The Effect of Social Rewards and Punishments on Deceptive Behavior

Authors: Yuan Bo, Zhao Jingshi, Qi Dan, Tong Zhao, Hu Jiaqi, Yuan Bo

Date: 2025-03-19T00:00:00+00:00

Abstract

Deception refers to the act of providing false information or concealing relevant information to others through verbal or non-verbal means, and rewards and punishments are key factors influencing deceptive behavior. However, previous research has primarily focused on monetary rewards and punishments, and it remains unclear how social rewards and punishments affect deceptive behavior. This study investigates the effects of social rewards and punishments on deceptive behavior and their mediating and moderating mechanisms through three experiments. Experiment 1 employs a signaling game task to examine whether social rewards and punishments can influence deceptive behavior. The results reveal that, similar to monetary rewards and punishments, social rewards and punishments can reduce deceptive behavior, and social punishment is more effective than social reward. Analysis using the drift-diffusion model reveals that the drift rate v under social reward/punishment and monetary reward/punishment conditions is significantly smaller than under no-reward/punishment conditions, indicating that social and monetary rewards/punishments prompt individuals to accumulate evidence more favorably toward the non-deceptive choice. Experiments 2 and 3 use the same experimental task to further examine the mediating role of reputation concern in the effect of social rewards and punishments on deceptive behavior, and the moderating role of social value orientation. The results reveal that reputation concern mediates the effect of social rewards and punishments on deceptive behavior, and social value orientation moderates the mediating process through which social rewards and punishments affect deceptive behavior via reputation concern. These results indicate that social rewards and punishments can serve as effective means to inhibit deceptive behavior.

Full Text

The Influence of Social Reward and Punishment on Deception

YUAN Bo, ZHAO Jingshi, QI Dan, ZHAO Tong, HU Jiaqi
(Department of Psychology, Ningbo University, Ningbo 315211, China)

Abstract

Deception refers to behaviors that provide others with false information or conceal relevant information through verbal or nonverbal means. Rewards and punishments are key factors influencing deceptive behavior. However, previous research has primarily focused on monetary rewards and punishments, leaving unclear how social rewards and punishments affect deception. This study investigated the impact of social rewards and punishments on deceptive behavior and its mediating and moderating mechanisms through three experiments. Experiment 1 employed a signaling game task to examine whether social rewards and punishments could influence deception. The results revealed that, similar to monetary incentives, social rewards and punishments reduced deceptive behavior, with social punishment proving more effective than social reward. Drift-Diffusion Model analysis showed that the drift rate under both social and monetary reward/punishment conditions was significantly lower than under the no-reward/punishment condition, indicating that both types of incentives prompted individuals to accumulate evidence favoring non-deceptive choices. Experiments 2 and 3 used the same experimental task to further examine the mediating role of reputation concern and the moderating role of social value orientation. The results demonstrated that reputation concern mediated the effect of social rewards and punishments on deception, while social value orientation moderated this mediating process. These findings suggest that social rewards and punishments can serve as effective means to inhibit deceptive behavior.

Keywords: social reward and punishment, deception, reputation concern, social value orientation, Drift-Diffusion Model
Classification Number: B849: C91

1 Introduction

Deception is a behavior that provides others with false information or conceals relevant information through verbal or nonverbal means to achieve certain goals (Depaulo et al., 2003). It involves a series of cognitive and behavioral processes, including information selection, processing, and transmission (Carr et al., 2019). Deception is ubiquitous in social life and imposes substantial costs on organizations and society (Mazar & Ariely, 2006). To better formulate policies and institutions for reducing deceptive behavior, we must understand the factors that influence people' s decisions to deceive. Under what circumstances do individuals choose to deceive? According to the rational actor assumption in standard

economic models, individuals weigh the expected external benefits and costs when engaging in deception (Allingham & Sandmo, 1972; Becker, 1968). Consequently, rewards and punishments, as mechanisms that increase or decrease specific motivations or behaviors, represent important tools for intervening in deceptive behavior.

Numerous studies have examined the impact of material rewards and punishments, such as monetary incentives, on deception (Gneezy, 2005; Kaushik et al., 2022). For instance, Mazar et al. (2008) found that individuals exhibited more deception when it yielded greater benefits. Conversely, when honesty was rewarded monetarily, deceptive behavior decreased (Rosenbaum et al., 2014). Similarly, monetary punishment suppressed deception by increasing its costs. Research has shown that increasing the likelihood and severity of monetary punishment for deception reduces individuals' deceptive behavior (Behnk et al., 2018; Nagin & Pogarsky, 2003). However, monetary incentives are a double-edged sword: while they inhibit deception, they also incur high costs. Moreover, once individuals adapt to external monetary reinforcement, reducing or eliminating such reinforcement may cause the desired behavior to revert to baseline or even lower levels (Mulder et al., 2006). For example, monetary punishment can transform the intrinsic motivation for prosocial behaviors like cooperation and trust into extrinsic motivation, potentially decreasing prosocial behavior after punishment removal (Gächter & Herrmann, 2009). Monetary incentives may also lead individuals to view deception as an acceptable risk. When the benefits of deception outweigh the monetary punishment, individuals may rationalize deception, thereby inadvertently encouraging it. Is there a more promising form of reward and punishment for inhibiting deceptive behavior?

1.1 The Concept and Manipulation of Social Rewards and Punishments

Humans are inherently social beings. From simple verbal communication to exchanges of interests between individuals, interaction with others is essential, and deceptive behavior typically occurs within such interactions (Kohls et al., 2013). Social reward refers to socially desirable outcomes that individuals expect to obtain without material return, including respect, politeness, acceptance, praise, or recognition through behaviors and verbal expressions (Ramirez-Marin & Shafa, 2018). Social reward signifies acceptance and liking by others or groups, eliciting positive emotions such as satisfaction, happiness, and pride (De Cremer & Tyler, 2005). Social punishment involves criticism, exclusion, verbal abuse, gossip, and similar forms of sanction (Kim & Jeong, 2020). Social punishment triggers negative emotions such as loss, depression, anxiety, jealousy, and may activate neural responses similar to physical pain (Beston, 2019).

Previous studies have manipulated social rewards and punishments in various ways. For example, some have used happy or sad faces to represent social rewards and punishments (Wang et al., 2017), others have used upward and downward arrows (Wang et al., 2020), and still others have used varying de-

grees of happy facial expressions to represent social rewards (Spreckelmeyer et al., 2009). Emojis function as independent expressions that convey meaning (Kaye et al., 2017), effectively serving similar functions to nonverbal cues in traditional face-to-face interactions and communicating rich social and emotional information (Boutet et al., 2021; Cherbonnier & Michinov, 2021; Fischer & Herbert, 2021; Hand et al., 2023). According to the Social Information Model (Walther, 1992), emotions recognized from emojis can guide subsequent social behavior (Van Kleef, 2009). Research has shown that emoji cues help individuals express specific emotional states and reinforce the emotional content embedded in communication messages (Kaye et al., 2016). Emojis can simulate facial feedback in face-to-face communication, providing positive or negative evaluative information that influences recipients' emotional responses and information comprehension (Jin et al., 2022; Kaye et al., 2016). Can social rewards and punishments conveyed through emojis influence deceptive behavior? What psychological mechanisms might underlie this influence?

