The impact of lying aversion and prosocial lies on cheating*

Daniel Parra†

Abstract

This paper examines how prosocial motivations shape lying behavior in a strategic setting where two players privately observe and report a random draw. I develop a theoretical model predicting that individuals will lie less when others can lie on their behalf (strategic substitution), but more when their dishonesty generates positive externalities (prosocial lying). To test these predictions, I design a laboratory experiment using a two-player mind game in which both players benefit if at least one reports a match. Four treatments isolate the roles of strategic avoidance and prosocial incentives. The results show that prosocial motives outweigh lying aversion: participants lie significantly more when their dishonesty benefits others, even when they could avoid lying costs by allowing their partner to lie. These findings suggest that prosocial preferences can substantially offset the psychological costs of lying, but only when individuals' actions are instrumental in producing gains for others.

JEL Codes: C91, D02, D90.

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

Statements and Declarations

Competing Interests: This project was jointly funded by the WZB Berlin Social Science Center, which covered the experimental costs, and the Pontificia Universidad Javeriana, which supported the researcher's time to finalize the manuscript. The author has no relevant material or financial interests related to this paper.

*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, BEEC 2022, the 2021 ESA Global Online Meetings, the Rady School of Management, ASFEE 2022, and ESA Europe Bologna 2022 the for helpful comments. This research used generic funds provided by the WZB Berlin Social Science Center. The usual disclaimer applies.

†Department of Business Administration, School of Economics and Management, Pontificia Universidad Javeriana. E-mail: danielfparra@javeriana.edu.co.

1 Introduction

The subject of lying raises significant ethical and legal questions and evokes strong reactions in debates over political rhetoric, legal testimony, and everyday life. This debate likewise extends to economic markets and institutions. For instance, in his seminal work, Akerlof (1970) emphasizes the central role of dishonesty in markets with asymmetric information. He argues that dishonesty can lead to market failures in the presence of information asymmetries. Specifically, he shows that dishonesty imposes social costs: direct losses borne by deceived parties and indirect costs, notably the erosion of incentives to supply high-quality goods.

Recently, there has been a significant increase in interest in lying and dishonesty (Abeler et al., 2019). Since lying is inherently social, prosocial preferences, central to models of other-regarding behavior, can influence individuals' willingness to lie. In classical economic models, such as that of Akerlof (1970), individuals are assumed to lie when they have monetary incentives. However, evidence on lying has shown that, despite the presence of monetary incentives, the frequency of lying among people in the population is moderate (Abeler et al., 2019). This moderate level of dishonesty can be explained if some individuals exhibit aversion to lying. Theoretical models typically capture lying aversion through a psychological cost term. Psychological costs reflect a disutility from violating internal moral norms of being honest or because they at least want to appear honest. At the same time, prosocial preferences generate a positive utility whenever a lie produces a benefit for another party (e.g., Fehr and Schmidt (1999); Charness and Rabin (2002) Wiltermuth (2011)). When a dishonest act benefits another person, such as a partner or group member, this utility gain may partially offset the psychological cost of lying. In strategic settings, where one player's lie may reduce the burden on another (e.g., allowing them to avoid lying themselves), prosocial and lying-averse motives interact. As the value of the social benefit increases, the threshold at which lying becomes optimal decreases, and vice versa.

This paper examines how prosocial preferences influence lying aversion in strategic settings. Specifically, I explore two primary questions. First, do individuals lie less when another person can lie on their behalf, thereby allowing them to avoid incurring the psychological costs of lying? Second, do people lie more frequently when their lies create positive externalities, and how does this prosocial lying respond to opportunities for strategic avoidance?

