

# The impact of lying aversion and prosociality on cheating\*

Daniel Parra<sup>†</sup>

[\[Click here for the latest version of the paper\]](#)

## Abstract

In some situations, when an agent faces the decision of whether to lie or not, their decision only affects the agent themselves. However, in other situations, lying can also benefit third parties as well as the liar. In these latter situations, there is a trade-off between the disutility of being a liar and the utility of a prosocial lie. This paper studies the interaction of two behavioral traits that are well established in the behavioral economics literature: the aversion to lying and prosociality. I first present a model that incorporates heterogeneous lying costs and prosociality as a part of the individual's preferences. I show that individuals are more likely to lie when they have prosocial motives in addition to their self-interested motives. Using online experiments, I show that participants are more dishonest when their lies benefit others and themselves. More importantly, I present evidence that, on average, the prosocial motive facilitates lying even for people with relatively high lying costs.

JEL Codes: C91, D02, D90.

Keywords: Cheating; Dishonesty; Prosociality; Psychological lying costs.

This version: August 1, 2022

---

\*I am grateful to Kai Barron, Tilman Fries, Uri Gneezy, Johann Graf Lambsdorff, Jeanne Hagenbach, Agne Kajackaite, Johannes Leutgeb, Cesar Mantilla, Robert Stüber, Christian Traxler, Roel van Veldhuizen, Yuliet Verbel, and audiences at WZB, BEBES, the 2021 ESA Global Online Meetings, and the Rady School of Management for helpful comments. This research used generic funds provided by the WZB Berlin Social Science Center. The usual disclaimer applies.

<sup>†</sup>WZB Berlin Social Science Center and Berlin School of Economics. E-mail: [daniel.parra@wzb.eu](mailto:daniel.parra@wzb.eu).

# 1 Introduction

The domain of lying is rife with ethical and legal implications, evoking strong emotions among those on all sides of debates on issues such as lying in political campaigns, legal testimony, and daily life. The economic sphere is not an exception to this debate. For instance, in his classical model, [Akerlof \(1970\)](#) highlights dishonesty's central role in asymmetric information markets. He argues that dishonesty leads to market failures in the presence of information asymmetries. In particular, he shows that the social damage generated by dishonesty includes the direct cost to the deceived individual and other indirect costs, such as the erosion of the incentives to produce high-quality goods.

In these classical models, it has been assumed that individuals lie whenever they have monetary incentives. However, evidence on lying has shown that, even if there is no punishment for lying and there are personal benefits from doing so, the proportion of people in the population who decide to lie is moderate ([Abeler et al., 2019](#)). This moderate extent of people's dishonesty can be explained if some individuals have some aversion to lying. Theoretical models usually represent this lying aversion as the psychological cost of lying.<sup>1</sup> This idea of psychological costs means that some individuals do not lie because they dislike violating their internal moral norms of being honest or because they at least want to appear honest.

However, human beings are social beings and depend on cooperation with others to achieve many of their goals. Being a social being makes people also have social preferences in addition to their pure self-interested incentives. Hence, the psychological lying cost might be reduced when one can lie to benefit themselves and others. In other words, people might feel that telling Pareto improving lies is not as bad as telling pure selfish lies. This effect might occur because people might use prosociality to make lying easier.

To illustrate how lying aversion and prosociality may interact, imagine a car broker who is selling a used car on behalf of the owner in exchange for a commission. The broker has incentives to lie about the actual quality of the car. They earn a higher commission if they sell the car for a higher price but incur psychological lying costs if they lie. However, all else equal, they may also feel less unethical by lying about the car's quality because the lie benefits the owner. A sales representative faces a similar trade-off between prosociality and lying aversion when they can lie to get a team bonus the CEO promised the sales team for reaching a certain threshold. This duality is present in several situations, for instance, when a taxpayer misreports their family earnings and hence gains benefits that also extend to their family themselves or when a lie told by an underage drinker to procure alcohol

---

<sup>1</sup>What I refer to as psychological cost can capture both an intrinsic distaste of lying and potential image concerns (having a dislike for being perceived as a liar).

benefits the liar and their friends. In general, when a lie benefits others, two effects go in opposite directions. Lying aversion makes telling a lie costly, and conversely, prosociality generates some utility. This utility generated by prosociality has been vastly studied and is a well-established behavioral trait shown by different studies such as [Andreoni \(1990\)](#); [Andreoni and Miller \(2002\)](#); [Charness and Rabin \(2002\)](#); [Bénabou and Tirole \(2006\)](#); [Ariely et al. \(2009\)](#); [DellaVigna et al. \(2012\)](#).

In this paper, I compare the impact of lying aversion and prosociality on cheating decisions. To do so, I study lying in a two-person game. If at least one partner lies in the game, both benefit from the lie, but there are no additional gains if both lie. More specifically, participants will report a random draw which is private information. The random draw can be either 1 or 0. In each couple, if at least one group member reports drawing a 1 instead of a 0, both members will earn a higher monetary reward. I use this two-person game because it creates a situation in which people can avoid lying by relying on others' incentives or they can tell a prosocial lie. Hence, in this strategic situation, there is a trade-off. On the one hand, lying aversion implies that people are primarily honest when others are likely to lie on their behalf. On the other hand, prosociality implies that people are prone to lying when their lies benefit others. The paper's main contribution is to present a theoretical framework that explains the interaction of these behavioral traits in the individual utility function and then uses an experimental design that allows me to disentangle them.

I first present the theoretical framework incorporating heterogeneous psychological lying costs and prosociality in individual preferences. I include prosociality as a parameter that reduces the lying costs. Then, I use experimental data to assess the model's predictions empirically. To provide more privacy to participants than the standard die roll game ([Fischbacher and Föllmi-Heusi, 2013](#)), I ran an experiment in which participants chose one color out of five in their minds. Then, they randomly drew one of the five colors. They had to report whether the color chosen in their minds was the same as the color drawn. The random draw was known by the individual but not by their partner. Therefore, the random draw was not relevant for the monetary payoffs but just the individuals' reports.

The experiment had four treatments pre-registered in the AEA RCT Registry.<sup>2</sup> In the first treatment, called AVOID, two participants reported the result of the private random draw sequentially. Both got a higher monetary payoff if at least one individual reported that the colors matched. Therefore, the first mover could avoid the cost of lying by telling the truth. In a second treatment, called NO AVOID, the first mover also reported whether the colors matched. However, the second mover was asked to report the color they picked on their mind before the random draw; the computer program then reported truthfully

---

<sup>2</sup>The registration number is AEARCTR-0007214.

whether the colors matched. In other words, by design, in NO AVOID it is common knowledge that the second player's report will necessarily be truthful. This variation makes it more difficult to avoid the lying costs in NO AVOID than in AVOID. To disentangle the impact of the positive externality from the psychological cost of lying, I included a third treatment called NO EXTERNALITY. In this treatment, the first mover report does not benefit the second mover. Finally, I used a fourth treatment called SIMULTANEOUS where both players reported whether their colors matched without knowing their partner's report. In SIMULTANEOUS, participants could no longer be sure that if they reported that their colors matched, their action would give the other participant a direct benefit. Put another way, if both participants said that their colors matched, no one could claim that their action benefited the other player, given that the payoffs would have been the same if the report had been different. Hence, in SIMULTANEOUS, there was a trade-off between strategically lying about whether their colors matched to secure a higher payoff and losing the prosocial motive because both reported matching colors. The prosocial motive might have been lost in this situation because, in my theoretical framework, I assume that participants hold a consequentialist view of prosociality and then care about the consequence of their actions rather than the intention behind them.

The experiment results show that the second mover in AVOID lied less when the first-mover reported that the colors matched. Even if this result shows that the first mover has a strategic advantage when trying to avoid the cost of lying, surprisingly, I did not find any difference in the lying rates of the first movers in AVOID and NO AVOID. Consequently, having high and similar lying rates in AVOID and NO AVOID suggests that prosociality might be a strong driver of lying behavior even in the presence of lying costs. This conjecture is supported by the result of NO EXTERNALITY, where I found that first-movers lie more in NO AVOID than in NO EXTERNALITY, indicating that people lied more when they benefited others as well as themselves. This result is in line with [Wiltermuth \(2011\)](#), [Gino et al. \(2013\)](#), and [Levine and Schweitzer \(2015\)](#). Furthermore, combining these results, I show that, on average, prosocial lying outweighs lying aversion. In other words, even relatively honest people tend to lie more often in situations where they benefit others and themselves. However, this result does not imply that lying aversion does not matter in the presence of prosocial motives because even in AVOID and NO AVOID, more than half of the people did not lie. The final result of the experiment was that lying was not lower in AVOID than in SIMULTANEOUS. This result might suggest that individuals care about the actual benefit they generate in others rather than the prosocial intention of their actions which is consistent with consequentialist prosocial lying.

This paper contributes to the fast-growing literature on lying behavior. This body of literature has argued that the deviation from the world full of liars that the classical economic

theory assumed can be explained by people's disutility when they are dishonest. [Kajackaite and Gneezy \(2017\)](#) show that individuals follow a cost-benefit analysis in which they evaluate the psychological cost of lying and the incentives to lie. [Gneezy et al. \(2018\)](#), [Abeler et al. \(2019\)](#), and [Khalmetski and Sliwka \(2019\)](#) present evidence that individuals indeed have psychological costs of lying that can be divided into intrinsic costs of lying and reputation. [Dufwenberg and Dufwenberg \(2018\)](#) also explains lying behavior by the cost of lying. However, they argue that this cost increases proportionally to how the individual is perceived to cheat, making the lying costs extrinsic. I contribute to this literature by showing that individuals do seek to avoid lying when someone lies on their behalf, as shown by the behavior of the second mover in AVOID. However, I found evidence that prosocial lying decreases lying costs. Hence, I show that even if the psychological costs of lying make people lie less, prosocial lying might reduce the impact of lying aversion. This result does not imply that lying aversion is not a significant motivator but broadens our knowledge about how some factors interact with the psychological cost of lying.

The second strand of literature to which this paper is closely related is the literature on collaborative lying (for a survey see [Leib et al., 2021](#)). These studies use games in which participants play in groups, and each member of the group needs to lie to increase everyone's earnings (e.g. [Conrads et al., 2013](#); [Weisel and Shalvi, 2015](#); [Muehlheusser et al., 2015](#); [Kocher et al., 2018](#); [Rilke et al., 2021](#)). In other words, in this body of research, lies are strategic complements. Hence, collaborative lying research focuses on situations where coordination in dishonesty is central; thus, it is impossible to rely on others not to lie. The collaborative lying games suggest that social preferences make people lie more, implying that cooperation enhances dishonest behaviors. Although I also use a group setting, I study a different situation in which dishonesty is not complementary but a substitute. Therefore, individuals can avoid being dishonest by relying on others. Additionally, in my setting, prosociality rather than cooperation is the social preference that plays an important role. I show that individuals use prosociality to justify lying even when collaboration is unnecessary to increase their payoffs.

Finally, this paper also relates to the studies that analyze how positive externalities impact lying behavior. [Wiltermuth \(2011\)](#) and [Levine and Schweitzer \(2015\)](#) have shown that people are more likely to lie if their lies benefit others. [Levine and Schweitzer \(2015\)](#) showed that prosocial lying enhances trust in group settings, which may explain why people are willing to lie for others. I add to this body of evidence by showing that prosociality is strong enough to outweigh lying aversion for a significant part of the population. Nevertheless, the effect of prosociality on dishonesty vanishes when the actual impact on others' payoffs is uncertain. This result provide evidence in favor of a consequentialist view of prosocial lies. To put it another way, individuals' utility depends on the actual consequences of their

actions for others rather than on the intention of benefiting them. Therefore, this paper sheds some light on the mechanisms explaining why people are more likely to lie when their lies are Pareto improving.

The paper proceeds as follows. Section 2 presents a theoretical model of lying with heterogeneous psychological lying costs and prosociality. It also presents four treatments that disentangle lying aversion and prosociality alongside their respective hypotheses. Section 3 explain the details of the online experiment and its procedures. Section 4 presents evidence of an experimental study that tests the hypotheses of the model. Section 5 discusses the findings from the experiment and interprets them using the benchmark model presented in Section 2. Additionally, it uses the model to compare the results with a collaborative lying game. Section 6 concludes.

## 2 Theoretical framework, experimental design, and hypotheses

### 2.1 Individual's preferences

The lying models presented by [Dufwenberg and Dufwenberg \(2018\)](#), [Gneezy et al. \(2018\)](#), [Abeler et al. \(2019\)](#), and [Khalmetski and Sliwka \(2019\)](#) are fundamental to understand why people lie in situations without externalities.<sup>3</sup> I use them as a starting point and add strategic interaction to study the willingness to avoid the lying costs and lie prosocially. Specifically, I study situations where individuals interact in dyads. I denote the members of each dyad as  $P_i$  where  $i \in \{1, 2\}$ .