1.2 The Influence of Social Rewards and Punishments on Deception and Its Mediating Mechanism

Social rewards and punishments are crucial motivational factors in human interaction. To obtain social rewards or avoid social punishments, people may forgo or sacrifice their monetary interests to gain opportunities for group inclusion or social interaction (Tamir & Mitchell, 2012; Tamir et al., 2015). Maintaining a positive social image aligns with long-term personal interests (Gintis, 2000), and those who successfully demonstrate generosity typically receive positive returns in reputation and status (Flynn et al., 2006; Hardy & Van Vugt, 2006). Moreover, individuals employ various interpersonal communication strategies to manage their self-image and gain acceptance, recognition, and liking from others (Jones & Pittman, 1982; Berman et al., 2015). Thus, satisfying social needs such as belonging and respect is as vital for human survival as meeting material needs (Romano et al., 2017). Simultaneously, to avoid social punishments like exclusion or reputational damage, individuals are often willing to contribute more money in public goods games (Feinberg et al., 2014; Guala, 2012). The threat of social exclusion is an important means of promoting cooperation. Research has found that in public goods games, when low contributors receive warnings before potential exclusion, most respond with higher contributions to avoid exclusion (Cinyabuguma et al., 2005). These findings indicate that individuals possess strong motivations to pursue social rewards and avoid social punishments. Therefore, we hypothesize that social rewards and punishments can also serve as effective means to inhibit deceptive behavior, reducing deception in interpersonal interactions.

If social rewards and punishments can reduce deception, what psychological mechanisms might explain this effect? Mazar et al. (2008) proposed the theory of self-concept maintenance, which suggests that individuals face psychological conflict when deceiving (benefit from deception vs. maintain positive self-

concept). To resolve this conflict, individuals adjust the degree of deception to avoid excessive threat to their self-concept (e.g., reputation). Reputation is a general impression formed through observing an individual' s historical behavior and represents an intangible asset requiring continuous investment and maintenance (Pfeiffer et al., 2012). Reputation concern—the degree to which individuals care about their own reputation—is an important way to maintain a positive self-concept. In situations involving reputation, individuals typically adjust their behavior to conform to social norms, thereby avoiding negative evaluation and maintaining a positive self-concept (Leary & Kowalski, 1990).

Research has shown that reputation concern is an important psychological factor influencing whether individuals choose to deceive (Russell et al., 2008). The prerequisite for reputation concern is the possibility of social evaluation. Compared to monetary incentives that emphasize economic trade-offs, social rewards and punishments involve more social evaluation processes (Deci, 1971). Social rewards from others may enhance an individual' s reputation, while social punishments from others can damage it. For example, research has found that publicly disclosing donors' names to increase their reputation concern significantly increases charitable donations (Karlan & Mcconnell, 2014). Vabba et al. (2022) found that participants typically adjust their behavior according to the risk of reputational damage, showing significantly less deception when they realize their behavior may be observed. Similarly, Ariely and Gino (2012) found that when individuals' moral standards are reminded or they are in situations where they are observed and evaluated by others, their deceptive behavior decreases significantly. In summary, social rewards and punishments may reduce deception by enhancing individuals' reputation concern. Specifically, social rewards may enhance individuals' positive reputation, thereby reducing their motivation to deceive, while social punishments threaten individuals' reputation, prompting them to return to normative behavior. Therefore, we hypothesize that reputation concern mediates the effect of social rewards and punishments on deceptive behavior.

1.3 The Moderating Role of Social Value Orientation in the Influence of Social Rewards and Punishments on Deception

The influence of social rewards and punishments on deception may be moderated by personality factors. Social value orientation (SVO), as a stable social preference, reflects individuals' tendencies in distributing benefits between themselves and others (Balliet et al., 2009; Grosch & Rau, 2017; Steinel, 2015). SVO can be simplified into two categories: (1) proself orientation (including individualistic and competitive orientations), which emphasizes self-interest over others' interests; and (2) prosocial orientation (including cooperative and altruistic orientations), which focuses on mutual benefits or even prioritizes others' interests (Zhang et al., 2015). Research has shown that SVO significantly influences responses to rewards and punishments. For example, Balliet and Van Lange (2013) found that monetary rewards have stronger incentive effects on cooper-

ation among proself individuals compared to prosocial individuals. Conversely, prosocial individuals show higher sensitivity to social rewards and punishments related to social acceptance and rejection (Li et al., 2020). This sensitivity to social rewards and punishments may cause prosocial individuals to pay more attention to their reputation when facing social incentives, thereby influencing their deceptive behavior. Reputation, as a form of "social currency," is particularly important for prosocial individuals who value social relationships (Milinski et al., 2002). Prosocial individuals typically value relationships with others more highly and tend to maintain a positive reputation to promote trust and cooperation. They are more concerned about their reputation in interactions to ensure positive evaluation from others (Simpson & Willer, 2008). In contrast, proself individuals tend to maximize their own interests rather than build long-term relationships. In situations offering short-term gains, proself individuals are more likely to ignore the risk of reputational damage and show relatively low concern for reputation (Van Lange, 1999). Based on this, we hypothesize that SVO moderates the mediating process through which social rewards and punishments influence deception via reputation concern. Specifically, compared to proself individuals, social rewards and punishments are more likely to trigger reputation concern among prosocial individuals, thereby reducing their deceptive behavior.

In summary, this study uses three experiments to explore the influence of social rewards and punishments on deception and its psychological mechanisms. Experiment 1 examines whether social rewards and punishments can influence deception, Experiment 2 further investigates the mediating role of reputation concern in this relationship, and Experiment 3 examines the moderating role of SVO.

2 Experiment 1: The Effect of Social Rewards and Punishments on Deception

2.1 Participants

Based on the experimental design (single-factor within-subjects) with significance level $\alpha = 0.05$, statistical power 1-$\beta = 0.8$, and medium effect size $p^2 = 0.13$, G*Power 3.1 software calculation indicated that a minimum sample size of 28 participants was required. Therefore, 30 university students were recruited (Mage = 20.43, SD = 1.63; 10 males, 20 females).

2.2 Experimental Design

A single-factor within-subjects design was employed, with reward/punishment type as the independent variable at three levels: social reward/punishment, monetary reward/punishment, and no reward/punishment. The dependent variable was deceptive behavior, measured as the proportion of false information sent by participants in the signaling game task.

2.3 Experimental Task

The signaling game task was used to measure deceptive behavior. This task involves two roles: information sender and information receiver. The sender possesses information, while the receiver can only rely on the sender' s message. This creates an information asymmetry situation that captures the core dilemma of deception: a conflict of interest between sender and receiver. If the sender chooses to send truthful information, they receive lower payoffs while the receiver gains more; if the sender chooses to send false information, they may increase their own payoff while reducing the receiver' s payoff (Zhu et al., 2014).

Participants always served as the information sender. At task onset, four options were presented, each corresponding to different monetary payoffs (e.g., Option 1: sender receives ¥10, receiver receives ¥7; Option 3: sender receives ¥7, receiver receives ¥10; Options 2 and 4 offered zero payoff, see Figure 1 [Figure 1: see original paper]). Two zero-payoff options were included to ensure participants believed that even if they chose to deceive, the receiver—motivated by self-interest—would still select based on the sender' s message. Otherwise, there would be a 50% chance of selecting a zero-payoff option, resulting in zero earnings for both parties and greater loss, thereby convincing participants that the receiver would always follow their recommendation. Participants had to recommend one option to the receiver by sending a message claiming it would yield higher payoff for the receiver. They could either send truthful information to help the receiver gain more or deceive the receiver to secure higher payoff for themselves. Participants were informed that the system had a certain probability of revealing the truthfulness of their message to the receiver, who would then reward or punish them accordingly. If the system did not reveal message truthfulness, the receiver could not know whether the message was truthful, resulting in no reward or punishment feedback.