To illustrate the interaction between lying aversion and prosocial preferences, consider a car broker selling a used car on behalf of an owner in exchange for a commission. The broker has an incentive to misrepresent the car's quality to increase the sale price and earn a higher commission, but doing so imposes a psychological cost. However, because the lie also benefits the car owner, the broker may experience less guilt, reducing the overall disutility of lying. A similar trade-off arises for a sales representative who can lie to secure a team bonus promised by the CEO if the group reaches a performance threshold. In such cases, the individual weighs the aversion to lying

against the utility from benefiting others. This dynamic is present across a range of everyday settings: a taxpayer who misreports income to increase family benefits; an underage drinker who lies to obtain alcohol for themselves and their friends; or a job reference who embellishes a candidate's qualifications. In general, when a lie benefits others, two opposing forces shape behavior: lying aversion imposes a psychological cost (Abeler et al., 2019; Gneezy et al., 2018; Khalmetski and Sliwka, 2019; Dufwenberg and Dufwenberg, 2018), while prosocial preferences generate positive utility. These other-regarding preferences are a well-established component of economic behavior, as documented by numerous studies (Andreoni, 1990; Andreoni and Miller, 2002; Charness and Rabin, 2002; Bénabou and Tirole, 2006; DellaVigna et al., 2012).

The key insight is that strategic interaction fundamentally alters the lying decision. When lies act as strategic substitutes–such that one person's dishonesty can replace another's–individuals face a trade-off between incurring lying costs and preserving prosocial outcomes. This creates opportunities for strategic moral free-riding, a form of behavior that does not arise in individual (non-strategic) lying contexts.

To address these questions, I develop a theoretical framework that embeds lying costs and prosocial preferences into individual utility. The model yields three key predictions. First, under conditions of strategic substitution, individuals are expected to lie less when they know others can lie on their behalf, as delegating the dishonest act allows them to avoid the psychological cost of lying. Second, consistent with the notion of prosocial lying, misreporting rates should increase when the lie directly benefits another person, since the utility from helping others offsets part of the moral cost. Third, the model predicts conditional prosociality: this increase in lying is most pronounced when the benefits to others are guaranteed rather than merely intended, as only realized social gains can fully counteract the disutility associated with deception.

Previous studies, such as Kerschbamer et al. (2019) and Hurkens and Kartik (2009), used two separate experimental tasks to elicit social preferences and lying aversion. However, as shown by more recent work, this approach introduces measurement error due to the endogeneity of both constructs, which biases estimates of their effects (Gillen et al., 2019; Van Veldhuizen, 2022). To address this issue, my experimental cleanly separates strategic avoidance from prosocial motives.

The experimental design involves a two-player game in which participants first mentally selected one color from a set of five, and then drew one color at random from a deck of cards. The probability of a match between the mentally selected and the drawn color was fixed at 0.2, regardless of the color chosen.. Participants then reported whether their mentally selected color matched the one drawn. The random draw was known only to each individual and did not affect monetary rewards, which depended solely on participants' reports. Participants earn higher rewards if at least one member of the group reports a match, creating a scenario where individuals must balance lying aversion against prosocial benefits. Four treatment conditions allow for

isolating the effects of prosocial lies and strategic avoidance. Specifically, the four treatments are designed as follows: 1) SEQUENTIAL, where participants report their outcomes sequentially, enabling the second mover to condition their report on the first mover's decision; 2) SEQUENTIAL-COMPUTER, where the second mover's truthful report is automated, eliminating the first mover's ability to strategically avoid lying costs; 3) SEQUENTIAL-NOEXTERNALITY, where the first mover's report does not affect the second mover's payoff, removing the prosocial externality associated with reporting a match; and 4) SIMULTANEOUS, where participants report outcomes simultaneously, introducing strategic uncertainty about the partner's reporting behavior.

The experimental results show that the second mover in SEQUENTIAL lied less when the first mover reported that the colors matched.[1] While this finding suggests that first movers may possess a strategic advantage in avoiding lying costs, surprisingly, I found no significant difference in the lying rates of the first movers in SEQUENTIAL and SEQUENTIAL-COMPUTER. The similarly high lying rates in SEQUENTIAL and SEQUENTIAL-COMPUTER suggest 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 SEQUENTIAL-NOEXTERNALITY, where I found that first-movers lie more in SEQUENTIAL-COMPUTER than in SEQUENTIAL-NOEXTERNALITY, indicating that people lied more when they benefited others as well as themselves. Notably, lying rates remain elevated even when individuals have the opportunity to strategically avoid lying costs, suggesting prosociality strongly offsets lying aversion. This result is in line with Wiltermuth (2011), Levine and Schweitzer (2015), and Kerschbamer et al. (2019). 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 SEQUENTIAL and SEQUENTIAL-COMPUTER, more than half of the people did not lie. Finally, lying rates were not lower in SEQUENTIAL than in SIMULTANEOUS. This may indicate that individuals prioritize actual outcomes over intentions, but only when their actions are instrumental in producing those outcomes. This reflects a form of *conditional* or *instrumental prosociality*, rather than pure consequentialist altruism.

