher, M. G., Schudy, S., & Spantig, L. (2017). I lie? We lie! Why? Experimental evidence on a dishonesty shift in groups. *Management Science*, *64*(9), 3971–4470.
Konow, J. (2000). Fair shares: Accountability and cognitive dissonance in allocation decisions. *American Economic Review, 90*(4), 1072–1091.
Krueger, A. B., & Mas, A. (2004). Strikes, scabs, and tread separations: Labor strife and the production of defective Bridgestone/Firestone tires. *Journal of Political Economy, 112*(2), 253–289.
Lin-Healy, F., & Small, D. A. (2012). Cheapened altruism: Discounting personally affected prosocial actors. *Organizational Behavior and Human Decision Processes, 117*(2), 269–274.
López-Pérez, R., & Spiegelman, E. (2013). Why do people tell the truth? Experimental evidence for pure lie aversion. *Experimental Economics, 16*(3), 233–247.
Lundquist, T., Ellingsen, T., Gribbe, E., & Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization, 70*(1), 81–92.
Mas, A. (2006). Pay, reference points, and police performance. *The Quarterly Journal of Economics, 121*(3), 783–821.
Mas, A. (2008). Labour unrest and the quality of production: Evidence from the construction equipment resale market. *The Review of Economic Studies, 75*(1), 229–258.
Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research, 45*(6), 633–644.
Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological Science, 25*(3), 648–655.
Nosenzo, D., Offerman, T., Sefton, M., & van der Veen, A. (2014). Encouraging compliance: Bonuses versus fines in inspection games. *Journal of Law Economics and Organization, 30*(3), 623–648.
Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review, 46*(8), 1423–1437.
Rabin, M. (1993). Incorporating fairness into game theory and economics. *The American Economic Review, 83*(5), 1281–1302.
Rand, D. G., Dreber, A., Ellingsen, T., Fudenberg, D., & Nowak, M. A. (2009). Positive interactions promote public cooperation. *Science, 325*(5945), 1272–1275.
Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (pp. 136–168). Tübingen: J.C.B. Mohr (Paul Siebeck).
Sutter, M. (2009). Deception through telling the truth?! Experimental evidence from individuals and teams. *The Economic Journal, 119*(534), 47–60.
Treviño, L. K., den Nieuwenboer, N. A., & Kish-Gephart, J. J. (2014). (Un) ethical behavior in organizations. *Annual Review of Psychology, 65,* 635–660.
Van de Ven, J., & Villeval, M. C. (2015). Dishonesty under scrutiny. *Journal of the Economic Science Association, 1*(1), 86–99.
Weisel, O., & Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences, 112*(34), 10651–10656.