Following Gneezy (2005), the benefits participants could gain through deception were divided into three levels (high, medium, low profit), as were the losses inflicted on receivers (high, medium, low loss). Amounts below ¥5 were considered low, ¥5-10 medium, and above ¥10 high. For social reward/punishment manipulation, previous studies have used approval/disapproval comments, thumbs up/down gestures, happy/sad faces or cartoon images (Matyjek et al., 2020). This study adopted the method of Wang et al. (2017), using happy or sad emojis as social rewards and punishments. Monetary rewards and punishments were manipulated using "+" or "-" coin images, with the amount set at ¥1—lower than the benefits from deception—to prevent participants from choosing honesty merely for greater profit.

2.4 Experimental Procedure

The experimental task consisted of three blocks (social reward/punishment, monetary reward/punishment, and no reward/punishment), each containing 40 trials. Block order was randomized across participants to balance order effects.

Participants were told they would interact with a randomly matched partner in another laboratory room, though the partner was actually a computer program. Participants were informed that the system had a certain probability of revealing message truthfulness to the receiver; in reality, this probability was set at 80%, though participants were unaware of the exact value. If truthfulness was revealed in a trial, the receiver would reward or punish based on message veracity; if not revealed, the trial proceeded without reward or punishment. Feedback differed across blocks: happy or sad faces in the social block, "+"or "-"¥1 coins in the monetary block, and no feedback in the no-reward/punishment block. Additionally, in 20% of trials across each block, participants' emotions were measured ("To what extent did the receiver' s response make you feel happy? Please select a number: 1 = very unhappy to 9 = very happy). Emotion measurement trials were pseudo-randomly distributed, with 8 fixed truth-revealed trials (20%) in each block selected for emotion measurement.

The trial procedure was as follows: First, a "+" appeared for 800 ms, signaling trial onset. After a 600 ms blank screen, four options were presented, prompting participants to select and send information. Following their choice, the selected option was highlighted for 1200 ms to confirm the selection. Next, participants were informed whether the system revealed message truthfulness to the receiver, with duration randomly varying between 1000-2500 ms to simulate the receiver' s decision time. If truthfulness was revealed, reward or punishment feedback was presented for 1500 ms based on message veracity.

After the experiment, the computer randomly selected one trial from each block, and participants were paid according to their choice and any reward or punishment in that trial. Figure 2 [Figure 2: see original paper] shows the experimental procedure using the social reward/punishment block as an example. The task was presented using PsychoPy software (Peirce, 2009).

2.5 Data Analysis

Traditional data analysis was first conducted on reaction time data. Subsequently, Drift-Diffusion Modeling (DDM) was applied to analyze reaction time data. DDM describes decision-making as a continuous sampling process in which noisy information accumulates from a starting point to a boundary or threshold corresponding to an option, which is then selected (Ratcliff & McKoon, 2008). DDM parameters include drift rate ( ), boundary height (α), starting point bias (z), and non-decision time (τ). Drift rate represents the rate at which evidence for a choice accumulates; stronger preference for an option yields faster information accumulation toward that option. Boundary height α indicates the amount of information required before a response is made. Starting point bias z quantifies prior bias before decision-making. Non-decision time τ reflects other factors affecting reaction time, including stimulus encoding and motor response time (Yuan et al., 2023; Zhang et al., 2020). In this study, DDM used choice and reaction time distributions to describe how participants accumulated evidence for deceptive versus non-deceptive choices. Drift rate quantified the strength

of evidence favoring deception or non-deception obtained through processing payoff information—that is, the degree of value trade-off between deceptive and non-deceptive choices. In this study, choosing deception was coded as 1 and non-deception as 0. Therefore, more positive drift rates indicated stronger preference for deception, while more negative drift rates indicated stronger preference for non-deception. Starting point bias z quantified participants' tendency toward deception/non-deception before accumulating any evidence. Boundary height α quantified the amount of evidence required to make a choice, reflecting decision caution across conditions.

The Python-based HSSM (Hierarchical Sequential Sampling Modeling) package was used for Bayesian parameter estimation of the reaction time-based DDM model. Bayesian estimation allows direct comparison of posterior parameter distributions without relying on traditional frequentist statistics. For two conditions of interest (e.g., drift rates under social reward/punishment vs. no reward/punishment), if the 95% highest density intervals (HDI) of the posterior distributions did not overlap, the difference between conditions was considered credible (Yuan et al., 2023).

2.6 Results

First, a repeated-measures ANOVA on emotion levels across the three conditions (social, monetary, no reward/punishment) revealed no significant differences, $F(2, 58) = 2.99$, $p = 0.058$, $p^2 = 0.09$.

Second, a repeated-measures ANOVA on deception rates across the three conditions showed significant differences, $F(2, 58) = 12.66$, $p < 0.001$, $p^2 = 0.30$. Deception rates were significantly lower under social reward/punishment ($M = 0.34$, $SD = 0.27$) and monetary reward/punishment ($M = 0.34$, $SD = 0.25$) compared to the no-reward/punishment condition ($M = 0.48$, $SD = 0.31$), $t(29) = -3.69$, $p < 0.001$, and $t(29) = -3.40$, $p < 0.001$, respectively. No significant difference emerged between social and monetary conditions, $t(29) = 0.12$, $p = 0.909$ (Figure 3 [Figure 3: see original paper]). An ANCOVA with gender as a covariate still showed a significant main effect of reward/punishment type, $F(2, 28) = 8.37$, $p = 0.001$, $p^2 = 0.374$. These results indicate that, similar to monetary incentives, social rewards and punishments effectively reduce deception, with equivalent effectiveness.

To further understand how social reward/punishment feedback influences subsequent deception, we compared deception rates after receiving social feedback versus no-feedback trials. The post-feedback deception rate was calculated as the number of deceptive choices in the trial immediately following social reward or punishment feedback divided by the total number of post-feedback trials in that block (averaged across all relevant trials, not based on a single trial). The no-feedback deception rate was the number of deceptive choices in no-reward/punishment trials divided by the total number of no-feedback trials in that block. ANOVA revealed significant differences across conditions, $F(2,$

48) = 24.50, p < 0.001, p^2 = 0.51. Compared to the no-reward/punishment condition (M = 0.57, SD = 0.24), deception rates were significantly lower after receiving social reward (M = 0.44, SD = 0.24) or social punishment (M = 0.27, SD = 0.22), t(24) = −2.47, p = 0.021, and t(24) = −7.29, p < 0.001, respectively. Moreover, deception rates after social punishment were significantly lower than after social reward, t(24) = −5.16, p < 0.001 (Figure 4 [Figure 4: see original paper]). Thus, compared to no social feedback, receiving social reward or punishment led participants to prefer sending truthful information in subsequent trials, with social punishment proving more effective than social reward.

A two-way repeated-measures ANOVA on reaction times for honest versus deceptive choices across the three conditions revealed no significant main effect of reward/punishment type, F(2, 44) = 0.51, p = 0.604, p^2 = 0.02, and no significant main effect of choice type, F(1, 22) = 0.44, p = 0.512, p^2 = 0.02. However, the interaction was significant, F(2, 44) = 5.89, p = 0.005, p^2 = 0.21. Simple effects analysis showed that in the no-reward/punishment condition, honest choices (M = 5.31, SD = 2.49) had longer reaction times than deceptive choices (M = 4.40, SD = 2.18), t(24) = 3.43, p = 0.002. In contrast, no significant differences emerged between honest and deceptive reaction times in the social or monetary reward/punishment conditions.