This paper contributes to the fast-growing literature on lying behavior. This literature argues that individuals' disutility from dishonesty explains the deviation from the world of universal dishonesty assumed by classical economic theory. Kajackaite and Gneezy (2017) show that individuals follow a cost-benefit analysis to 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 which can be decomposed into intrinsic costs and image concerns. Dufwenberg and Dufwenberg (2018) also explains lying behavior

[1] I conducted an initial experiment where participants' random draws were observed by the experimenter but this design yielded very low lying rates, preventing effective hypothesis testing. Therefore, in this paper I will present the results of the mind game experiment only.

by the cost of lying. However, they argue that the cost of lying increases proportionately to the individual's perceived level of cheating, making the costs extrinsic. This paper extends standard lying-aversion models by incorporating strategic interaction and prosocial preferences. In the experiment, I found evidence that prosocial lying offsets lying costs. Hence, I show that even if the psychological costs reduce lying on average, prosocial lying might offset the impact of lying aversion. This finding does not suggest that lying aversion is unimportant, but rather highlights how prosocial incentives can interact with psychological costs of lying.

A second closely related strand of literature examines collaborative lying (for a survey see Leib et al., 2021). These studies use games in which participants play in groups, and each group member must lie to increase collective earnings (e.g. Conrads et al., 2013; Weisel and Shalvi, 2015; Muehlheusser et al., 2015; Kocher et al., 2018; Rilke et al., 2021). That is, lies function as strategic complements in these settings. Hence, collaborative lying research focuses on situations where coordination in dishonesty is central. This implies that coordinated dishonesty is essential and that relying on others to abstain from lying is not feasible. The collaborative lying games suggest that social preferences make people lie more, implying that cooperation can amplify dishonest behavior. Although I also use a group setting, I study a different situation in which lies are not complementary but substitutes. Thus, individuals can maintain honesty by relying on others to act dishonestly. In contrast, my setting emphasizes prosociality rather than cooperation as the relevant social preference. I show that individuals use prosocial motives to justify dishonest behavior, even when cooperation is not needed to increase their payoffs.

Finally, this paper contributes to the literature examining how other-regarding preferences influence dishonest behavior. Behavioral implications vary depending on the type of lie. For instance, Biziou-van Pol et al. (2015) explore the relationship between white lies and prosocial preferences, showing that different types of white lies–Pareto white lies versus altruistic white lies–reflect distinct underlying moral motivations and prosocial tendencies. Hurkens and Kartik (2009) present evidence that deception is primarily driven by individuals' social preferences over outcomes, rather than by aversion to lying. Similarly, Wiltermuth (2011), Levine and Schweitzer (2015), and Kerschbamer et al. (2019) show that individuals are more likely to lie when their dishonesty benefits others, and less likely to lie when the gain is solely personal. Levine and Schweitzer (2015) further demonstrates that prosocial lies can increase trust in group contexts, potentially explaining their prevalence. This paper extends this evidence by showing that, even when lying can be avoided, the prosocial motive is strong enough to outweigh lying aversion for a significant part of the participants. Moreover, I show that the effect of prosocial behavior on dishonesty is diminished when the actual impact on others' payoffs is uncertain. The findings suggest that when individuals decide whether to lie, their 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 yield

Pareto improvements and why individuals respond primarily to the actual consequences of their actions on others' payoffs.