Players play a binary lying game. I use a binary game because I am interested in whether people lie, not in the size of the lie. Therefore, the standard die-roll game ([Fischbacher and Föllmi-Heusi, 2013](#)) creates an unnecessary noise that decreases the statistical power and does not add anything to the paper's main question. Specifically, each player draws a state  $x_i \in \mathcal{X} = \{0, 1\}$ . The probability of  $x_i = 0$  is 0.8 and the probability of  $x_i = 1$  is 0.2. I use these probabilities because, in the experiments, they will generate more players drawing 0 than in a coin toss; therefore, more individuals will face the situation where they can lie to improve their payoffs. Players send a report  $r_i \in \mathcal{X}$ . The players' payoffs are interdependent. If at least one player,  $P_1$  or  $P_2$ , reports 1, each of them will get a monetary payoff  $v_h$ . If both report 0, they will get  $v_l$ . To ease the notation, I normalize  $v_l$  to zero and  $v_h$  to 1. In this context, lies are under the category of Pareto White Lies ([Erat and Gneezy,](#)

---

<sup>3</sup>These models make use of psychological game theory to model behavior. They use the experimenter as an observer that affects the individual utility. Hence, they study strategic situations with one player making decisions and a third party who does not take any particular action but can affect the decision maker through beliefs.

2011) because they help others and benefit the liar.

Individuals' preferences depend on three elements that determine the willingness to lie or tell the truth. First, they get utility from the monetary payoff  $v_i \in \{v_h, v_l\}$  that depends on their report. All else equal, they have extrinsic incentives to report 1 regardless of their actual random draw  $x_i$ . Second, individuals dislike lying. Lying aversion is represented by some psychological costs ( $c_i$ ) that they incur when they misreport their random draw (Gneezy et al., 2018; Abeler et al., 2019; Khalmetski and Sliwka, 2019). The psychological lying cost includes the intrinsic costs of lying and the image costs. Let these costs, represented by  $c_i$ , be distributed among the population according to  $c_i \sim U[0, \bar{c}]$ . Hence, the cumulative density function of  $c_i$  is  $F(c_i) = \frac{c_i}{\bar{c}}$ . The heterogeneity in the psychological lying costs captures the fact that some people are more morally inclined than others.

Third, I formally include prosociality in the utility function inspired by the insights provided by Wiltermuth (2011), Gino et al. (2013), and Levine and Schweitzer (2015). In particular, individuals get some satisfaction ( $\theta$ ) when they benefit others with their report, i.e. when they generate a positive externality. In the case of  $\theta$ , I impose  $0 \leq \theta \leq 1$  so that the prosocial lying utility is non-negative but never higher than the utility of the own monetary reward. The standard models omit prosociality in the decision on whether to lie or not. In my model, when an agent lies and benefits others, the prosocial lie reduces individuals' psychological lying cost. Moreover, I assume that  $c_i(r_i = x_i) = 0$ , and  $1 + \theta < \bar{c}$ . The last condition is used to rule out the uninteresting case where all individuals have a psychological cost of lying so small that everyone lies. With this assumption, the individual with the highest lying cost will always tell the truth.

The remaining question at this point is: does the utility derived from prosocial lies depend only on actions or also on consequences? Some models use warm-glow and altruism to explain giving (Andreoni, 1990) which implies that people care about their intentions to give. Conversely, I assume that individuals' utility depends on the outcome of their actions. Therefore, the positive externality reduces the cost of lying only if the marginal benefit of one's report on the partner is 1. In other words, the utility from prosociality ( $\theta$ ) when their partner reports 1 is 0 regardless of one's report.<sup>4</sup> Arguably, it is more difficult to justify a dishonest behavior with prosociality when in the absence of one's report, the payoff of the other would be the same. With these three elements, I represent individuals' preferences by the following function:

$$U_i(x_i, r_i, r_j) = r_i + r_j - r_i r_j - 1_{x_i \neq r_i} (c_i - \theta r_i (1 - r_j)) \quad (1)$$

---

<sup>4</sup>The alternative way to incorporate the positive externality's impact would be to assume that only by lying and reporting 1 they feel good. This view would represent deontological prosocial lying where the intention of benefiting others matters regardless of the actual consequence.

## 2.2 Treatments

### 2.2.1 Treatment 1: AVOID

In the main treatment, AVOID, I study a two stage lying game where players' lies are substitutes.  $P_1$  draws  $x_1 \in \mathcal{X}$  and sends a report  $r_1 \in \mathcal{X}$  to  $P_2$ . After learning  $r_1$ ,  $P_2$  draws  $x_2 \in \mathcal{X}$  and sends a report  $r_2 \in \mathcal{X}$ . Note that  $x_i$  is only known by  $P_i$ , but not by the other player.

I use backward induction to analyze the strategic context of the game. When  $P_1$  reports 1,  $P_2$  has no strict incentives to lie. Whereas, when  $P_1$  reports 0,  $P_2$ 's best response is 1 if  $1 + \theta > c_i$ . That is, if the second-mover considers that the combination of the monetary incentives and the satisfaction of benefiting others exceed the costs of lying, they will report 1 regardless of  $x_2$ . Importantly, in this game there is no downward lying in equilibrium. If individuals draw 1 and report 0 they incur the cost of lying without getting the monetary payoffs or the benefit of the positive externality. So, in equilibrium individuals only lie if they draw  $x_i = 0$  by reporting  $r_i = 1$ .

Let  $\hat{c}_i$  be the lying cost threshold where individuals are indifferent between lying or not. This threshold for  $P_2$ , when  $P_1$  reports 0, is  $\hat{c}_2(r_1 = 0) = 1 + \theta$ . Hence, the probability that  $P_2$  lies after  $P_1$  reports 0 is the expected proportion of players with  $\hat{c}_2(r_1 = 0) < 1 + \theta$ , namely:

$$F(\hat{c}_2(r_1 = 0)) = \frac{1 + \theta}{\bar{c}} \quad (2)$$

The decision of  $P_1$  depends on their beliefs about  $P_2$ 's report.  $P_1$  lies if  $E(U_1(r_1 = 1)) > E(U_1(r_1 = 0))$ . Let  $b_0$  be  $P_1$ 's belief that  $P_2$  reports 1 after  $r_1 = 0$ , and  $b_1$  be  $P_1$ 's belief that  $P_2$  reports 1 after  $r_1 = 1$ . Then, taking into account the utility presented in (1),  $P_1$  lies if  $1 - c_i + \theta(1 - b_1) > b_0$ . This implies that the lying threshold that divides those who lie from those who do not in AVOID is:

$$\hat{c}_1 = 1 - b_0 + \theta(1 - b_1) \quad (3)$$

In equilibrium, the beliefs about the response of  $P_2$  are  $b_1 = 0.2$ , given that this is the probability of drawing 1, and  $b_0 = 0.2 + 0.8 \frac{1+\theta}{\bar{c}}$ . Thus, replacing  $b_1$  and  $b_2$  in equation  $\hat{c}_1$ , I get that the lying threshold at equilibrium for  $P_1$  in AVOID is:

$$\hat{c}_1 = 0.8 \left( 1 + \theta - \frac{1 + \theta}{\bar{c}} \right) \quad (4)$$

### 2.2.2 Treatment 2: NO AVOID

In a second treatment, NO AVOID, I remove the  $P_1$ 's capacity of relying on  $P_2$  possibility to lie by imposing  $x_2 = r_2$ . To do so, in NO AVOID, participants with the role of  $P_2$  do not have the possibility of reporting their random draw, but the computer will have the information to do it on their behalf. This procedure is common knowledge. The payoff structure is the same as in AVOID, and even if a computer makes the report, a human participant bears the consequences in terms of payoffs. Thus, with this procedure, I ensure that  $P_1$  has an objective probability of  $r_2$ . This feature implies that  $b_1 = b_0 = 0.2$ . Then, using (3) and substituting the new values of  $b_1$  and  $b_2$ , the threshold of the lying cost at equilibrium for  $P_1$  in NO AVOID is:

$$\hat{c}_1 = 0.8(1 + \theta) \quad (5)$$

Comparing the lying cost thresholds presented in (4) and (5), it follows that more  $P_1$  will lie when they can not rely on  $P_2$ 's incentives to lie. This result holds because the utility by prosociality and the maximum lying cost are non-negative.

**Hypothesis 1** (No cost avoidance). *In NO AVOID, the proportion of  $P_1$  who lie will be higher compared with AVOID.*

### 2.2.3 Treatment 3: NO EXTERNALITY

In a third treatment, I investigate the role of the positive externality on  $P_1$ 's decision. I use the same structure as in NO AVOID but change the payoff scheme to eliminate the benefit on others, so that in this treatment lies are no longer Pareto White Lies but pure selfish lies. I keep  $P_1$ 's monetary payoffs identical as in NO AVOID but make  $P_2$ 's monetary payoffs only dependent on  $x_2$ . In particular, in NO EXTERNALITY,  $P_2$  gets 1 only if  $x_2 = 1$  and 0 otherwise. The variation in this treatment implies that in the utility function  $\theta = 0$ . Therefore, the threshold of the lying cost at equilibrium for  $P_1$  in NO EXTERNALITY is:

$$\hat{c}_1 = 0.8 \quad (6)$$

From the comparison between (5) and (6) it follows that lying is more pronounced in NO AVOID than in NO EXTERNALITY.

**Hypothesis 2** (Positive externality). *In NO EXTERNALITY, the proportion of  $P_1$  who lie will be less compared with NO AVOID.*

Given that in NO EXTERNALITY lying decreases compared with NO AVOID, one question remaining is whether NO EXTERNALITY accounts for the same effect than AVOID. However, this effect depends on  $\bar{c}$ . When  $\bar{c}$  is lower than 1, lying will be higher in NO EXTERNALITY than in AVOID. This means that the comparison in lying rates between AVOID and NO EXTERNALITY will depend on the proportion of people with high lying cost in the population.

Until this point, I have presented the main treatments that allow me to assess the impact of lying aversion and prosocial lying on the preferences to lie. To sum up, in NO AVOID the probability of  $P_2$  reporting 1 is fixed at 0.2 (in contrast to AVOID where it was a subjective probability). So,  $P_1$  has more room to avoid the lying cost in AVOID than in NO AVOID, but the prosocial motive is still present in both conditions. Therefore, NO EXTERNALITY lets me assess the role of prosocial lying. Table 1 illustrates how the experimental design isolates each potential explanation allowing me to assess each motive.

**Table 1. Comparison of lying aversion and prosociality across sequential treatments.**

	AVOID	NO AVOID	NO EXTERNALITY
<b>Avoid the Lying Costs</b> $P(r_2 = 1   r_1 = 0)$	$0.2 + b_0$	0.2	0.2
<b>Prosociality</b>	✓	✓	×

Note: the row *Avoid the Lying Cost* refers to how likely is to effectively avoid the lying cost while getting the high payoff. For AVOID it uses  $b_0$  to represent the subjective probability  $P_1$  attributes to  $P_2$  reporting 1.

#### 2.2.4 Treatment 4: SIMULTANEOUS

The last treatment, SIMULTANEOUS, uses the same payoffs structure as in AVOID but participants report simultaneously instead of sequentially. Playing sequentially allows  $P_1$  to transmit their action  $r_1$  to  $P_2$  and gives some strategic advantage to  $P_1$ . In contrast, in SIMULTANEOUS participants need to act without any information about the partner's actual decision. In the utility function presented in (1), I assume that lies that benefit others generate some utility represented by  $\theta$ . However,  $\theta$  only counts if the benefited individual does not report 1. This assumption implies that individuals use consequentialist norms when lying for others, where the action itself does not matter but only the consequence on others payoffs.

In SIMULTANEOUS, players are symmetric and no information is learned before deciding. Hence, I do not use  $b_0$  and  $b_1$ , but define  $b_{ij}$  as the belief of  $P_i$  that  $P_j$  reports 1. As in AVOID,  $P_i$  lies if  $E(U_i(r_i = 1)) > E(U_i(r_i = 0))$ . That is,  $P_i$  lies if  $1 + \theta(1 - b_{ij}) - c_i > b_{ij}$ ,

which leads to the lying threshold  $\hat{c}_i = (1 - b_{ij})(1 + \theta)$ . In equilibrium,  $b_{ij} = 0.2 + 0.8\frac{\hat{c}_i}{\bar{c}}$ . Thus, I plug  $b_{ij}$  in the threshold equation to get  $\hat{c}_i = (0.8 - 0.8\frac{\hat{c}_i}{\bar{c}})(1 + \theta)$ . It follows that the lying threshold in SIMULTANEOUS is:

$$\hat{c}_1 = 0.8 \left( \frac{0.8\bar{c}(1 + \theta)}{1 + 0.8(1 + \theta)} \right) \quad (7)$$

The resulting lying threshold in SIMULTANEOUS presented in (7) needs to be compared with (4). However, this comparison is not as trivial as in the other treatments. Individuals have two competing motives when deciding whether to lie in SIMULTANEOUS and AVOID. On the one hand, they hope to be able to rely on their partner's incentives and avoid the psychological cost of lying. On the other hand, they have the prosocial motive when lying for others and then can use it to decrease their cost of lying. The first motive, lying aversion, implies that  $P_1$ 's motive to lie out of own-payoff consideration is stronger in SIMULTANEOUS than in AVOID because of sequentiality. However, it is more difficult for  $P_i$  to use the prosocial motive in SIMULTANEOUS than in AVOID because consequential prosocial lying implies that reporting 1 only increases individuals' payoffs if their partner reports 0. In other words, in SIMULTANEOUS an individual  $P_i$  may be willing to lie and use the prosociality to decrease the cost of lying, but  $P_j$  is likely doing the same, and none of them gets  $\theta$  which can be anticipated for both players and leads to no one lying.