DDM fitting results showed good model convergence, with R-hat values below 1.05. Model comparison indicated that the model allowing all four parameters (, α, z, τ) to vary across reward/punishment conditions was optimal, showing the highest expected log predictive density for Leave-One-Out cross-validation (elpd_{loo}). Parameter analysis revealed that drift rates under social reward/punishment (M = −0.19, 95% HDI [−0.23, −0.14]) and monetary reward/punishment (M = −0.18, 95% HDI [−0.23, −0.14]) were significantly lower than under no reward/punishment (M = −0.03, 95% HDI [−0.07, 0.02]), indicating that both incentive types prompted individuals to accumulate evidence favoring non-deception. Additionally, non-decision times were significantly longer under social (M = 0.95, 95% HDI [0.89, 1.01]) and monetary (M = 1.03, 95% HDI [0.98, 1.10]) conditions compared to the no-reward/punishment condition (M = 0.75, 95% HDI [0.69, 0.80]) (Figure 5 [Figure 5: see original paper]).

3 Experiment 2: The Mediating Role of Reputation Concern

Experiment 1 demonstrated that social rewards and punishments reduce deceptive behavior similarly to monetary incentives. But what mechanisms underlie this effect? Social rewards and punishments involve social evaluation processes that may heighten reputation concern and thereby influence deception. Experiment 2 examined whether reputation concern mediates the relationship between social incentives and deception.

3.1 Participants

To examine mediation effects, 60 university students were recruited (Mage = 20.25, SD = 1.74; 20 males, 40 females).

3.2 Experimental Design

This experiment used a single-factor within-subjects design, with reward/punishment type (social, monetary, no reward/punishment) as the independent variable, deception rate as the dependent variable, and reputation concern as the mediator.

3.3 Materials and Task

We adapted the reputation concern questionnaire from Wu, Balliet, and Lange (2016) to measure participants' concern for their reputation. The original questionnaire comprised six items with good reliability and validity. Based on the nature of social rewards and punishments in this study, two items were selected to measure reputation concern using a 9-point scale (1 = strongly disagree, 9 = strongly agree), with the second item reverse-scored. Higher scores indicated greater reputation concern (Appendix A).

3.4 Procedure

The procedure was identical to Experiment 1, with participants completing the signaling game task and completing the reputation concern measure after each block.

3.5 Results

First, a repeated-measures ANOVA on emotion levels across the three conditions showed no significant differences, $F(2, 118) = 2.05$, $p = 0.133$, $p^2 = 0.03$.

Second, a repeated-measures ANOVA on deception rates revealed significant differences across conditions, $F(2, 118) = 12.99$, $p < 0.001$, $p^2 = 0.18$. Deception rates were significantly lower under social reward/punishment ($M = 0.35$, $SD = 0.25$) and monetary reward/punishment ($M = 0.33$, $SD = 0.25$) compared to the no-reward/punishment condition ($M = 0.46$, $SD = 0.30$), $t(59) = -3.87$, $p < 0.001$, and $t(59) = -4.33$, $p < 0.001$, respectively. No significant difference emerged between social and monetary conditions, $t(59) = 1.01$, $p = 0.315$ (Figure 6 [Figure 6: see original paper]). An ANCOVA with gender as a covariate still showed a significant main effect of reward/punishment type, $F(2, 116) = 11.45$, $p < 0.001$, $p^2 = 0.17$. These results replicate Experiment 1, showing that social rewards and punishments reduce deception similarly to monetary incentives.

To analyze the mediating effect of reputation concern, we used PROCESS 3.1 to test the path "presence of social reward/punishment → reputation concern

→ deception behavior" (Model 4, 5,000 bootstrap samples). Presence of social reward/punishment was coded as 1 and no reward/punishment as 0, with reputation concern scores as the mediator and deception rate as the outcome. The results showed a significant indirect effect, Mean bootstrapped indirect effect ab = -0.12, BootSE = 0.04, 95% CI [LLCI = −0.22, ULCI = −0.05], which did not include zero (Figure 7 [Figure 7: see original paper]).

We also examined whether reputation concern mediated the effect of monetary incentives on deception. Using presence of monetary reward/punishment (monetary = 1, no reward/punishment = 0) as the independent variable, reputation concern as the mediator, and deception rate as the outcome, the 95% confidence interval for the indirect effect was [LLCI = −0.10, ULCI = 0.02], which included zero, indicating a non-significant mediation effect (Figure 8 [Figure 8: see original paper]).

A two-way repeated-measures ANOVA on reaction times for honest versus deceptive choices revealed no significant main effect of reward/punishment type, $F(2, 94) = 0.12$, $p = 0.892$, $p^2 = 0.002$, and no significant main effect of choice type, $F(1, 47) = 0.07$, $p = 0.788$, $p^2 = 0.002$. However, the interaction was significant, $F(2, 94) = 6.72$, $p = 0.002$, $p^2 = 0.13$. Simple effects analysis showed that in the no-reward/punishment condition, honest choices ($M = 4.69$, $SD = 2.51$) had longer reaction times than deceptive choices ($M = 4.21$, $SD = 2.28$), $t(47) = 2.98$, $p = 0.005$. In the monetary condition, honest choices ($M = 4.35$, $SD = 1.69$) had shorter reaction times than deceptive choices ($M = 4.78$, $SD = 2.41$), $t(47) = 2.04$, $p = 0.047$. In the social condition, no significant difference emerged between honest and deceptive reaction times, $t(47) = 0.75$, $p = 0.459$.

DDM fitting results showed good model convergence (R-hat < 1.05). Model comparison indicated that the model allowing all four parameters to vary across conditions was optimal, with the highest elpd_{loo} value. Parameter analysis revealed that drift rates under social reward/punishment ($M = −0.16$, 95% HDI [−0.19, −0.13]) and monetary reward/punishment ($M = −0.21$, 95% HDI [−0.24, −0.17]) were significantly lower than under no reward/punishment ($M = −0.05$, 95% HDI [−0.08, 0.02]), indicating that both incentive types prompted evidence accumulation favoring non-deception. Additionally, non-decision times were significantly longer under social ($M = 0.92$, 95% HDI [0.86, 0.97]) and monetary ($M = 0.94$, 95% HDI [0.88, 0.99]) conditions compared to the no-reward/punishment condition ($M = 0.76$, 95% HDI [0.71, 0.80]) (Figure 9 [Figure 9: see original paper]).

4 Experiment 3: The Moderating Role of Social Value Orientation

Experiment 2 found that social rewards and punishments increase reputation concern and thereby reduce deception, whereas monetary incentives do not. This process may be moderated by personality traits, particularly social value orientation. Prosocial individuals show greater sensitivity to social rewards and

punishments (acceptance or rejection) compared to proself individuals. Experiment 3 examined whether SVO moderates the mediating process through which social incentives influence deception via reputation concern.

4.1 Participants

A total of 193 participants completed the SVO slider measure, yielding 125 proself individuals and 68 prosocial individuals (Mage = 22.83, SD = 3.75; 96 males, 97 females).

4.2 Experimental Design

A 2 (SVO: proself vs. prosocial) × 3 (reward/punishment type: social, monetary, no reward/punishment) mixed design was employed, with reward/punishment type as the within-subjects factor and SVO as the between-subjects factor. The dependent variable was the proportion of deceptive behavior.

4.3 Materials and Task

(1) Social Value Orientation Measurement
The SVO slider measure developed by Murphy, Ackermann, and Handgraaf (2011) was used. This measure consists of 15 items: six primary items and nine secondary items. Each item presents nine allocation options (e.g., ¥150 distribution), from which participants select their preferred distribution (see Appendix B). Primary items assess SVO angle or type (altruistic, prosocial, individualistic, competitive), with larger angle values indicating greater concern for others' payoffs. The maximum value of 61.39° represents pure altruism, while the minimum of -16.26° represents pure competition (Zhang et al., 2015). Secondary items assess inequality aversion. This study used the two most common SVO types: (1) proself orientation (individualistic and competitive) and (2) prosocial orientation (altruistic and prosocial).