The paper proceeds as follows. Section 2 presents a theoretical model of lying with psychological lying costs and prosociality. It also introduces four experimental treatments designed to disentangle lying aversion from prosocial motives, along with their respective hypotheses. Section 3 explains the details of the online experiment and its procedures. Section 4 presents experimental evidence that tests the model's hypotheses and discusses the experimental findings and interprets them through the lens of the benchmark model presented in Section 2. It further uses the model to contrast these findings with results from a collaborative lying game. Section 5 concludes.

2 Theoretical framework, experimental design, and hypotheses

Understanding behavior in contexts where lying yields Pareto improvements necessitates extending existing models of lying behavior. Standard models of lying behavior exhibit significant limitations when applied to strategic environments involving other-regarding preferences. Pure lying cost models (Gneezy et al., 2018; Dufwenberg and Dufwenberg, 2018; Abeler et al., 2019; Khalmetski and Sliwka, 2019) successfully explain limited lying behavior when psychological costs are sufficiently high. However, these models cannot explain why lying rates decrease when both the liar and another party benefit. These models focus on individual decision-making and abstract away from strategic interaction that may shape lying behavior. For instance, Gneezy et al. (2018)'s model assumes that lying costs depend solely on individual-specific fixed and outcome-based components, but it does not account for contexts in which lying may generate benefits for others, as it omits prosocial intent. Similarly, Abeler et al. (2019) emphasize truth-telling preferences driven by intrinsic honesty and reputation concerns; nonetheless, their framework does not directly capture settings in which individuals may lie to generate positive externalities.

On the other hand, standard social preference models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002) can partially rationalize increased lying when it produces positive externalities for others, but they do not account for lying aversion. In settings where lies benefit everyone, these models would predict full lying, which contradicts robust empirical evidence of truth-telling even in Pareto-improving contexts. For instance, Fehr and Schmidt (1999) model agents as inequity-averse, meaning they dislike both advantageous and disadvantageous inequality, typically with a stronger aversion to the latter. Therefore, if a report increases both players' utility, everyone should lie. Thus, explaining empirically observed patterns of partial prosocial lying requires incorporating explicit lying costs alongside social preferences.

A key omission in existing frameworks is the interaction between prosocial motivations and lying costs in strategic settings. When lying generates positive externalities, individuals face a

trade-off between lying aversion and prosocial concerns. Moreover, in strategic contexts, reliance on others' prosocial lies introduces a coordination problem absent from existing models. Understanding behavior in such settings requires a framework that integrates three elements: psychological lying costs, prosocial motivations, and strategic reasoning about others' actions.

2.1 Individual's preferences

The present model can be viewed as an extension of the lying-aversion framework of Gneezy et al. (2018) by explicitly incorporating prosocial motivations into the structure of intrinsic lying costs. Specifically, whereas Gneezy et al. (2018) model intrinsic lying costs as functions of individual-specific fixed and outcome-contingent components, the present framework allows these costs to be partially offset when lying generates positive externalities. The analysis focuses on environments in which individuals interact in dyads. Let P_i, where $i \in \{1,2\}$, denote the members of each dyad.

The experiment employs a binary lying game. This format was chosen to focus on the incidence rather than the magnitude of lying. The standard die-roll game, as described in Fischbacher and Föllmi-Heusi (2013), generates unnecessary noise when there are no hypotheses regarding the intensive margin, which weakens the statistical power and does not contribute to the central question of the paper. Each player $i \in \{1,2\}$ privately draws a state $x_i \in \{0,1\}$, with $\Pr(x_i = 1) = 0.2$, and sends a report $r_i \in \{0,1\}$. Their payoffs are interdependent, with both players receiving a monetary payoff

$$v(r_i, r_j) = \begin{cases} v_h & \text{if } r_i + r_j \geq 1, \\ v_l & \text{if } r_i + r_j = 0, \end{cases}$$

In this context, lies are characterized as Pareto White Lies (Erat and Gneezy, 2011) because they benefit both the liar and others.

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, following the standard lying models by Gneezy et al. (2018), Abeler et al. (2019), and Khalmetski and Sliwka (2019).[2] 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 and follows the

[2]The psychological lying cost includes the intrinsic costs of lying. While image concerns could, in principle, contribute to lying costs, they are not captured in the model because they would depend endogenously on equilibrium reporting probabilities.

convention introduced by Gneezy et al. (2018), Abeler et al. (2019), and Khalmetski and Sliwka (2019).