By observing the lying thresholds in (4) and (7), one can see that determining which motive dominates the other depends on the combination of  $\theta$  and  $\bar{c}$ . To understand this relation, I calculate numerically the values of  $\theta$  and  $\bar{c}$  that imply the same lying rates in AVOID and SIMULTANEOUS. Figure 1 shows that when  $\theta$  is high, the lying threshold is higher in AVOID than in SIMULTANEOUS.<sup>5</sup> Hence, lying rates are higher in AVOID than in SIMULTANEOUS. Conversely, if  $\theta$  is low enough, it is more likely that the cost avoidance motive plays a central role, and thus lying would be higher in SIMULTANEOUS than in AVOID (shaded area in Figure 1). As Figure 1 shows, lying can be higher or lower in AVOID compared with SIMULTANEOUS. Thus, in order to have a hypothesis that can be empirically tested, I have as the null hypothesis that individuals will lie more in SIMULTANEOUS than in AVOID under the conjecture that the motive of lying aversion is strong enough to make individuals more willing to avoid their costs of lying than benefiting their partner.<sup>6</sup>

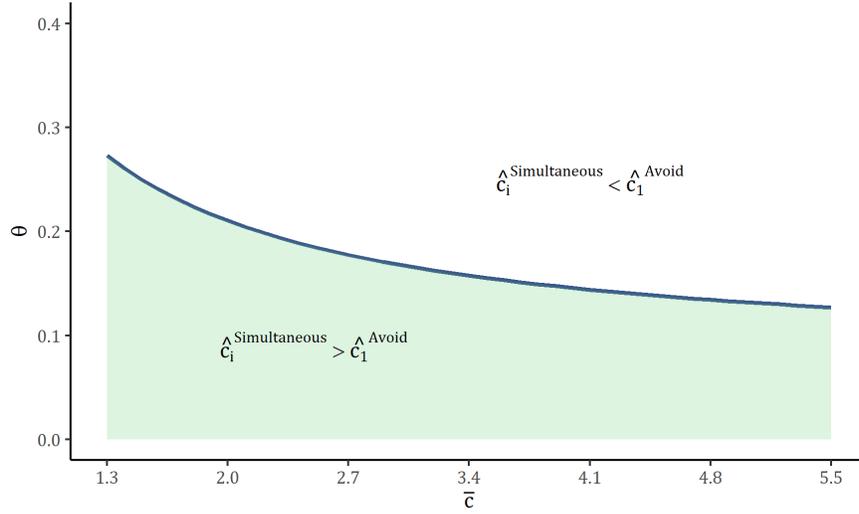
**Hypothesis 3** (No cost avoidance SIMULTANEOUS). *In SIMULTANEOUS, the proportion of  $P_1$  who lie will be more compared with AVOID.*

Table 2 summarizes the decisions each player has to make across the four conditions,

<sup>5</sup>Note that a value of 0.3 means that the utility generated by the positive externality is equal to a 30% of the utility generated by the monetary payoff.

<sup>6</sup>All the hypotheses were preregistered in the AEA RCT Registry.

**Figure 1. Comparison of Lying Thresholds in AVOID and SIMULTANEOUS**



the payoff function, and the hypotheses based on the model. Finally, it is important to note that direct treatment comparisons are only possible in the following pairs: AVOID-NO AVOID, AVOID-SIMULTANEOUS, NO AVOID-NO EXTERNALITY. In the comparisons AVOID-NO EXTERNALITY, NO AVOID-SIMULTANEOUS, and NO EXTERNALITY-SIMULTANEOUS, more than one variable changes.

**Table 2. Summary of the actions, payoffs, and hypotheses in each treatment**

	$P_1$	$P_2$	Payoffs	$H_0$
AVOID	reports $r_1$	learns $r_1$ and then reports $r_2$	$v_i = \begin{cases} 0 & \text{if } r_i = r_j = 0 \\ 1 & \text{otherwise} \end{cases}$	-
NO AVOID	as in AVOID	learns $r_1$ but the report is made by the computer.	as in AVOID	$P_1$ lies more than in AVOID
NO EXTERNALITY	as in NO AVOID	as in NO AVOID	$v_1 = \begin{cases} 0 & \text{if } r_i = r_j = 0 \\ 1 & \text{otherwise} \end{cases}$ $v_2 = \begin{cases} 0 & \text{if } r_2 = 0 \\ 1 & \text{otherwise} \end{cases}$	$P_1$ lies less than in NO AVOID
SIMULTANEOUS	Players make simultaneous decisions		as in AVOID	$P_1$ lies less than in AVOID

### 3 The mind-cheating game

The treatments presented in section 2.2 were first implemented in an online experiment where the experimenter has the information about the random draw and the report. In this *Observed Game*, I used a cheating game where participants drew a Black card or an Orange card and then reported their color. A more detailed presentation of this experiment and the main result is in Appendix A. The procedure used in the *Observed Game* allows identifying lying at the individual level and, in principle, could provide a richer data set to explore. However, in this experiment, lying rates were meager, making it difficult to find any treatment difference. Hence, I changed the procedure to generate the random draw.

#### 3.1 Overview and design

One potential reason why the *Observed Game* presented minimal lying rates, making detecting treatment differences difficult, is that people might be concerned about the random draw's observability. Gneezy et al. (2018) and Fries et al. (2021) show that, in laboratory experiments, the observability of the random draw decreases lying. In the *Observed Game*, maybe because it was an online experiment, and given that Prolific emphasizes the importance that their participants respond to everything honestly, this effect was exacerbated.<sup>7</sup> To assess whether behavior changes when lying is not observed, I designed a second study where the random draw is private and not observed, not even by the experimenter.

In this second study, I use a mind-cheating game<sup>8</sup> in which participants choose one color out of five in their minds (see colors in Figure 2).<sup>9</sup> Then, they draw a color from a deck of cards presented on their computer's screen. The deck of cards contains two cards for each one of the colors. Participants then report whether the color they drew from the deck is the same as their mentally chosen color. If participants want to report that the colors match, they report *Yes*; otherwise, they report *No*. Thus, in this game, *Yes* represents  $x_i = 1$  and *No* represents  $x_i = 0$ . For the payoffs, the rewards I use are  $v_h = \pounds 2.5$  and  $v_l = \pounds 0.3$ . In this study, the state of nature is in participants' minds, so I can only compare distributions of groups based on the known theoretical distribution. However, I cannot identify whether an individual lies or not. One advantage of the mind-cheating game is

---

<sup>7</sup>For instance, when people register in Prolific, they have to complete a study before participating in further studies, which includes the following statement: "...we want to build a world where people and organisations can make important decisions based on trustworthy data and solid evidence. We can't build that world without your contribution: The data you provide, combined with your honesty, your integrity and your effort, is a precious piece of the research puzzle. And together, those pieces help advance human knowledge."

<sup>8</sup>Mind games were previously implemented using die rolls (Jiang, 2013; Shalvi and De Dreu, 2014; Potters and Stoop, 2016; Kajackaite and Gneezy, 2017; Dimant et al., 2020) or coin tosses (Shalvi et al., 2012; Garbarino et al., 2019).

<sup>9</sup>The colors were chosen such that people with colorblindness can see five different colors.

that independent of the color chosen by participants, the probability of matching is always 0.2, which is the same as drawing *Orange* in the *Observed Game*.

**Figure 2. Colors used in mind game**



In the experiment, I use the treatments presented in Table 2. Specifically, in AVOID,  $P_1$  reports to  $P_2$  whether the colors match or not. Once  $P_2$  learns  $r_1$ , they follow the same sequence of decisions: think a color, draw a color from a deck of cards, and report whether the colors match. In NO AVOID,  $P_1$ 's decisions are the same, but  $P_2$  do not select their card in their mind, but they selected it from a list presented on their screens. Then, they draw a color from a deck of cards. Finally, using the selected color and the drawn color, the computer reports whether the colors match or not. Participants know that the computer's report will always be truthful. In NO EXTERNALITY, decisions are identical to NO AVOID, and the variation is that  $P_2$ 's payoffs only depend on whether their selected color and their drawn color match regardless of  $P_1$ 's payoffs. Finally, in SIMULTANEOUS, both participants think of a color, draw a color, and report at the same time whether the colors match.

While  $P_2$  is reporting, I elicit  $P_1$ 's beliefs about  $P_2$ 's report. I use a mechanism proposed by Karni (2009) and implemented experimentally first by Mobius et al. (2011) which allows eliciting probabilities in an incentive-compatible way. Specifically, I use a similar implementation as the one proposed by Coffman (2011). Participants are asked to guess the whether  $P_2$  reports *Yes* or *No* and then ask how likely they think their guess is correct. This procedure allows me to elicit the probability of the  $P_2$  reporting *Yes*. Participants are told that they do not need to read the instructions about the mechanism or understand it if they do not want to. I use this option to reduce the risk of people leaving the experiment because of the complexity of the mechanism. They can, however, click on a button to see the detailed explanation.<sup>10</sup>

Specifically, the elicitation mechanism is based on robots that can guess on behalf of the participants. There are 100 robots, each with integer probability between 1 and 100 of correctly guessing  $P_2$ 's report. A robot from this interval is drawn randomly, and it can guess on the participant's behalf with an accuracy level determined by its number. Robot 1 is accurate 1% of the time; robot 2 is accurate 2% of the time, all the way up to the robot that is accurate 100% of the time. The reported likelihood of their guess being correct is used as an "accuracy threshold." That is, if the robot has an accuracy greater than or equal to the

<sup>10</sup>From the total of participants in  $P_1$  role, 31.79% clicked once in the info button, 1.2% clicked twice, and 0.17% click three times.

threshold, the robot guesses  $r_2$  for  $P_1$ . If the robot has an accuracy less than the threshold,  $P_1$ 's guess is submitted. If the guess is correct, whether it is the participant's or the robot's, it gives a payoff of £0.3.

### 3.2 Procedures

I pre-registered the experiment in AEA RCT Registry under the number AEARCTR-0007214. I calculated the power of the target sample size using computer simulations. I used a minimum detectable effect size of 0.15 percentage points from people reporting *Yes*. The power reached with a sample size of 140 observations by treatment is about 0.8 when simulating 1500 Fisher tests. The experiment was conducted online on Prolific (Palan and Schitter, 2018) in February 2021. The experiment was programmed in oTree (Chen et al., 2016). A total of 992 people participated in five sessions.<sup>11</sup> I did not run the whole experiment in one session to avoid overloading the server and minimize the probability of technical issues. Table B.1 in the Appendix presents the number of observations for people on the role of  $P_1$  in each session. The computer program assigned a treatment to each participant. Participants participated only in one treatment, and the game was played only one time. Among the participants, 54.71% identified themselves as male, 44.60% as female, 0.30% as other, and 0.40% preferred not to report it. The average age of participants was 26.25, and 47.28% were students. Participants spent about 7 minutes on average to complete the experiment. In addition to the mind-cheating game earnings and the guessing task, participants earned a completion fee of £1.15.

## 4 Results

Given that I used a mind game in this study, the 'state of the world' is the participants' private information, and I can only compare the reports at an aggregate level. Theoretically, the random draw follows a binomial distribution with a probability of the high-paying state of 0.2. Therefore, I will use this theoretical distribution as a benchmark under truth-telling. Table 3 presents the main outcome variables' mean and standard deviation.  $P_1$ 's *report* is the report by the first mover that can be either 0 or 1. *Yes is more likely* is a binary variable that takes a value of 1 when the reported belief of  $P_2$  reporting 1 is higher than 0.5, which would mean that the participant thought that their partner is more likely to report 1 than 0. *Belief about  $Pr(r_2 = 1)$*  is the subjective probability reported by  $P_1$  that  $P_2$  reports 1. *Times clicked in info* is the number of times a participant clicked the info button in the belief

---

<sup>11</sup>A total of 1009 people showed up, but some left in the middle of the session

elicitation task. Finally, *Time Spent Reporting* is the time in seconds a participant took to decide whether their colors match.