(2) Reputation Concern Measurement
Identical to Experiment 2.

4.4 Procedure

First, participants completed the SVO slider measure to determine their SVO angle. Participants with SVO angles $> 22.45°$ were classified as prosocial, while those $< 22.45°$ were classified as proself (Murphy et al., 2011). The subsequent procedure mirrored Experiment 2, using the signaling game task with reputation concern measured after each block.

4.5 Results

A two-way ANOVA on deception rates revealed significant main effects of reward/punishment type, $F(2, 382) = 40.66$, $p < 0.001$, $p^2 = 0.183$, and SVO, $F(1, 191) = 30.99$, $p < 0.001$, $p^2 = 0.14$. Prosocial individuals (M = 0.35,

SD = 0.29) showed significantly lower deception rates than proself individuals (M = 0.57, SD = 0.33). The SVO × reward/punishment type interaction was significant, $F(2, 382) = 5.44$, $p = 0.005$, $p^2 = 0.03$.

Simple effects analysis showed that for proself individuals, the main effect of reward/punishment type was significant, $F(2, 191) = 19.66$, $p < 0.001$, $p^2 = 0.17$. Multiple comparisons revealed no significant difference between social reward/punishment (M = 0.60, SD = 0.32) and no reward/punishment (M = 0.65, SD = 0.31), $t(191) = -1.76$, $p = 0.079$. However, monetary reward/punishment (M = 0.46, SD = 0.34) significantly reduced deception compared to no reward/punishment, $t(191) = -5.98$, $p < 0.001$.

For prosocial individuals, the main effect of reward/punishment type was also significant, $F(2, 191) = 23.70$, $p < 0.001$, $p^2 = 0.20$. Both social reward/punishment (M = 0.32, SD = 0.27) and monetary reward/punishment (M = 0.30, SD = 0.26) significantly reduced deception compared to no reward/punishment (M = 0.45, SD = 0.32), $t(191) = -5.95$, $p < 0.001$, and $t(191) = -6.19$, $p < 0.001$, respectively. No significant difference emerged between social and monetary conditions, $t(191) = 0.81$, $p = 0.420$ (Figure 10 [Figure 10: see original paper]).

We re-examined the mediating effect of reputation concern in the social incentive condition. Using presence of social reward/punishment as the independent variable, reputation concern as the mediator, and deception rate as the outcome, the indirect effect was significant, 95% CI [LLCI = −0.092, ULCI = −0.032], not including zero. In contrast, the mediating effect of reputation concern in the monetary incentive condition was non-significant, 95% CI [LLCI = −0.012, ULCI = 0.031], including zero.

We then used PROCESS Model 8 to test whether SVO moderated the mediating effect of reputation concern (moderated mediation). Results showed that social reward/punishment positively predicted reputation concern ($\beta = 0.88$, $p < 0.001$) and negatively predicted deception ($\beta = -0.05$, $p < 0.001$). SVO positively predicted reputation concern ($\beta = 0.82$, $p < 0.001$) and negatively predicted deception ($\beta = -0.20$, $p < 0.001$). Critically, the interaction between social reward/punishment and SVO significantly predicted reputation concern ($\beta = 0.90$, $p = 0.030$) (Table 1). For proself individuals, the indirect effect of reputation concern was non-significant, ab = -0.018, BootSE = 0.021, z = -0.82, p = 0.415, 95% CI [LLCI = −0.061, ULCI = 0.022]. For prosocial individuals, the indirect effect was significant, ab = -0.070, BootSE = 0.016, z = -4.29, $p < 0.001$, 95% CI [LLCI = −0.104, ULCI = −0.040].

DDM fitting results showed good convergence (R-hat < 1.05). Model comparison indicated that the model allowing all four parameters to vary across conditions was optimal, with the highest elpd_{loo} value. Parameter analysis revealed that drift rates under social reward/punishment (M = −0.13, 95% HDI [−0.14, −0.11]) and monetary reward/punishment (M = −0.22, 95% HDI [−0.24, −0.20]) were significantly lower than under no reward/punishment (M

= 0.05, 95% HDI [0.03, 0.06]), indicating that both incentive types prompted evidence accumulation favoring non-deception. Additionally, non-decision times were significantly longer under social (M = 0.81, 95% HDI [0.79, 0.83]) and monetary (M = 0.71, 95% HDI [0.70, 0.73]) conditions compared to no reward/punishment (M = 0.76, 95% HDI [0.71, 0.80]) (Figure 12 [Figure 12: see original paper]).

5 General Discussion

This study systematically investigated the influence of social rewards and punishments on deceptive behavior and its mediating and moderating mechanisms through three experiments using a signaling game task. The findings demonstrate that social rewards and punishments reduce deception similarly to monetary incentives, with social punishment proving more effective than social reward. Reputation concern mediates the effect of social incentives on deception, and social value orientation moderates this mediating process.

5.1 The Effect of Social Rewards and Punishments on Deception

This study found that social and monetary incentives have similar effects in reducing deception, both significantly decreasing deceptive behavior. This aligns with previous research showing that social rewards enhance prosocial behavior. When individuals follow social norms and exhibit prosocial behavior, they often receive social affirmation such as smiles, praise, recognition, acceptance, support, and care, which satisfies interpersonal needs and belongingness, thereby reinforcing prosocial behavior (Lü et al., 2021; Kringelbach & Rolls, 2003). Research indicates that anticipating others' verbal praise can increase prosocial behavior (Ellingsen & Johannesson, 2008), and some individuals are willing to forgo monetary gains to obtain social recognition and acceptance (Shore & Heerey, 2011). Conversely, social punishment triggers negative emotions and activates neural responses similar to physical pain (Beston, 2019). To avoid social exclusion and reputational damage, individuals exhibit more prosocial behavior.

Further analysis showed that after receiving social reward or punishment, participants were more likely to send truthful information in subsequent trials, with social punishment demonstrating stronger inhibitory effects than social reward. This likely stems from social punishment' s stronger emotional activation, which more effectively suppresses norm-violating behavior. These results support the negativity bias effect—the tendency for negative information to have greater impact than positive information. Research shows that negative events produce more enduring and intense effects at emotional, behavioral, and cognitive levels, with negative stimuli (e.g., social punishment) rapidly and significantly activating emotional responses (Baumeister et al., 2001; Rozin & Royzman, 2001). Compared to social reward, social punishment more readily triggers shame, guilt, or anxiety, producing more lasting effects on psychological states (Eisenberger et al., 2003; Kujawa et al., 2015).

Additionally, the significant interaction between reward/punishment type and choice type on reaction times revealed that in the no-reward/punishment condition, honest choices took longer than deceptive choices, whereas no such difference emerged in the social or monetary conditions. Debey et al. (2015) analyzed reaction times to reveal the cognitive mechanisms of honesty and deception, finding that honesty typically involves longer reaction times, suggesting it relies on a slower, reflective system requiring careful consideration of moral and social norms, while deception may depend more on a fast, intuitive system. In the absence of incentives, the reflective system dominates, resulting in longer honest response times (Greene et al., 2008). However, explicit motivation to gain rewards or avoid punishment simplifies decision-making and reduces reflective system involvement (Gino & Ariely, 2012), enabling faster honest choices. DDM results support this cognitive process: drift rates under social and monetary conditions were significantly lower (more negative) than under no-reward/punishment, indicating stronger preference for honest options. Thus, both social and monetary incentives prompt individuals to accumulate evidence favoring honesty, leading to more honest choices.