Third, I incorporate prosociality into the utility function, drawing on insights from Wiltermuth (2011), Levine and Schweitzer (2015), and Janezic (2020). To capture the heterogeneity in prosocial motivations observed in real populations, I assume that individuals derive utility θ_i when they benefit another agent with their report, i.e., when they generate a positive externality. The prosocial parameter θ_i is distributed among the population according to

$$\theta_i \sim U[0, \bar{\theta}].$$

This heterogeneity reflects agents' varying concern for others' welfare, analogous to heterogeneity in agents' intrinsic lying costs. I adopt the taxonomy introduced by Sobel (2020), which distinguishes among lying (making a statement one believes to be false), deception (inducing the receiver to form incorrect beliefs), and damage (the harm inflicted on the receiver by the falsehood). Under this framework, any lie that yields a positive recipient payoff ($I_i > 0$) partially offsets the intrinsic moral cost through prosocial utility. Consequently, non-damaging misreports impose a lower net cost, capturing the attenuated moral weight of prosocial lies.

For simplicity, I do not consider the possibility of individuals being prosocial towards the experimenter. While this paper I focus exclusively on positive externalities–where prosociality serves only to reduce the net cost of lying–the framework can be readily extended to allow $\theta_i < 0$, thereby capturing the case where a lie harms others and increases the net psychological cost of deception (as in Gneezy and Kajackaite, 2020; Kerschbamer et al., 2019). I impose

$$0 \leq \bar{\theta} \leq v_h$$

so that the utility from prosocial lying is non-negative but never exceeds the utility of one's own monetary reward. Finally, I assume $c_i(r_i = x_i) = 0$ and

$$v_h - v_l + \bar{\theta} < \bar{c}\,.$$

This last condition guarantees that some individuals with the highest lying costs will always choose to tell the truth, ruling out the trivial equilibrium in which all individuals lie due to negligible lying costs. In other words, it ensures a non-trivial equilibrium by excluding scenarios where prosocial preferences universally dominate lying costs; absent it, even the most morally averse would misreport, contradicting empirical findings that people often tell the truth even when dishonesty yields Pareto improvements.

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 derive utility from their actions' outcomes, rather than from their intentions to be prosocial. Therefore, individuals get the utility from being prosocial only if the marginal benefit of one's report on the partner is positive. Put another way, the utility from prosociality (θ_i) when their partner reports 1 is reduced or eliminated regardless of one's report.[3] This approach is consistent with my conceptual focus on the moral justification of lying, rather than on pure altruistic concern.

To capture this instrumental consequentialism view while allowing for treatment-specific considerations, I introduce an intention-to-help indicator I_i that varies by treatment contexts. This indicator reflects a form of instrumental consequentialism, whereby agents derive utility from prosocial outcomes only when their lie is pivotal in producing them. It is worth emphasizing that the model does not assume standard outcome-based social preferences of the form $U_i = \text{payoff}_i + \theta \cdot \text{payoff}_j$, instead, prosocial motivation enters the model indirectly by offsetting the intrinsic cost of lying: agents find it less morally costly to lie when their misreport generates a positive externality. With these three elements, I represent individuals' preferences using the following function:

$$U_i(x_i, r_i, r_j) = v(r_i, r_j) - \mathbf{1}_{x_i \neq r_i}(c_i - \theta_i \cdot I_i) \tag{1}$$

where $v(r_i, r_j) = v_h$ if $r_i + r_j \geq 1$, and v_l otherwise. I_i captures whether lying has prosocial intent and varies by treatment as specified in the following sections. For algebraic clarity in the subsequent proofs, I normalize the low payoff such that $v_l = 0$. The high payoff v_h then represents the net monetary gain from a successful outcome. This normalization is standard and does not affect the model's qualitative predictions, as all results can be generalized by substituting $v_h - v_l$ for v_h where necessary.