**Table 3. Summary statistics of the main variables in the Mind Game**

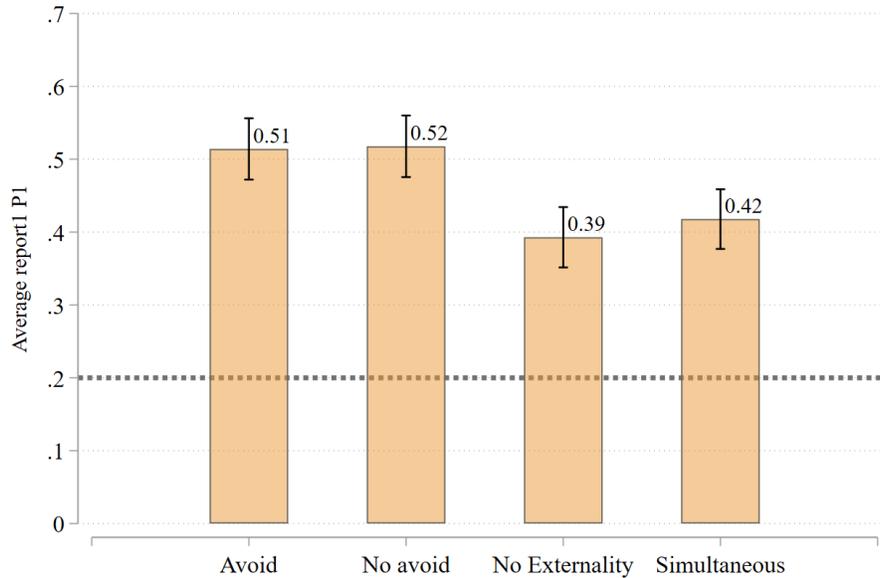
	AVOID	NO AVOID	NO EXTERNALITY	SIMULTANEOUS
$P_1$ 's report=1	0.514 (0.502)	0.518 (0.501)	0.393 (0.490)	0.418 (0.495)
Yes is more likely	0.451 (0.499)	0.326 (0.471)	0.250 (0.435)	0.527 (0.501)
Belief about $Pr(r_2 = 1)$	53.148 (24.353)	46.809 (24.138)	42.843 (23.370)	59.144 (24.835)
Times clicked info	0.338 (0.504)	0.376 (0.515)	0.371 (0.514)	0.329 (0.527)
Time Spent Reporting	15.796 (11.423)	16.823 (9.821)	18.957 (25.469)	9.473 (6.952)
$N$	142	141	140	146

Standard deviations in parenthesis.

A visual representation of the  $P_1$ 's report is presented in Figure 3. This Figure shows the proportion of those participants with the role  $P_1$  that reported Yes. Using the Binomial test, I confirmed that all treatments' actual reports are statistically different from those expected under full honesty. I calculate the expected lying rates of the reports in Figure 3 by taking the average of the reports, then subtracting 0.2 (the expected proportion of people actually matching colors), and finally, I divide the result over 0.8. The resulting expected lying rates are 38.75% in AVOID, 40% in NO AVOID, 23.75% in NO EXTERNALITY, and 27.50% in SIMULTANEOUS. The pairwise comparisons using one-sided Fisher Exact test show that the difference between AVOID and NO AVOID is not significant ( $p = 0.523$ ), the difference between NO AVOID and NO EXTERNALITY is significant ( $p = 0.024$ ), and the difference between AVOID and SIMULTANEOUS is weakly significant ( $p = 0.064$ ).

I use a Linear Probability Model estimation to assess the treatment effects allowing me to control for demographic fixed effects. In columns 1 and 2 of Table 4, I present two regressions with  $r_1$  as the dependent variable. The regressors are the treatment dummies and some demographic variables, which include gender, age, number of experiments in which they have participated, and their student status. The Table also presents, in the row AVOID *mean* the mean of  $r_1$  in the the treatment AVOID. This information makes it easier to interpret the coefficients of the treatment variables. Additionally, the row NO AVOID *vs* NO EXTERNALITY presents the p-value of a Chi-square test assessing the difference between the coefficients of NO AVOID and NO EXTERNALITY. This test was performed post-estimation because the direct comparison between AVOID and NO EXTERNALITY is not clean as ex-

**Figure 3.  $P_1$ 's Yes reports across treatments in Study 2**



Note: The dashed horizontal lines display the underlying theoretical proportion of Yes under truth-telling.

plained before. The regressions *Report  $P_1$  1* and *Report  $P_1$  2* confirm the result derived from Figure 3 that there is no difference in lying between AVOID and NO AVOID leading to Result 1.

**Result 1** (Related to Hypothesis 1).  *$P_1$  lying behavior is not different in AVOID than in NO AVOID. I reject Hypothesis 1.*

A second finding derived from Table 4 is that the impact of SIMULTANEOUS on lying is significant only at the 10% level. In regression *Report  $P_1$  1* the coefficient SIMULTANEOUS shows a null effect when we do not add individual fixed effects. However, once I control for the demographic fixed effects that may account for some heterogeneity, in AVOID  $P_1$  lied more than in SIMULTANEOUS. The coefficient of SIMULTANEOUS implies that participants reported about 10% less frequently having matching colors compared with AVOID. This effect is only weakly significant, but even if the effect does not exist, this finding is in the opposite direction of hypothesis 3 which states that lying will be more pronounced in SIMULTANEOUS than in AVOID.

**Result 2** (Related to Hypothesis 3).  *$P_1$  does not lie less in AVOID than in SIMULTANEOUS. I reject Hypothesis 3.*

The column NO AVOID *vs* NO EXTERNALITY reports the p-value of the test assessing the difference between NO AVOID and NO EXTERNALITY. I use this postestimation test because

**Table 4. Regressions testing the differences across treatments**

	Report $P_1$ 1	Report $P_1$ 2	Report $P_2$ AVOID
NO AVOID	0.004 (0.058)	0.001 (0.059)	
NO EXTERNALITY	-0.121** (0.059)	-0.127** (0.058)	
SIMULTANEOUS	-0.096 (0.059)	-0.102* (0.058)	
$P_1$ 's report=1			-0.242*** (0.083)
AVOID mean	0.514	0.511	
NO AVOID vs NO EXTERNALITY	0.032	0.028	
Demographic FE	No	Yes	Yes
$R^2$	0.013	0.031	0.091
Observations	569	567	142

Note: Regressions Report  $P_1$  1, Report  $P_1$  2, Report  $P_2$  are Linear Probability models. Report  $P_1$  1 and Report  $P_1$  2 use data from all the treatments. Report  $P_2$  AVOID uses data from  $P_2$  in AVOID. Bootstrap standard errors are presented in parentheses.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

NO EXTERNALITY is only directly comparable with NO AVOID. These p-values reveal that  $P_1$  lying is lower in NO EXTERNALITY than NO AVOID which is consistent with hypothesis 2. In particular, it is about 12% less likely that a participant reports that their colors match when reporting 1 has no benefits in a third party.

**Result 3** (Related to Hypothesis 2).  $P_1$  lies less in NO EXTERNALITY than in NO AVOID. I do not reject Hypothesis 2.

Besides the results concerning  $P_1$ 's reports, Table 4 also presents important evidence regarding  $P_2$  reports. In the column *Report  $P_2$* , it presents the relation between  $r_1$  and  $r_2$ . In this regression, I only used data from AVOID because it is the only treatment where  $P_2$  can lie after learning  $r_1$ . The coefficient  $P_1$ 's report=1 shows that  $P_2$  was significantly more likely to report *Yes* when  $r_1 = \text{No}$ . The mean for  $P_1$ 's report=0 is 0.373 which puts in perspective the coefficient for  $P_1$ 's report=1. Using [Hugh-Jones \(2019\)](#) Bayesian method, I estimate the lying rates of  $P_2$  conditional on  $r_1$ . The lying rate when  $P_2$  observed  $r_1 = \text{Yes}$  is 9.39% while the lying rate when observing  $r_1 = \text{No}$  is 36.63%.

**Result 4.**  $P_2$  lie more when they observe that  $P_1$  reported No than when they observe that  $P_1$  reported Yes.

Finally, one remaining question is whether groups coordinated into having only one liar who bears the cost of lying. In Table 5, I present the number of participants per group who reported *Yes* which is simply the sum of  $r_1$  and  $r_2$ . Given that lies are substitutes, one should expect in a few groups, both participants report *Yes*, except in SIMULTANEOUS where coordination was more difficult than in other treatments. Table 5 shows that given the incentives in AVOID, the proportion of groups with at least one liar was higher in this treatment than in any other. Even though the results also show that lying aversion is still important for some individuals, given that in 24.65% of the groups, no one lied. Additionally, one can see that despite SIMULTANEOUS creating the highest number of groups with both participants reporting *Yes*, the proportion of groups with no liar was similar as in NO AVOID.

**Table 5. Number of participants per group who reports *Yes***

$r_1 + r_2$	AVOID		NO AVOID		NO EXTERNALITY		SIMULTANEOUS	
	Freq.	Percent	Freq.	Percent	Freq.	Percent	Freq.	Percent
0	35	24.65	51	36.17	68	48.57	52	35.62
1	88	61.97	75	53.19	65	46.43	66	45.21
2	19	13.38	15	10.64	7	5	28	19.18
Total	142	100	141	100	140	100	146	100

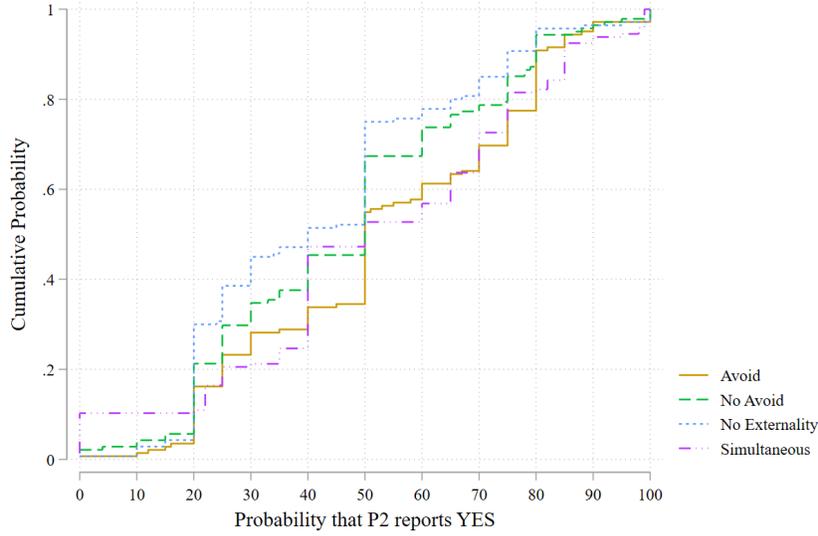
#### 4.1 Secondary outcomes

Another outcome of the experiment was the elicited beliefs about  $P_2$ 's report. Figure 4 presents the cumulative density function of the implied probabilities of  $P_2$  reporting *Yes* in each treatment. I use a Kolmogorov-Smirnov test to test whether these distributions are equal. The pairwise comparisons using this test shows that in the pairs AVOID-SIMULTANEOUS and NO AVOID-NO EXTERNALITY, there is no statistically significant difference, but in the pair AVOID-NO AVOID, there is a weakly significant difference. This test implies that there are lower belief levels in NO AVOID than in ( $p = 0.084$ ).

To deepen the insights from Figure 4, I use OLS regressions to study the differences in beliefs across treatments. Table B.3 in the appendix presents two regressions (one for each reference treatment). In both regressions, the belief about  $P_2$  reporting *Yes* is the dependent variable. I also include interaction terms of the report by  $P_1$  and each treatment. Regression *Beliefs 1* in Table B.3 shows that those participants in SIMULTANEOUS who reported *No* believed that their partner was more likely to report *Yes* than *No*. Table B.3 also shows that beliefs were not self-serving in the mind-cheating game.<sup>12</sup>

<sup>12</sup>Using the Kolmogorov-Smirnov test to assess that there is no differences in the distributions gives the

**Figure 4. Cumulative Density Function of  $P_1$ 's subjective probability that  $P_2$  reports Yes**



**Result 5** (Beliefs in mind game).  $P_1$ 's subjective probability of  $r_2 = 1$  is only positively correlated with  $r_1$  in the mind game. Participants in SIMULTANEOUS believe that it is more likely that their partner reports Yes than in other treatments.