5.2 The Mediating Role of Reputation Concern

This study found that reputation concern mediates the effect of social incentives on deception. Specifically, social rewards and punishments increase reputation concern, thereby reducing deception, whereas monetary incentives do not operate through this mechanism. These results support indirect reciprocity theory and altruistic reputation theory. Indirect reciprocity theory posits that reputation is key to explaining human altruism (Nowak & Sigmund, 2005); under reputation systems, individuals attend to their reputation and others' impressions, making reputation a behavioral evaluation standard (Leimar & Hammerstein, 2001). Altruistic reputation theory suggests that people build positive reputations through publicly visible altruistic acts (e.g., helping, generous donations), which confer social and survival advantages (Trivers, 1971). Compared to monetary incentives, social incentives involve more social evaluation processes, heightening reputation concern and influencing deception. Reputation concern refers to the attentional target activated by others' evaluations of one' s behavior, prompting behavioral adjustment to gain good reputation and increase future long-term benefits (Sperber & Baumard, 2012). Research shows that when individuals can gain reputational incentives through indirect reciprocity, they are more willing to cooperate (Milinski et al., 2006). Furthermore, individuals adjust their cooperative behavior based on group members' evaluations to gain acceptance and recognition (Brady et al., 2017; Sommerfeld et al., 2007). When people expect their reputation information to be disseminated, they attend more to others' opinions and contribute more; when receiving negative evaluations, they realize these may harm their reputation, increasing contributions and adherence to group norms (Wu et al., 2016).

5.3 Social Value Orientation Moderates the Effect of Social Incentives

This study also found that SVO moderates the effect of social incentives on deception. For proself individuals, monetary incentives significantly reduced deception, but social incentives had limited effect. Conversely, for prosocial individuals, both social and monetary incentives significantly reduced deception. Moreover, SVO moderated the mediating process through which social incentives affect deception via reputation concern. Previous research shows that proself individuals prioritize self-interest, while prosocial individuals prioritize group interests and exhibit more cooperative, prosocial behavior in social dilemmas (Liu & Hao, 2011). SVO influences reputation concern, which in turn moderates deceptive behavior in response to incentives. Reputation serves as an important mechanism for prosocial individuals to maintain social connections; for them, reputation is not only external recognition but also an internal driver for maintaining relationships and promoting cooperation (De Cremer & Van Lange, 2001). Prosocial individuals are more attentive to social and others' expectations, showing higher reputation concern in interpersonal interactions to ensure positive evaluation from others (Van Lange, 1999; Simpson & Willer, 2008; Cameron & Payne, 2011). Reputation management theory suggests that individuals manipulate their image in others'eyes to maintain social status. SVO influences sensitivity to reputation gains and losses; compared to proself individuals, prosocial individuals are more likely to regulate their behavior through social incentives to avoid reputational loss from dishonesty (Leary & Kowalski, 1990). This suggests that prosocial individuals may be more norm-compliant in social incentive contexts to protect their reputation and reduce deception, while proself individuals may be less responsive to reputational threats from social incentives, showing higher deception tendencies.

5.4 Theoretical Contributions and Practical Implications

This study reveals unique mechanisms through which social incentives inhibit deception compared to monetary incentives. Although both effectively reduce deception, their pathways differ significantly. Monetary incentives primarily reduce deception through external incentives or increased potential costs (Gneezy, 2005), whereas social incentives rely on individuals' reputation concern to activate internal moral responsibility, thereby inhibiting deception (Feinberg et al., 2014). While monetary incentives influence decision-making through material motivation, their effects are often constrained by external conditions. Social incentives, based on reputation concern, represent a more enduring internal driving force (Feinberg et al., 2014). This study found that even when using emojis as social incentive signals, individuals still showed reputation concern and reduced deception, confirming the broad applicability of social incentives and demonstrating that even lightweight social signals can influence moral decision-making. This provides a new perspective for future research on social behavior and highlights the importance of social incentive mechanisms in maintaining moral norms.

These findings offer important implications for effectively curbing deception. The results indicate that social rewards and punishments, as a more promising and low-cost approach, can effectively inhibit deception. In educational contexts, educators should recognize the role of social incentives in moral education. Traditional educational management often relies on material rewards to motivate honesty. However, this study shows that social incentives also effectively reduce unethical behavior. Since adolescents' values are still developing, over-reliance on monetary rewards may foster short-term utilitarian orientation while neglecting cultivation of internal moral responsibility (Frey & Jegen, 2001). Social incentives can have more profound effects on students' value formation. Research shows that monetary rewards may undermine intrinsic motivation, making behavior more dependent on external incentives rather than moral beliefs (Deci et al., 1999). In contrast, social rewards satisfy needs for belonging and self-identity (Ryan & Deci, 2000), strengthening students' internal identification with integrity and social norms, thereby promoting more enduring moral behavior (Feinberg et al., 2014). Therefore, emphasizing social incentives in educational management can help shape students' moral values and reduce over-reliance on material rewards, making honest behavior more sustainable.

5.5 Limitations and Future Directions

This study has several limitations that suggest directions for future research. First, social rewards and punishments encompass various forms, including respect, recognition, praise, acceptance, criticism, opposition, exclusion, and gossip (Ramirez-Marin & Shafa, 2018; Kim & Jeong, 2020). However, the specific scope of social incentives remains unclear, lacking systematic structural analysis. This study only used smile and sad emojis as manipulations; the effects of other forms of social incentives on deception require further investigation to fully understand the mechanisms. Second, this study focused only on SVO as an individual difference variable, while other individual differences may also moderate the effects. For example, individuals high in social reward/punishment sensitivity show greater attention to social evaluation and reputation maintenance (Gino & Pierce, 2009), and social reward processing differs between depressed and healthy populations (Li et al., 2024). Future research should explore these moderating effects to more comprehensively understand how individual characteristics influence the effectiveness of social incentives.

Finally, deceptive behavior typically occurs in specific social contexts (Gneezy, 2005; Mazar et al., 2008). Future research should examine how different social contexts moderate the effects of social incentives. For instance, social exclusion may threaten individuals' sense of belonging and influence their responses to social incentives (Eisenberger et al., 2003). Additionally, situational publicity may affect the effectiveness of social incentives. In public situations, individuals show stronger self-awareness and greater activation in self-related neural regions (Somerville et al., 2013), making them more sensitive to evaluation and more attentive to reputation. Therefore, future research should explore how social

contextual factors shape the influence patterns of social incentives.

6 Conclusion

Compared to no-incentive conditions, both social and monetary rewards and punishments reduce deceptive behavior, with social punishment proving more effective than social reward. Social incentives increase reputation concern, which in turn reduces deception. Individuals' social value orientation moderates this mediating process, such that social incentives heighten reputation concern among prosocial individuals, thereby reducing their deception.

References

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of Public Economics, 1*(3-4), 323-338.

Ariely, D., & Gino, F. (2012). Cheating, self-signaling, and groups: Reminding people of their moral standards. *Social Psychological and Personality Science, 3*(3), 344-352.

Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations, 12*(4), 533-547.

Balliet, D., & Van Lange, P. A. (2013). Trust, conflict, and cooperation: A meta-analysis. *Psychological Bulletin, 139*(5), 1090-1112.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323-370.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy, 76*(2), 169-217.

Behnk, S., Barreda-Tarrazona, I., & Garcia-Gallego, A. (2018). Punishing liars-How monitoring affects honesty and trust. *PLoS One, 13*(10), e0205420.

Berman, J. Z., Levine, E. E., Barasch, A., & Small, D. A. (2015). The Braggart' s dilemma: On the social rewards and penalties of advertising prosocial behavior. *Journal of Marketing Research, 52*(1), 90-104.

Beston, P. (2019). *The effect of social rewards and punishments on learning and cooperative decision-making* (Master' s thesis). Bangor University, United Kingdom.

Boutet, I., LeBlanc, M., Chamberland, J. A., & Collin, C. A. (2021). Emojis influence emotional communication, social attributions, and information processing. *Computers in Human Behavior, 119*, 106722.

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313-7318.

Cameron, C. D., & Payne, B. K. (2011). Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *Journal of Personality and Social Psychology, 100*(1), 1-15.

Carr, Z. M., Solbu, A., & Frank, M. G. (2019). Why methods matter: Approaches to the study of deception and considerations for the future. In T. Docan-Morgan (Ed.), *The Palgrave Handbook of Deceptive Communication* (pp. 267-286). Palgrave Macmillan.

Cherbonnier, A., & Michinov, N. (2021). The recognition of emotions beyond facial expressions: Comparing emoticons specifically designed to convey basic emotions with other modes of expression. *Computers in Human Behavior, 118*, 106689.

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics, 89*(8), 1421-1435.

De Cremer, D., & Van Lange, P. A. (2001). Why prosocials exhibit greater cooperation than proselfs: The roles of social responsibility and reciprocity. *European Journal of Personality, 15*(S1), S5-S18.

De Cremer, D., & Tyler, T. R. (2005). Managing group behavior: The interplay between procedural justice, sense of self, and cooperation. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology, 37*(5), 151-218.

Debey, E., Ridderinkhof, R. K., De Houwer, J., De Schryver, M., & Verschuere, B. (2015). Suppressing the truth as a mechanism of deception: Delta plots reveal the role of response inhibition in lying. *Consciousness and Cognition, 37*, 148-159.

Deci, E. L. (1971). Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology, 18*(1), 105-115.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin, 125*(6), 627-668.

Depaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74-118.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science, 302*(5643), 290-292.

Ellingsen, T., & Johannesson, M. (2008). Pride and Prejudice: The human side of incentive theory. *American Economic Review, 98*(3), 990-1008.

Feinberg, M., Willer, R., & Schultz, M. (2014). Gossip and ostracism promote cooperation in groups. *Psychological Science, 25*(3), 656-664.

Fischer, B., & Herbert, C. (2021). Emoji as affective symbols: Affective judgments of emoji, emoticons, and human faces varying in emotional content. *Frontiers in Psychology, 12*, 645173.

Flynn, F. J., Reagans, R. E., Amanatullah, E. T., & Ames, D. R. (2006). Helping one' s way to the top: Self-monitors achieve status by helping others and knowing who helps whom. *Journal of Personality and Social Psychology, 91*(6), 1123–1137.

Frey, B. S., & Jegen, R. (2001). Motivation crowding theory. *Journal of Economic Surveys, 15*(5), 589-611.

Gächter, S., & Herrmann, B. (2009). Reciprocity, culture and human cooperation: Previous insights and a new cross-cultural experiment. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1518), 791–796.

Gino, F., & Pierce, L. (2009). Dishonesty in the name of equity. *Psychological Science, 20*(9), 1153–1160.

Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology, 206*(2), 169–179.

Gneezy, U. (2005). Deception: The role of consequences. *American Economic Review, 95*(1), 384–394.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154.

Grosch, K., & Rau, H. A. (2017). Gender differences in honesty: The role of social value orientation. *Journal of Economic Psychology, 62*, 258–267.

Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences, 35*(1), 1–15.

Hand, C. J., Kennedy, A., Filik, R., Pitchford, M., & Robus, C. M. (2023). Emoji identification and emoji effects on sentence emotionality in ASD-diagnosed adults and neurotypical controls. *Journal of Autism and Developmental Disorders, 53*(6), 2514–2528.

Hardy, C. L., & Van Vugt, M. (2006). Nice guys finish first: The competitive altruism hypothesis. *Personality and Social Psychology Bulletin, 32*(10), 1402–1413.

Jin, Y. C., Deng, C. L., Wu, P., Lin, X., Zheng, P. X., & An, J. X. (2022). Emoji image symbol' s social function and application. *Advances in Psychological Science, 30*(5), 1062–1077.

Jones, E. E., & Pittman, T. S. (1982). Toward a general theory of strategic self-presentation. *Psychological Perspectives on the Self, 1*(1), 231–262.

Karlan, D., & Mcconnell, M. A. (2014). Hey look at me: The effect of giving circles on giving. *Journal of Economic Behavior & Organization, 106*, 402–412.

Kaushik, M., Singh, V., & Chakravarty, S. (2022). Experimental evidence of the effect of financial incentives and detection on dishonesty. *Scientific Reports, 12*(1), 2680.

Kaye, L. K., Wall, H. J., & Malone, S. A. (2016). "Turn that frown upside-down" : A contextual account of emoticon usage on different virtual platforms. *Computers in Human Behavior, 60*, 463-467.

Kaye, L. K., Malone, S. A., & Wall, H. J. (2017). Emojis: Insights, affordances, and possibilities for psychological science. *Trends in Cognitive Sciences, 21*(2), 66-68.

Kim, J., & Jeong, B. (2020). Expecting social punishment facilitates control over a decision under uncertainty by recruiting medial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 15*(11), 1260-1270.

Kohls, G., Perino, M. T., Taylor, J. M., Madva, E. N., Cayless, S. J., Troiani, V., Price, E., Faja, S., Herrington, J. D., & Schultz, R. T. (2013). The nucleus accumbens is involved in both the pursuit of social reward and the avoidance of social punishment. *Neuropsychologia, 51*(11), 2062-2069.

Kringelbach, M. L., & Rolls, E. T. (2003). Neural correlates of rapid reversal learning in a simple model of human social interaction. *NeuroImage, 20*(2), 1371-1383.

Kujawa, A., Proudfit, G. H., Kessel, E. M., Dyson, M., Olino, T., & Klein, D. N. (2015). Neural reactivity to monetary rewards and losses in childhood: Longitudinal and concurrent associations with observed and self-reported positive emotionality. *Biological Psychology, 104*, 41-47.

Leary, M. R., & Kowalski, R. M. (1990). Impression management: A literature review and two-component model. *Psychological Bulletin, 107*(1), 34-47.

Leimar, O., & Hammerstein, P. (2001). Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 268*(1468), 745-753.

Li, J., Sun, Y., Yang, Z. L., & Zhong, Y. P. (2020). Social value orientation modulates the processing of social rewards for self: Evidence from ERPs study. *Acta Psychologica Sinica, 52*(6), 786-800.

Li, S. J., Tang, Y. Y., & Zhang, D. D. (2024). Neural mechanism of monetary and social reward processing in healthy and depressed populations. *Journal of Psychological Science, 47*(6), 1317-1327.

Liu, C. J., & Hao, F. (2011). Social value orientation and cooperation in asymmetric social dilemmas. *Acta Psychologica Sinica, 43*(4), 432-441.

Lü, F. Y., Tan, J. B., Xu, P. F., Xiong, X. L., Jin, Z. H., & Gao, D. G. (2021). The social reward and the neurocognitive mechanism of social reward processing. *Chinese Journal of Applied Psychology, 27*(3), 189-203.

Matyjek, M., Meliss, S., Dziobek, I., & Murayama, K. (2020). A multidimensional view on social and non-social rewards. *Frontiers in Psychiatry, 11*, 818.

Mazar, N., & Ariely, D. (2006). Dishonesty in everyday life and its policy implications. *Journal of Public Policy & Marketing, 25*(1), 117–126.