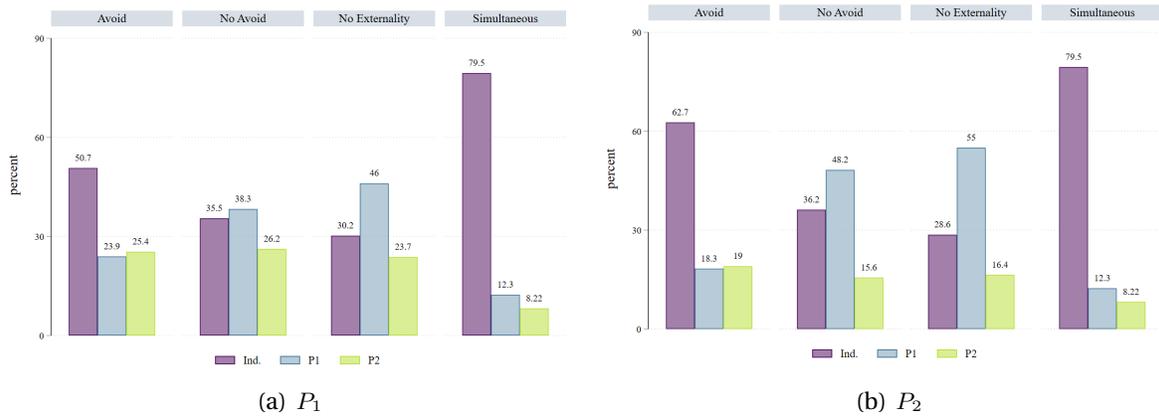
Finally, in the experimental sessions, I included a question in the final questionnaire where I asked them: "Imagine you were to play the same game again and had a choice, Would you rather be Participant A or Participant B?"<sup>13</sup> Their responses, distinguished by the role they had, are presented in Figure 5. Figure 6(a) shows that  $P_1$  did not, in general, interpret being the first mover as an advantage in AVOID or SIMULTANEOUS. Figure 6(b) shows that this is also true for  $P_2$  with even more people being completely indifferent between the two roles. Figure 6(b) also shows that  $P_2$  did not like the second mover position when they could not report and preferred being  $P_1$ .

## 5 Discussion

This paper aimed to analyze the impact of two different behavioral traits, lying aversion and prosociality, on the decision of whether to lie or not. In a first experiment, the observed-cheating game, I was not able to detect any treatment differences because a lot of participants were reluctant to lie. Previous evidence of laboratory experiments on cheating showed that, in observed cheating games, people lie less (Gneezy et al., 2018; Abeler et al., same results.

<sup>13</sup>This was the exact wording I used in both experiments to refer to  $P_1$  and  $P_2$ .

**Figure 5. Which role would participants choose if they play again?**



2019; Fries et al., 2021). However, they were able to detect treatment effects. I used this previous evidence as the basis for designing this experiment. My primary motivation for using an observed random draw was to get information at the individual level. In contrast to laboratory experiments, my online experiment only found extremely low lying rates: cheating was only observed in between 4% and 13% of the first-mover choices. Hence, the low lying rates meant that there was limited scope to detect treatment effects.

To make it easier to detect treatment differences, I used the mind-cheating game, where it was not possible to know whether a specific participant lied or not. The main advantage of the observed cheating game was that I could identify liars and get more information about liars and non liars. In contrast, participants were more willing to lie in the mind-cheating game, but I could only detect lying at the aggregate level. In the mind-cheating experiment, where participants' private information was not observed, it was possible to assess the hypotheses presented in 2 because participants were more sensitive to changes in the decision context. The main lesson we can draw when comparing the behavior from the observed-cheating game and the mind-cheating game is that it is crucial to reduce observability when using platforms such as Prolific (Palan and Schitter, 2018) to study lying behavior. A potential reason is that they stress the importance of being honest when participating in studies. Then, even if the only identification is their Prolific ID, participants may care about how they are perceived.

Regarding the main objective of this paper, I hypothesized that most of the participants assigned the role of  $P_1$  in the treatment AVOID would try to reduce their lying costs and pass the burden to  $P_2$ . Result 1 showed that this was not the case and that participants had similar lying rates in AVOID and NO AVOID. This finding was unexpected and suggested that either many participants in the role of  $P_1$  in AVOID expected that  $P_2$  would not lie and then decided to secure the highest payoff or that their utility derived from the prosocial lie

was high enough to make lying attractive. The beliefs of  $P_1$  in AVOID, presented in Figure 4 and Table B.3, ruled out the possibility that, in AVOID, most of  $P_1$  believed that  $P_2$  was particularly honest. Additionally, Result 4 showed that participants assigned the  $P_2$  role were willing to avoid lying when their actions did not affect others' payoffs or their payoffs. Result 3 showed indeed that in NO EXTERNALITY, where there is no prosocial motive,  $P_1$  lied considerably less. Thus, it also supports the notion that prosociality is the main channel explaining the similar lying rates of  $P_1$  in AVOID and NO AVOID.

Hence, the most important insight of the paper is that even for people who are lying averse, prosociality decreases their lying cost if their lies benefit others as well as themselves. Consequently, they are more likely to lie in this case. This result adds to the previous finding by [Wiltermuth \(2011\)](#), [Gino et al. \(2013\)](#), and [Levine and Schweitzer \(2015\)](#) by showing that individuals are more willing to lie when they create a positive externality and that this motive is strong enough to make some people with high lying costs more prone to lying. The question that consequently arises is how SIMULTANEOUS enters this picture. Arguably, SIMULTANEOUS provides two contributions to the paper's main result. First, it is in line with the assumption regarding how prosociality was included in the utility function. Specifically, I used a particular assumption about the type of prosocial preferences where participants benefit from the consequences of their actions and not from the actions themselves. These types of prosocial preferences are inspired by consequentialism and opposed to deontological ethics that evaluate the means instead of the ends. Even though SIMULTANEOUS was not intended to test this assumption, Result 2 suggests that participants hold consequentialist prosocial preferences when lying for others.

Second, and more importantly, SIMULTANEOUS confirms that the prosocial motive is strong enough to make some otherwise honest participants lie. As presented in Figure 1 of Section 2, when the utility generated by the positive externality ( $\theta$ ) is high enough, lying is higher in AVOID than in SIMULTANEOUS. This finding was not expected because I predicted that people would be more self-interested and more motivated to avoid the psychological cost of lying than trying to tell a prosocial lie. However, once we established that the prosocial motive was strong enough that some people would neglect their lying aversion motive, this result is consistent with the model. In other words, it would have been contradictory to have similar rates of lying in AVOID and NO AVOID but to have more lying in SIMULTANEOUS than in AVOID. Obtaining these hypothetical results would have run contrary to the theoretical model, but my results are consistent with it. The elicited beliefs about  $r_2$  in SIMULTANEOUS are also consistent with the model. More specifically, participants were more likely to believe that their partner would report *Yes* than *No*. Thinking that the other participant is likely to lie implies that it will be easier to avoid the lying cost and moreover that it is more likely to lose the utility of the positive externality if both lie.

Finally, I would like to compare the results of this paper theoretically with the collaborative lying games. In particular, I want to use my theoretical model to compare my results in AVOID with a game with a similar structure like the classical dyadic game of [Weisel and Shalvi \(2015\)](#). Imagine the same game as in AVOID; however, instead of being paid 1 if at least one group member reports 1, participants only get the monetary reward if both members of the group report 1. In this collaborative lying game, the utility function would be represented by:

$$U_i(x_i, r_i, r_j) = r_i r_j - 1_{x_i \neq r_i} (c_i - \theta r_i r_j) \quad (8)$$

This utility function includes consequentialist prosocial lying by assuming that people only get  $\theta$  when both participants lie because only then their action affects others' payoffs. I use the same logic as in section 2 to get the lying threshold for  $P_1$  in this game. In particular the theoretical cost of lying threshold would be  $\hat{c}_1 = (1 + \theta)(0.2 + 0.8 \frac{1 + \theta}{c})$ . Figure 6 compares this threshold with the one in equation (4). This figure shows that  $P_1$  would be more likely to lie in the complementaries game than in the substitutes game<sup>14</sup> the higher the prosocial motive is. In other words, if people are highly prosocial, then the first mover will lie and they will also expect that the second mover will lie not because of a dishonest population but because of a cooperative population. This theoretical implication of my model is consistent with the evidence from collaborative lying experiments where lying rates are high. Hence, the results of my paper, which finds that prosociality is a strong motivator of lying behavior in groups, combined with the theoretical implications of this empirical result, also shed some light on the channel that makes people lie more in collaborative lying games.

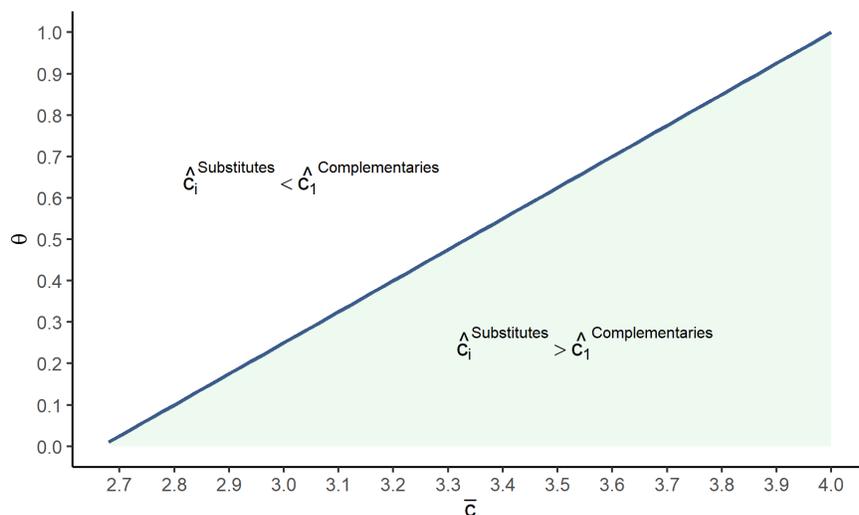
## 6 Conclusion

Does prosocial lying make people more likely to lie? In this paper, I found evidence that it does, even for some people with high lying costs. Studies by [Gneezy et al. \(2018\)](#), [Abeler et al. \(2019\)](#), and [Khalmetski and Sliwka \(2019\)](#) have indicated that one explanation for people not lying to maximize their monetary payoff is that they have psychological lying cost. However, it is unclear whether this effect holds in group settings where prosocial lying enters the picture. For instance, it is not the same lying to get an individual bonus a CEO has promised when achieving a goal than lying to reach a threshold that gives the bonus to a team. A similar dilemma arises when people use intermediaries in some situations, for instance, to complete tax declarations, sell a car, or sell a house. In these contexts, the

---

<sup>14</sup>AVOID in this paper.

**Figure 6. Comparison of Lying Thresholds in AVOID and Sequential Collaborative Lying Game.**



intermediary has a higher payoff if they lie. Therefore, when individuals lie to benefit both others and themselves, there are two competing motivators: lying aversion and prosociality.

I used experimental data to study these situations. I found meager lying rates in a first experiment, making it difficult to identify any treatment differences. Arguably, the main reason for the low lying rates I found was that I used an *Observed Game*, which makes it possible to know whether a participant lied. In a second experiment, I solved the observability problem by using a mind game. The mind-cheating game showed that individuals lie more when they can benefit others. In addition, it showed that this motivation is strong enough to prevail even when people can reduce their direct lying costs.

One additional finding was that prosocial lying is consistent with consequentialist rather than deontological views. In particular, it is only possible for prosociality to be a stronger motivator than lying aversion on average in the theoretical model if people care about the consequences of their acts rather than their intentions. However, the scope of this study was limited in terms of assessing whether consequentialist prosocial behavior is the only way to explain the results, and the experiment was not designed to assess it directly. Another issue not addressed in this study was whether the timing of the belief elicitation changed participants' guesses and their willingness to avoid their lying costs. Beliefs were elicited after  $P_1$  had reported their random draw. Therefore, beliefs might have been influenced by participants' reports. It was beyond the scope of the paper to test whether participants would be more prone to avoid their lying costs in AVOID when the elicitation task was done before reporting.

In spite of the mentioned limitations, the study certainly adds to our understanding of the role of prosociality on dishonesty. Although prosocial lying may seem trivial, it is, in fact, crucial in terms of today's concern over tax evasion and corruption. In practical terms, it suggests that it should be avoided having groups of people or intermediaries in positions where self-reports are central because people will be more likely to lie. For instance, continued efforts are needed to make declaring taxes easier for the general population so that they do not need to use an intermediary to declare for them. The same principle applies to situations in which one person is in charge of reporting the information on behalf of a team (for instance: a political party, a workgroup, or a firm). Individual reporting should always be preferred over creating dependencies between people's reports.

The findings provide important insights into the broader domain of dishonesty and prosociality. Nonetheless, some questions still remain to be answered. It is important to assess directly whether prosocial lying is a matter of intentions or consequences. I found insights into consequentialism in this paper, but further research is needed to confirm it. Another natural progression of this work is to analyze whether reciprocal lying exists in group settings. Finally, further research might explore the role of belief elicitation timing in the strategic avoidance of lying.