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.

Milinski, M., Semmann, D., & Krambeck, H. J. (2002). Reputation helps solve the 'tragedy of the commons' . *Nature, 415*(6870), 424–426.

Milinski, M., Semmann, D., Krambeck, H. J., & Marotzke, J. (2006). Stabilizing the Earth's climate is not a losing game: Supporting evidence from public goods experiments. *Proceedings of the National Academy of Sciences, 103*(11), 3994–3998.

Mulder, L. B., Dijk, E. V., Cremer, D. D., & Wilke, H. A. M. (2006). Undermining trust and cooperation: The paradox of sanctioning systems in social dilemmas. *Journal of Experimental Social Psychology, 42*(2), 147–162.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. (2011). Measuring social value orientation. *Judgment and Decision Making, 6*(8), 771–781.

Nagin, D. S., & Pogarsky, G. (2003). An experimental investigation of deterrence: Cheating, self-serving bias, and impulsivity. *Criminology, 41*(1), 167–194.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291–1298.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*, 343.

Pfeiffer, T., Tran, L., Krumme, C., & Rand, D. G. (2012). The value of reputation. *Journal of the Royal Society Interface, 9*(76), 2791–2797.

Ramirez-Marin, J. Y., & Shafa, S. (2018). Social rewards: The basis for collaboration in honor cultures. *Cross Cultural & Strategic Management, 25*(1), 53–69.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation, 20*(4), 873–922.

Romano, A., Balliet, D., Yamagishi, T., & Liu, J. H. (2017). Parochial trust and cooperation across 17 societies. *Proceedings of the National Academy of Sciences, 114*(48), 12702–12707.

Rosenbaum, S. M., Billinger, S., & Stieglitz, N. (2014). Let's be honest: A review of experimental evidence of honesty and truth-telling. *Journal of Economic Psychology, 45*, 181–196.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320.

Russell, Y. I., Call, J., & Dunbar, R. I. M. (2008). Image scoring in great apes. *Behavioural Processes, 78*(1), 108–111.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68–78.

Shore, D. M., & Heerey, E. A. (2011). The value of genuine and polite smiles. *Emotion, 11*(1), 169–174.

Simpson, B., & Willer, R. (2008). Altruism and indirect reciprocity: The interaction of person and situation in prosocial behavior. *Social Psychology Quarterly, 71*(1), 37–52.

Somerville, L. H., Jones, R. M., Ruberry, E. J., Dyke, J. P., Glover, G., & Casey, B. J. (2013). The medial prefrontal cortex and the emergence of self-conscious emotion in adolescence. *Psychological Science, 24*(8), 1554–1562.

Sommerfeld, R. D., Krambeck, H. J., Semmann, D., & Milinski, M. (2007). Gossip as an alternative for direct observation in games of indirect reciprocity. *Proceedings of the National Academy of Sciences, 104*(44), 17435–17440.

Spreckelmeyer, K. N., Krach, S., Kohls, G., Rademacher, L., Irmak, A., Konrad, K., ···& Gründer, G. (2009). Anticipation of monetary and social reward differently activates mesolimbic brain structures in men and women. *Social Cognitive and Affective Neuroscience, 4*(2), 158–165.

Sperber, D., & Baumard, N. (2012). Moral reputation: An evolutionary and cognitive perspective. *Mind & Language, 27*(5), 495–518.

Steinel, W. (2015). Social value orientation and deception: Are proselfs liars? *Current Opinion in Psychology, 6*, 86–90.

Tamir, D. I., & Mitchell, J. P. (2012). Disclosing information about the self is intrinsically rewarding. *Proceedings of the National Academy of Sciences, 109*(21), 8038–8043.

Tamir, D. I., Zaki, J., & Mitchell, J. P. (2015). Informing others is associated with behavioral and neural signatures of value. *Journal of Experimental Psychology: General, 144*(6), 1114–1123.

Trivers, R. L. (1971). The evolution of reciprocal altruism. *The Quarterly Review of Biology, 46*(1), 35–57.

Vabba, A., Porciello, G., Panasiti, M. S., & Aglioti, S. M. (2022). Interoceptive influences on the production of self-serving lies in reputation risk conditions. *International Journal of Psychophysiology, 177*, 34–42.

Van Lange, P. A. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and*

Social Psychology, 77(2), 337-349.

Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science, 18*(3), 184-188.

Walther, J. B. (1992). Interpersonal effects in computer-mediated interaction: A meta-analysis of social and antisocial communication. *Human Communication Research, 19*(4), 159-182.

Wang, D., Liu, T., & Shi, J. (2017). Development of monetary and social reward processes. *Scientific Reports, 7*(1), 1-9.

Wang, Z., Li, Q., Nie, L., & Zheng, Y. (2020). Neural dynamics of monetary and social reward processing in social anhedonia. *Social Cognitive and Affective Neuroscience, 15*(9), 991-1003.

Wu, J., Balliet, D., & Lange, P. A. M. V. (2016). Reputation management: Why and how gossip enhances generosity. *Evolution & Human Behavior, 37*(3), 193-201.

Yuan, B., Wang, X. P., Yin, J., & Li, W. Q. (2023). The role of cross-situational stimulus generalization in the formation of trust towards face: A perspective based on direct and observational learning. *Acta Psychologica Sinica, 55*(7), 1099-1114.

Zhang, Y. H., Li, H., & Wu, Y. (2020). The application of computational modelling in the studies of moral cognition. *Advances in Psychological Science, 28*(7), 1042-1055.

Zhang, Z., Zhang, F., Yuan, S., Guo, F. B., & Wang, Y. W. (2015). Psychometric analysis of the SVO slider measure in Chinese cultural context. *Studies of Psychology and Behavior, 13*(3), 404-409.

Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., King-Casas, B., & Hsu, M. (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience, 17*(10), 1319-1321.

Appendix A: Reputation Concern Scale

1. During the decision-making process in the previous task, I considered how the other person would evaluate me as a person.
   1 = completely disagree, 2 = disagree, 3 = somewhat disagree, 4 = slightly disagree, 5 = neutral, 6 = slightly agree, 7 = somewhat agree, 8 = agree, 9 = completely agree

2. During the decision-making process in the previous task, I didn' t care at all about the other person' s response to me.
   1 = completely disagree, 2 = disagree, 3 = somewhat disagree, 4 = slightly

disagree, 5 = neutral, 6 = slightly agree, 7 = somewhat agree, 8 = agree, 9 = completely agree

Appendix B: Social Value Orientation Slider Measure

In this task, please imagine that you are randomly paired with another person (referred to as "TA"). You do not know each other and will never meet. All your choices will be completely confidential. You will make a series of decisions about resource allocation between yourself and TA. For each question below, please select your preferred outcome distribution from the nine allocation options provided. Your decisions will determine the amount of money each of you receives. There are no right or wrong answers; they simply reflect personal preferences.

Appendix C: Data Analysis Tables

Table C1 Mixed ANOVA results for SVO × reward/punishment type in Experiment 3

Effect	df	F	p	p^2
SVO	1, 191	30.99	< 0.001	0.14
Reward/Punishment Type	2, 382	40.66	< 0.001	0.183
SVO × Reward/Punishment Type	2, 382	5.44	0.005	0.03

Table C2 Descriptive statistics (M ± SD) for SVO across reward/punishment types in Experiment 3

SVO	Social	Monetary	No Reward/Punishment
Proself	0.60 ± 0.32	0.46 ± 0.34	0.65 ± 0.31
Prosocial	0.32 ± 0.27	0.30 ± 0.26	0.45 ± 0.32

Note: Figure translations are in progress. See original paper for figures.