## References

- Abeler, J., Nosenzo, D., and Raymond, C. (2019). Preferences for Truth-Telling. *Econometrica*, 87(4):1115–1153.
- Akerlof, G. A. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3):488–500.
- Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal*, 100(401):464–477.
- Andreoni, J. and Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.
- Ariely, D., Bracha, A., and Meier, S. (2009). Doing good or doing well? image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99(1):544–55.
- Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *The American Economic Review*, 96(5):1652–1678.
- Charness, G. and Rabin, M. (2002). Understanding social preferences with simple tests. *The quarterly journal of economics*, 117(3):817–869.
- Chen, D. L., Schonger, M., and Wickens, C. (2016). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

- Coffman, L. C. (2011). *Intermediation reduces punishment (and reward)*, volume 3.
- Conrads, J., Irlenbusch, B., Rilke, R. M., and Walkowitz, G. (2013). Lying and team incentives. *Journal of Economic Psychology*, 34:1–7.
- DellaVigna, S., List, J. A., and Malmendier, U. (2012). Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, 127(1):1–56.
- Dimant, E., Van Kleef, G. A., and Shalvi, S. (2020). Requiem for a nudge: Framing effects in nudging honesty. *Journal of Economic Behavior & Organization*, 172:247–266.
- Dufwenberg, M. and Dufwenberg, M. A. (2018). Lies in disguise – A theoretical analysis of cheating. *Journal of Economic Theory*, 175:248–264.
- Erat, S. and Gneezy, U. (2011). White lies. *Management Science*, 58(4):723–733.
- Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, 11(3):525–547.
- Fries, T., Gneezy, U., Kajackaite, A., and Parra, D. (2021). Observability and lying. *Journal of Economic Behavior and Organization*, 189:132–149.
- Garbarino, E., Slonim, R., and Villeval, M. C. (2019). Loss aversion and lying behavior. *Journal of Economic Behavior & Organization*, 158:379–393.
- Gino, F., Ayal, S., and Ariely, D. (2013). Self-serving altruism? The lure of unethical actions that benefit others. *Journal of Economic Behavior and Organization*, 93:285–292.
- Gneezy, U., Kajackaite, A., and Sobel, J. (2018). Lying Aversion and the Size of the Lie. *American Economic Review*, 108(2):419–453.
- Hugh-Jones, D. (2019). True lies: Comment on garbarino, slonim and villeval (2018). *Journal of the Economic Science Association*.
- Jiang, T. (2013). Cheating in mind games: The subtlety of rules matters. *Journal of Economic Behavior and Organization*, 93:328–336.
- Kajackaite, A. and Gneezy, U. (2017). Incentives and cheating. *Games and Economic Behavior*, 102:433–444.
- Karni, E. (2009). A Mechanism for Eliciting Probabilities. *Econometrica*, 77(2):603–606.
- Khalmetski, K. and Sliwka, D. (2019). Disguising Lies—Image Concerns and Partial Lying in Cheating Games. *American Economic Journal: Microeconomics*, 11(4):79–110.
- Kocher, M. G., Schudy, S., and Spantig, L. (2018). I lie? we lie! why? experimental evidence on a dishonesty shift in groups. *Management Science*, 64(9):3995–4008.
- Leib, M., Köbis, N., Soraperra, I., Weisel, O., and Shalvi, S. (2021). Collaborative dishonesty: A meta-analytic review. *Psychological Bulletin*, 147(12):1241.
- Levine, E. E. and Schweitzer, M. E. (2015). Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106.

- Mobius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2011). Managing self-confidence: Theory and experimental evidence. Technical report, National Bureau of Economic Research.
- Muehlheusser, G., Roider, A., and Wallmeier, N. (2015). Gender differences in honesty: Groups versus individuals. *Economics Letters*, 128:25–29.
- Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.
- Potters, J. and Stoop, J. (2016). Do cheaters in the lab also cheat in the field? *European Economic Review*, 87:26–33.
- Rilke, R. M., Danilov, A., Weisel, O., Shalvi, S., and Irlenbusch, B. (2021). When leading by example leads to less corrupt collaboration. *Journal of Economic Behavior Organization*, 188:288–306.
- Shalvi, S. and De Dreu, C. K. (2014). Oxytocin promotes group-serving dishonesty. *Proceedings of the National Academy of Sciences*, 111(15):5503–5507.
- Shalvi, S., Eldar, O., and Bereby-Meyer, Y. (2012). Honesty Requires Time (and Lack of Justifications). *Psychological Science*, 23(10):1264–1270.
- Weisel, O. and Shalvi, S. (2015). The collaborative roots of corruption. *Proceedings of the National Academy of Sciences*, 112(34):10651–10656.
- Wiltermuth, S. S. (2011). Cheating more when the spoils are split. *Organizational Behavior and Human Decision Processes*, 115(2):157–168.

## Appendix A Two person cheating game – Observed

In this experiment, for the random draw, participants click on one card out of ten that reveals a color. There are two possible colors: *Orange* and *Black*. Reporting *Orange* pays £4, reporting *Black* pays £0.5. In total there are 8 cards with *Black* behind and 2 with *Orange*. In this study, given that the random draw  $x_i$  is observed by the experimenter, it is possible to identify whether  $P_1$  lied or not.

### A.1 Procedures

I pre-registered the experiment in AEA RCT Registry under the number AEARCTR-0006881. I targeted 120 observations per treatment ex-ante. I calculated the power of the target sample size using computer simulations. I used a minimum detectable effect size of 0.15 percentage points from people detected as liars. The power reached with a sample size of 120 observations by treatment is 0.8 when simulating 1500 Fisher tests. The experiment was conducted online on Prolific (Palan and Schitter, 2018) in December 2020. The experiment was programmed in oTree (Chen et al., 2016). A total of 878 people participated in five sessions.<sup>15</sup> I did not run the whole experiment in one session to avoid overloading the server and minimize the probability of technical issues. Table A.1 presents the number of valid observations for people on the role of  $P_1$  in each session. The computer program assigned a treatment to each participant.

Table A.1. Participants with the role of  $P_1$  in Study 1

	Session 1	Session 2	Session 3	Session 4	Session 5	Total
<b>AVOID</b>	22	23	22	23	35	125
<b>NO AVOID</b>	22	21	21	22	35	121
<b>NO EXTERNALITY</b>	23	24	22	21	37	127
<b>SIMULTANEOUS</b>	20	22	22	20	48	132

Participants read the instructions first and responded to some comprehension questions. After they answered the comprehension questions correctly, they waited until a second player was also ready, then they were matched together and proceeded to the observed cheating game. Roles were assigned randomly. After participants finished the observed cheating game and the elicitation task, they responded to a survey with demographic questions and a feedback question. Participants spent about 6 minutes on average to complete the experiment. Additional to the earnings on the cheating game and the guessing task, participants earned a completion fee of £2.5. Following Prolific rules, participants who left the experiment did not get the completion fee. Participants received their payoffs through the Prolific platform the same day they participated in their session.

<sup>15</sup>A total of 899 people showed up, but some left in the middle of the session.

## A.2 Results

The main outcome variable of interest to test the hypotheses presented in Section 2 is whether  $P_1$  lied or not. Given that I have the information about the random draw and the report in this experiment, I create a new variable called *Lied* that takes the value of 1 when  $x_i \neq r_i$ , and 0 when  $x_i = r_i$ . Table ?? presents the lying rates per treatment and some demographic variables for each treatment. Regarding the variable *Lied*, the results show that the lying rates are on average 7.52%, which is very low compared, for instance, with Gneezy et al. (2018) where lying rates were about 30%. The pairwise comparison of the comparable treatments using a Fisher exact test results in no significant differences at a 0.05 level across treatments.<sup>16</sup>

**Table A.2. Summary statistics Observed Game**

	Avoid	No Avoid	No Externality	Simultaneous
Lied=1	0.080 (0.272)	0.056 (0.230)	0.126 (0.333)	0.038 (0.192)
Student Status	0.472 (0.501)	0.437 (0.498)	0.512 (0.502)	0.455 (0.500)
Age	26.512 (8.395)	26.024 (8.235)	25.063 (7.433)	26.811 (9.414)
Gender	0.504 (0.548)	0.492 (0.562)	0.583 (0.635)	0.500 (0.586)
$N$	125	126	127	132

Standard deviations in parenthesis.

To confirm the result implied by Table ??, I use regressions that allow me to include some controls. Table A.3 presents, in columns 1 and 2, Linear Probability regressions with *Lied* as dependent variable. Lied 1 uses data from treatments AVOID, NO AVOID, and SIMULTANEOUS, with AVOID as the reference treatment. Lied 2 uses data from treatments AVOID, NO AVOID, and NO EXTERNALITY, with NO AVOID as the reference treatment. I use two regressions because AVOID is not directly comparable with NO EXTERNALITY which make it impossible to include NO EXTERNALITY in Lied 1. The same intuition applies to Lied 2 because between NO AVOID and SIMULTANEOUS two things change. In both regressions, I use only the data of  $P_1$  that drew *Black*. The independent variables are the treatment dummies, the time participants took to send the report, and some demographic variables.

The coefficients for the treatments dummies in Lied 1 and Lied 2 confirm no significant differences in lying rates across treatments. Additionally, I find that participants who spent more time reporting were likelier to lie. Interestingly, the coefficient of *Time Spent Reporting* shows that the probability reported by  $P_1$  is positively correlated with the probability that  $r_1 = 1$ . This result is not a surprise given the low lying rates. In this case, it is challenging to identify treatment differences because the power will be too low. In other words, I find no treatment differences between treatments leading me to reject Hypotheses 1, 2, and 3.

<sup>16</sup>To ensure that only one component changes, I only compare treatments in the following pairs: AVOID-NO AVOID, AVOID-SIMULTANEOUS, and NO AVOID-NO EXTERNALITY.

**Table A.3. Regressions testing the differences across treatments in the Observed Game**

	Lied 1	Lied 2	Beliefs 1	Beliefs 2
AVOID	<i>Reference</i>	0.013 (0.041)	<i>Reference</i>	7.875*** (2.616)
NO AVOID	-0.018 (0.037)	<i>Reference</i>	-7.607*** (2.759)	<i>Reference</i>
No Externality		0.061 (0.044)		3.183 (2.651)
SIMULTANEOUS	-0.030 (0.033)		-1.191 (3.159)	
Time Spent Reporting	0.009** (0.004)	0.012*** (0.004)		
Lied=1			24.105*** (5.414)	9.199** (4.522)
Constant	-0.017 (0.061)	-0.035 (0.082)	33.189*** (3.922)	26.030*** (4.033)
Controls	Yes	Yes	Yes	Yes
Observations	302	292	302	292
$R^2$	0.050	0.077	0.132	0.081

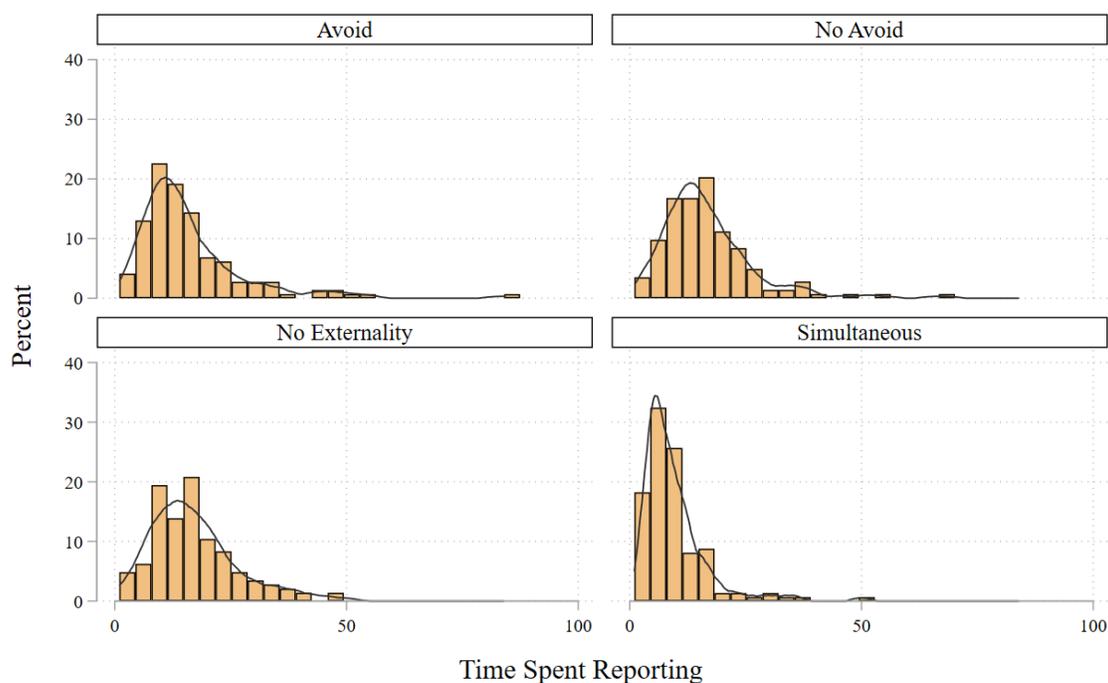
Note: Regressions Lied 1 and Lied 2 are Linear Probability models. Regressions Beliefs 1 and Beliefs 2 use OLS. All regressions use the data of participants with the role of  $P_1$  and who drew *Black*. Lied 1 and Belief 1 use data from treatments AVOID, NO AVOID, and SIMULTANEOUS. Lied 2 and Beliefs 2 use data from treatments AVOID, NO AVOID, and NO EXTERNALITY. Controls include gender, age, student status, education, number of experiments they participated in before, and their id in a session. Bootstrap standard errors are presented in parentheses.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

The mean beliefs by treatment are not significantly different in all pairwise comparisons using a Kolmogorov-Smirnov test. In columns Beliefs 1 and Beliefs 2 of Table A.3 I use a Ordinary Least Squares to assess whether the reported probability of reporting *Orange* varies across treatments. Belief 1 uses data from treatments AVOID, NO AVOID, and SIMULTANEOUS with AVOID as the reference treatment. Beliefs 2 uses data from treatments AVOID, NO AVOID, and NO EXTERNALITY with NO AVOID as the reference treatment. Beliefs 1 show that the reported probability in NO AVOID is lower than in AVOID while the difference between AVOID and SIMULTANEOUS is non-significant. Column Beliefs 2 shows that the difference between NO AVOID and NO EXTERNALITY is non-significant neither.

## Appendix B Further Results

**Figure B.1. Time  $P_1$  spent reporting whether their colors match.**



**Table B.1. Participants with the role of  $P_1$  in the Mind Game**

	Session 1	Session 2	Session 3	Session 4	Total
<b>Avoid</b>	29	38	38	37	142
<b>No Avoid</b>	29	36	39	37	141
<b>No Externality</b>	28	37	39	36	140
<b>Simultaneous</b>	30	38	40	38	146

**Table B.2. Treatment comparisons using Fisher Exact Tests**

	No Avoid	No Externality	Simultaneous
<b>Avoid</b>	$p = 0.523$	-	$p = 0.064$
<b>No Avoid</b>	-	$p = 0.024$	-

Note: I use 1-sided Fisher's exact tests given that I had directional hypotheses.

**Table B.3. Regressions testing the differences in Beliefs across treatments in Study 2**

	Beliefs 1	Beliefs 2
Avoid	<i>Reference</i>	3.711 (4.042)
No Avoid	-3.719 (4.025)	<i>Reference</i>
Simultaneous	7.895** (3.691)	
No Externality		-3.379 (3.721)
$P_1$ 's report=1	6.397 (4.090)	2.297 (4.102)
Avoid $\times P_1$ 's report=1		4.129 (5.791)
No Avoid $\times P_1$ 's report=1	-4.901 (5.785)	
Simultaneous $\times P_1$ 's report=1	-2.963 (6.192)	
No Externality $\times P_1$ 's report=1		-0.439 (5.834)
Constant	49.835*** (5.220)	55.810*** (5.499)
Controls	Yes	Yes
Observations	428	421
$R^2$	0.055	0.051

Note: Regressions Report  $P_1$  1, Report  $P_1$  2, Report  $P_2$  are Linear Probability models. Controls include gender, age, student status, education, religion, number of experiments they participated before, and their id in a session. Robust standard errors are presented in parentheses.

\*  $p < 0.1$ , \*\*  $p < 0.05$ , \*\*\*  $p < 0.01$ .

## Appendix C Instructions of the Mind Game

I present here the instructions used in the Mind Game. I provide the complete instructions for the treatment AVOID and the variations for the other treatments. The full instructions for each treatment, as well as the instructions for the Observed Game, are hosted in this [repository](#). You can also download the code to run the experiment using oTree.

### C.1 AVOID

[Screen 1]

During this study, you will interact in real-time with an anonymous partner. The game will last about 10 minutes (max. 15 minutes).

It is then crucial that you stay in front of the screen for the next 10-15 minutes. There will be some moments where you have to wait until your partner decides, so please be patient as your partner may take some minutes to decide.

Can you be in front of the screen for the next 15 minutes?

Yes\_\_\_ No\_\_\_

[Screen 2]

#### **Instructions**

Welcome to our study!

Please read the instructions carefully.

You will receive £1.15 after you complete the experiment and fill in a final questionnaire.

During the experiment, you will be able to earn additional money depending on the decisions you make. The decisions you make during the game will not be shared with Prolific at any time. We will explain how the game works on the next screen.

If something goes wrong, please make a screenshot and contact us via Prolific.

At the beginning of the experiment, you will be randomly matched with another participant.

The computer will then randomly assign a role to each of you. One participant will be Participant A and the other one will be Participant B. These labels will ensure that neither of you know the identity of the other participant.

[Screen 3]

The experiment works as follows:

- (i) Participant A sees 5 cards with different colors (black, orange, blue, yellow, and green) and chooses one color in their head.
- (ii) On the next screen, Participant A sees 10 cards with a question mark. Behind each card, there is a color. There are 2 black cards, 2 orange cards, 2 blue cards, 2 yellow

cards, and 2 green cards. The cards are placed in a random order. Participant A clicks on a card and the card flips and shows the color behind it. Only Participant A knows the color of the card she/he picked.

- (iii) After Participant A sees the color behind the picked card, she/he reports whether the color of the card she/he picked in the second stage is the same as the color she/he mentally chose in the first stage. Participant A has to report whether the colors of the picked card and the later seen card match. If both colors match, she/he reports "Yes"; otherwise, she/he reports "No". A message will later be sent to Participant B stating whether Participant A reported "Yes" or "No".
- (iv) Then, Participant B sees 5 cards with different colors (black, orange, blue, yellow, and green) and chooses one color in their head.
- (v) On the next screen, Participant B sees 10 cards with a question mark. Behind each card, there is a color. There are 2 black cards, 2 orange cards, 2 blue cards, 2 yellow cards, and 2 green cards. The cards are placed in a random order. Participant B clicks on a card and the card flips and shows the color behind it. Only Participant B knows the color of the card she/he picked.
- (vi) After picking a card, Participant B receives the message with the color reported by Participant A.
- (vii) Finally, Participant B reports whether the color of the card she/he picked in the second stage is the same as the color she/he mentally chose in the first stage. If the colors of the picked card and the later seen card match, she/he reports "Yes"; otherwise, she/he reports "No".

The reports by Participant A and B determine the payments in the experiment. Participants report "Yes" or "No" to the following statement: "Please indicate whether the color behind the card you picked is the color that you thought of." If either Participant A or B reports "Yes" (no matter who), both participants receive £2.50. If both report "No", both participants receive £0.30. All the possible report combinations are summarized in the table below.

Participant's report		Earnings	
A	B	A	B
No	No	£0.30	£0.30
No	Yes	£2.50	£2.50
Yes	No	£2.50	£2.50
Yes	Yes	£2.50	£2.50

There will be no further rounds in this experiment. That is, you will participate in the task described above only once.

Before starting, we'd like you to answer the questions on the next page to check your understanding.

[Screen 4]

1. Imagine the following scenario: Participant A reports No and Participant B reports Yes. What would the payments be?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

2. Now, imagine the following scenario: Participant A reports Yes and Participant B reports No. What would the payments be?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

3. What would the payments be if both participants report No?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

4. What would the payments be if both participants report Yes?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

5. When does Participant B learn what Participant A has reported?

- Before Participant B reports whether his or her card colors match.
- After Participant B reports whether his or her card colors match.

Once you have answered all the questions correctly, you can continue.

Please note that the experiment does not contain any further comprehension questions or attention checks from this point on!

[Screen 5]

**You are Participant A.**

Please choose in your head one from the five cards below.

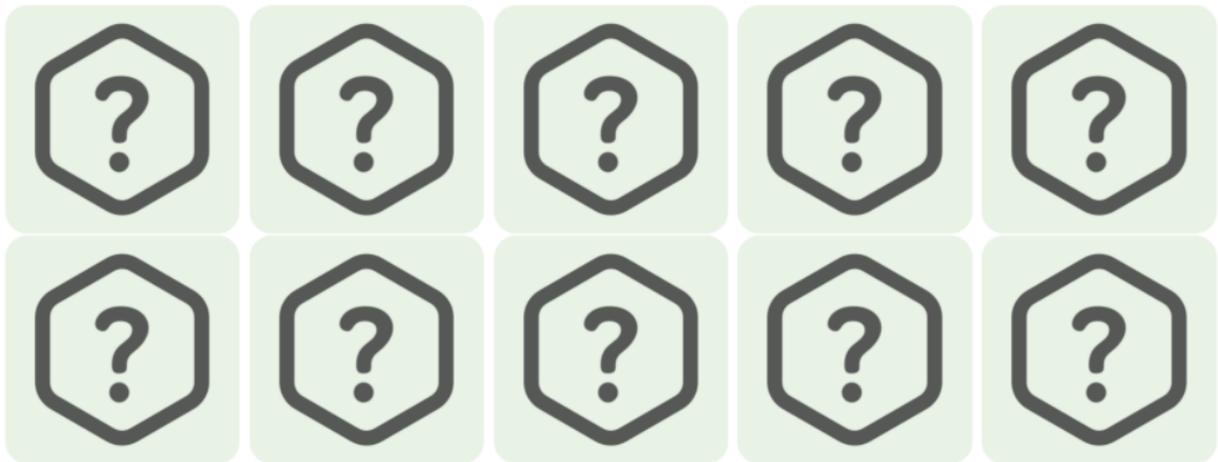


It is important that you remember your color for the rest of the game.

Once you have mentally chosen your card, you can click on "Next" to continue.

[Screen 6]

Please pick one card by clicking on it.



Once you have picked a card and seen the color behind it, you can click on the button "Show all" if you want to see what color was behind each card. You don't have to do so; it is only a tool to show you how the cards were distributed.

[Screen 7]

We now ask you to report whether the color you chose in the first stage was the same as the color you drew in the second stage.

Participant B will receive a message with your report.

Participant B will receive the message before she/he reports whether her/his colors match.

Please indicate whether the color behind the card you picked is the color that you thought of:

- Yes
- No

[Screen 8]

### **Guessing Participant B's report.**

Participant B in your group has already picked a card and is now reporting. In the meantime, we want to know what you think she or he will report and how confident you are of your guess.

You can earn £0.30 by guessing correctly. A robot may help you to increase your chances of earning this additional money. The robot will only help you from the point where you are not sure. In particular, you only need to select whether you think the other participant will report "Yes" or "No", and how likely you think your guess is to be correct (i.e. if you believe there's a 75% chance your guess is correct, you should write down 75).

The robot's selection is based in an algorithm, so you only have to tell us your guess and the chance it is correct. You don't need to know how exactly the robot's algorithm works to continue with the experiment. However, if you want to find out how it works, click on "more information". If not click directly on "Next".

#### More information pop up:

How do the robots work?

We have 100 different robots; each has a different level of accuracy. Each robot has an accuracy corresponding to an integer between 1 and 100. That is, there is a robot that is accurate 1% of the time, a robot that is accurate 2% of the time, a robot that is accurate 3% of the time, ... , all the way up to a robot that is accurate 100% of the time. A robot that is accurate 75% of the time correctly guesses the other participant's report 75% of the time and guess wrongly 25% of the time.

By reporting how confident you are with your guess, you decide which robots you would allow to guess for you.

Here's how it will work.

First, you will select whether you think Participant B will report "Yes" or "No". Then, you will decide how confident you are in this guess. You will do this by choosing an accuracy threshold (a number between 1 and 100) for your answer. For any robot that has accuracy greater than or equal to your threshold, you would prefer to have the robot answering instead of submitting your guess. For any robot that has an accuracy lower than your threshold, you would prefer to submit your guess instead of letting the robot answer.

Then, the computer will randomly select a robot. Each robot is equally likely to be chosen. If the robot has an accuracy greater than or equal to your threshold, the robot will guess the other participant's report for you. If the robot has an accuracy less than your threshold, your guess will be submitted and you will receive £0.30 additional based upon that guess.

For example, if you chose 75% as your accuracy threshold, and the randomly selected robot had an accuracy of 90%, this robot would answer for you. The robot would have a 90% chance of guessing Participant B's report correctly. If you chose 75% as your accuracy threshold, and the robot randomly selected had an accuracy of 20%, your answer would be submitted instead of the robot's.

[Screen 9]

### **Guessing Participant B's choice**

When Participant B was asked whether the color she/he picked is the color she/he thought of, I think Participant B reported:

- Yes
- No

I think the chance that my answer is correct is (write a number between 0 and 100):

[Screen 5 Participant B]

### **You are Participant B.**

Please choose in your head one from the five cards below.

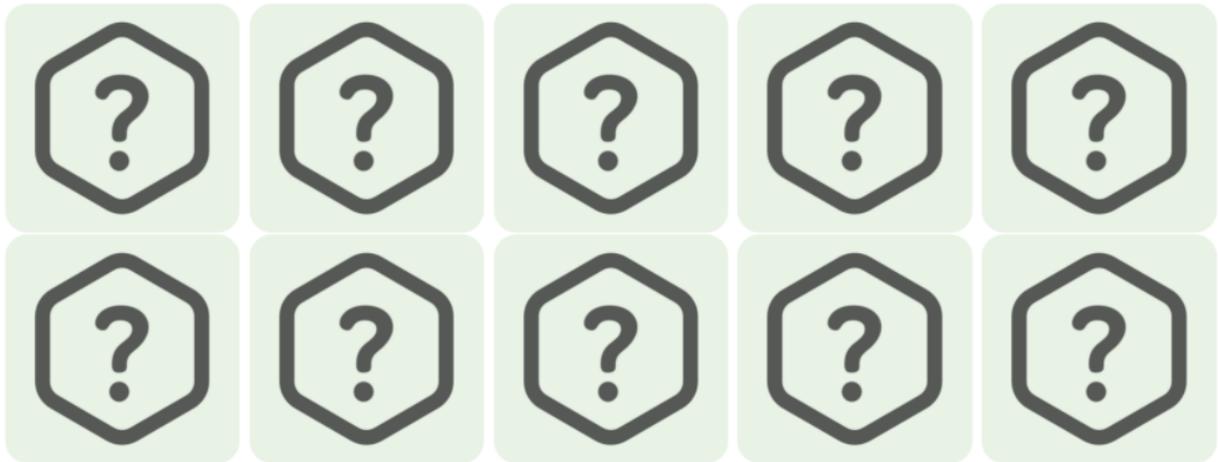


It is important that you remember your color for the rest of the game.

Once you have mentally chosen your card, you can click on "Next" to continue.

[Screen 6 Participant B]

Please pick one card by clicking on it.



Once you have picked a card and seen the color behind it, you can click on the button "Show all" if you want to see what color was behind each card. You don't have to do so; it is only a tool to show you how the cards were distributed.

[Screen 7 Participant B]

*Participant A's report*

Participant A was asked whether the color she/he picked is the color she/he thought of. Participant A reported: "Yes/No".

We now ask you to report whether the color you chose was the same as the color you drew. Please indicate whether the color behind the card you picked is the color that you thought of:

- Yes
- No

[Screen 10]

### **Results**

Participant A and you reported whether the color you picked is the color you thought of.

Participant A reported: "**Yes/No**". You reported: "**Yes/No**".

In the guessing part, you earned **XX** (*only for Participant A*).

After clicking Next, you will fill in a demographic survey to finish the experiment.

Thank you for participation in this study!

## **C.2 No AVOID**

[Screen 2]

## **Instructions**

Welcome to our study!

Please read the instructions carefully.

You will receive £1.15 after you complete the experiment and fill in a final questionnaire.

During the experiment, you will be able to earn additional money depending on the decisions you make. The decisions you make during the game will not be shared with Prolific at any time. We will explain how the game works on the next screen.

If something goes wrong, please make a screenshot and contact us via Prolific.

At the beginning of the experiment, you will be randomly matched with another participant.

The computer will then randomly assign a role to each of you. One participant will be Participant A and the other one will be Participant B. These labels will ensure that neither of you know the identity of the other participant.

The experiment works as follows:

- (i) Participant A sees 5 cards with different colors (black, orange, blue, yellow, and green) and chooses one color in their head.
- (ii) On the next screen, Participant A sees 10 cards with a question mark. Behind each card, there is a color. There are 2 black cards, 2 orange cards, 2 blue cards, 2 yellow cards, and 2 green cards. The cards are placed in a random order. Participant A clicks on a card and the card flips and shows the color behind it. Only Participant A knows the color of the card she/he picked.
- (iii) After Participant A sees the color behind the picked card, she/he reports whether the color of the card she/he picked in the second stage is the same as the color she/he mentally chose in the first stage. Participant A has to report whether the colors of the picked card and the later seen card match. If both colors match, she/he reports "Yes"; otherwise, she/he reports "No". A message will later be sent to Participant B stating whether Participant A reported "Yes" or "No".
- (iv) Then, Participant B sees 5 cards with different colors (black, orange, blue, yellow, and green) and chooses one color. In contrast to Participant A, Participant B reveals the chosen color.
- (v) After choosing a card, Participant B receives the message with the color reported by Participant A.
- (vi) On the next screen, Participant B sees 10 cards with a question mark. Behind each card, there is a color. There are 2 black cards, 2 orange cards, 2 blue cards, 2 yellow cards, and 2 green cards. The cards are placed in a random order. Participant B clicks on a card and the card flips and shows the color behind it.
- (vii) Finally, the computer automatically reports whether the color of the flipped card is the color that Participant B picked in the first stage. This means that, in contrast

to Participant A, Participant B will not report whether the cards matched – the computer will do so on behalf of Participant B.

The reports by Participant A and the computer (on behalf of Participant B) determine the payments in the experiment.

The computer knows the color chosen and the color picked by Participant B. Then, it reports whether the chosen color and the drawn color by Participant B match using this information.

Participant A and the computer report “Yes” or “No” to the following statement: “Please indicate whether the color behind the card you/Participant B picked is the color that you/Participant B thought of.” If either Participant A or the computer reports “Yes” (no matter who), both participants receive £2.50. If both report “No”, both participants receive £0.30. All the possible report combinations are summarized in the table below.

Participant A	Report	Earnings	
	Computer (on behalf of Participant B)	A	B
No	No	£0.30	£0.30
No	Yes	£2.50	£2.50
Yes	No	£2.50	£2.50
Yes	Yes	£2.50	£2.50

There will be no further rounds in this experiment. That is, you will participate in the task described above only once.

Before starting, we’d like you to answer the questions on the next page to check your understanding.

[Screen 3]

1. Imagine the following scenario: Participant A reports No and the computer (on behalf of Participant B) reports Yes. What would the payments be?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

2. Now, imagine the following scenario: Participant A reports Yes and the computer (on behalf of Participant B) reports No. What would the payments be?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

3. What would the payments be if both, Participant A and the computer, report No?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

4. What would the payments be if both, Participant A and the computer, report Yes?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

5. The computer uses the color selected by Participant B to report whether the chosen color and the drawn color match.

- True
- False

*Once you have answered all the questions correctly, you can continue.*

Please note that the experiment does not contain any further comprehension questions or attention checks from this point on!

### **C.3 NO EXTERNALITY**

[Screen 2]

#### **Instructions**

Welcome to our study!

Please read the instructions carefully.

You will receive £1.15 after you complete the experiment and fill in a final questionnaire.

During the experiment, you will be able to earn additional money depending on the decisions you make. The decisions you make during the game will not be shared with Prolific at any time. We will explain how the game works on the next screen.

If something goes wrong, please make a screenshot and contact us via Prolific.

At the beginning of the experiment, you will be randomly matched with another participant.

The computer will then randomly assign a role to each of you. One participant will be Participant A and the other one will be Participant B. These labels will ensure that neither of you know the identity of the other participant.

The experiment works as follows:

- (i) Participant A sees 5 cards with different colors (black, orange, blue, yellow, and green) and chooses one color in their head.
- (ii) On the next screen, Participant A sees 10 cards with a question mark. Behind each card, there is a color. There are 2 black cards, 2 orange cards, 2 blue cards, 2 yellow cards, and 2 green cards. The cards are placed in a random order. Participant A clicks on a card and the card flips and shows the color behind it. Only Participant A knows the color of the card she/he picked.
- (iii) After Participant A sees the color behind the picked card, she/he reports whether the color of the card she/he picked in the second stage is the same as the color she/he mentally chose in the first stage. Participant A has to report whether the colors of the picked card and the later seen card match. If both colors match, she/he reports "Yes"; otherwise, she/he reports "No". A message will later be sent to Participant B stating whether Participant A reported "Yes" or "No".
- (iv) Then, Participant B sees 5 cards with different colors (black, orange, blue, yellow, and green) and chooses one color. In contrast to Participant A, Participant B reveals the chosen color.
- (v) After choosing a card, Participant B receives the message with the color reported by Participant A.
- (vi) On the next screen, Participant B sees 10 cards with a question mark. Behind each card, there is a color. There are 2 black cards, 2 orange cards, 2 blue cards, 2 yellow cards, and 2 green cards. The cards are placed in a random order. Participant B clicks on a card and the card flips and shows the color behind it.
- (vii) Finally, the computer automatically reports whether the color of the flipped card is the color that Participant B picked in the first stage. This means that, in contrast to Participant A, Participant B will not report whether the cards matched – the computer will do so on behalf of Participant B.

The reports by Participant A and the computer (on behalf of Participant B) determine the payments in the experiment.

The computer knows the color chosen and the color picked by Participant B. Then, it reports whether the chosen color and the drawn color by Participant B match using this information.

Participant A and the computer report "Yes" or "No" to the following statement: "Please indicate whether the color behind the card you/Participant B picked is the color that you/Participant B thought of." If either Participant A or the computer reports "Yes" (no matter who), Participant A receives £2.50. If both report "No", Participant A receives £0.30.

On the other hand, the payment for Participant B only depends on the computer's report. Participant B receives £2.50 if the computer reports "Yes", and £0.30 if the computer reports "No". The possible combinations are summarized in the table below.

Report		Earnings	
Participant A	Computer(on behalf of Participant B)	A	B
No	No	£0.30	£0.30
No	Yes	£2.50	£2.50
Yes	No	£2.50	£0.30
Yes	Yes	£2.50	£2.50

There will be no further rounds in this experiment. That is, you will participate in the task described above only once.

Before starting, we'd like you to answer the questions on the next page to check your understanding.

[Screen 3]

1. Imagine the following scenario: Participant A reports No and the computer (on behalf of Participant B) reports Yes. What would the payments be?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

2. Now, imagine the following scenario: Participant A reports Yes and the computer (on behalf of Participant B) reports No. What would the payments be?

- Both would get £0.30.

- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

3. What would the payments be if both, Participant A and the computer, report No?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

4. What would the payments be if both, Participant A and the computer, report Yes?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

5. The computer uses the color selected by Participant B to report whether the chosen color and the drawn color match.

- True
- False

*Once you have answered all the questions correctly, you can continue.*

Please note that the experiment does not contain any further comprehension questions or attention checks from this point on!

#### **C.4 SIMULTANEOUS**

[Screen 2]

##### **Instructions**

Welcome to our study!

Please read the instructions carefully.

You will receive £1.15 after you complete the experiment and fill in a final questionnaire.

During the experiment, you will be able to earn additional money depending on the decisions you make. The decisions you make during the game will not be shared with Prolific at any time. We will explain how the game works on the next screen.

If something goes wrong, please make a screenshot and contact us via Prolific

At the beginning of the experiment, you will be randomly matched with another participant.

The computer will then randomly assign a role to each of you. One participant will be Participant A and the other one will be Participant B. These labels will ensure that neither of you know the identity of the other participant.

The experiment works as follows:

- (i) Participant A and Participant B see 5 cards with different colors (black, orange, blue, yellow, and green) and choose one color in their head.
- (ii) On the next screen, they see 10 cards with a question mark. Each participant sees a different set of cards. Behind each card, there is a color. There are 2 black cards, 2 orange cards, 2 blue cards, 2 yellow cards, and 2 green cards. The cards are placed in a random order. Participants click on a card and the card flips and shows the color behind it. Participants only know the color of the card she/he picked.
- (iii) After participants see the color behind their picked card, they report whether the color of the card they picked in the second stage is the same as the color they mentally chose in the first stage. Participants have to report whether the colors of the picked card and the later seen card match. If both colors match, they report "Yes"; otherwise, they report "No".

The reports by Participant A and B determine the payments in the experiment. Participants report "Yes" or "No" to the following statement: "Please indicate whether the color behind the card you picked is the color that you thought of." If either Participant A or B reports "Yes" (no matter who), both participants receive £2.50. If both report "No", both participants receive £0.30. All the possible report combinations are summarized in the table below.

Participant's report		Earnings	
A	B	A	B
No	No	£0.30	£0.30
No	Yes	£2.50	£2.50
Yes	No	£2.50	£2.50
Yes	Yes	£2.50	£2.50

There will be no further rounds in this experiment. That is, you will participate in the task described above only once.

Before starting, we'd like you to answer the questions on the next page to check your understanding.

[Screen 3]

1. Imagine the following scenario: Participant A reports No and Participant B reports Yes. What would the payments be?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

2. Now, imagine the following scenario: Participant A reports Yes and Participant B reports No. What would the payments be?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

3. What would the payments be if both participants report No?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

4. What would the payments be if both participants report Yes?

- Both would get £0.30.
- Participant A would get £0.30 and Participant B would get £2.50.
- Participant A would get £2.50 and Participant B would get £0.30.
- Both would get £2.50.

5. When does Participant B learn what Participant A has reported?

- Before Participant B reports her or his own card's color.
- After Participant B reports her or his own card's color.

*Once you have answered all the questions correctly, you can continue.*

Please note that the experiment does not contain any further comprehension questions or attention checks from this point on!