2.2 Treatments

2.2.1 Treatment 1: SEQUENTIAL

In the main treatment, SEQUENTIAL, 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 observing r_1, P_2 draws $x_2 \in \mathcal{X}$ and sends a report $r_2 \in \mathcal{X}$. Note that x_i is known only to P_i, not to the other player.

[3] 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.

In this treatment, the intention-to-help indicator is:

$$I_i = \begin{cases} r_i & \text{for } P_1 \\ r_i \cdot \mathbf{1}_{r_1=0} & \text{for } P_2 \end{cases} \tag{2}$$

I apply backward induction to analyze the game's strategic environment. Importantly, in this game, there is no downward lying in equilibrium. An individual who draws $x_i = 1$ and reports $r_i = 0$ incurs the lying cost without receiving either the monetary payoff or the benefit of a positive externality. Thus, in equilibrium individuals only lie if they draw $x_i = 0$ by reporting $r_i = 1$.

Proposition 1 (P2's Best Response in SEQUENTIAL)**.** *After observing r_1, P_2's best response is:*

- *If $r_1 = 1$: P_2 reports $r_2 = 1$ if $x_2 = 1$, and $r_2 = 0$ if $x_2 = 0$*
- *If $r_1 = 0$: P_2 reports $r_2 = 1$ if $c_2 < v_h + \theta_2$*

Proof. Recall that I have normalized the low payoff to zero, $v_l = 0$.

Case 1: $r_1 = 1$. If $x_2 = 0$ and P_2 reports truthfully ($r_2 = 0$), their payoff is

$$U_2 = v(r_1 = 1, r_2 = 0) = v_h.$$

If instead they lie ($r_2 = 1$), since $I_2 = 0$ there is no prosocial benefit, so

$$U_2 = v_h - c_2.$$

Because $c_2 > 0$, truth-telling ($r_2 = 0$) strictly dominates lying when $r_1 = 1$.

Case 2: $r_1 = 0$ and $x_2 = 0$. Then

$$U_2(r_2 = 0) = v_l = 0, \qquad U_2(r_2 = 1) = v_h - \big(c_2 - \theta_2\big) = v_h - c_2 + \theta_2.$$

Hence P_2 will lie (report $r_2 = 1$) if and only if

$$v_h - c_2 + \theta_2 > 0 \quad \Longleftrightarrow \quad c_2 < v_h + \theta_2.$$

■

Let $\hat{c}_i(\theta_i)$ be the lying cost threshold where individuals with prosociality θ_i are indifferent between lying and telling the truth. For P_2, when P_1 reports 0, I have $\hat{c}_2(\theta_2, r_1 = 0) = v_h + \theta_2$. Hence, the probability that P_2 lies after P_1 reports 0, conditional on prosociality level θ_2, is:

$$\Pr(P_2 \text{ lies} | r_1 = 0, \theta_2) = F(\hat{c}_2(\theta_2, r_1 = 0)) = \frac{v_h + \theta_2}{\bar{c}} \tag{3}$$

Proposition 2 (P1's Equilibrium Strategy in SEQUENTIAL)**.** *P_1 lies (reports $r_1 = 1$ when $x_1 = 0$) if their lying cost c_1 is below a threshold $\hat{c}_1^{\text{SEQUENTIAL}}(\theta_1)$ that depends on their prosociality θ_1:*

$$\hat{c}_1^{\text{SEQUENTIAL}}(\theta_1) = v_h + \theta_1 - \frac{v_h\left(v_h + \bar{\theta}/2\right)}{\bar{c}}. \tag{4}$$

Proof. P_1 anticipates P_2's strategy. Let b_0 denote P_1's belief that P_2 reports 1 after observing $r_1 = 0$, and b_1 be P_1's belief that P_2 reports 1 after $r_1 = 1$. In equilibrium:

$$b_1 = 0.2, \quad b_0 = 0.2 + 0.8 \cdot \frac{v_h + \bar{\theta}/2}{\bar{c}}.$$

P_1 with $x_1 = 0$ lies if

$$v_h - (c_1 - \theta_1)(1 - b_1) > b_0\, v_h,$$

which rearranges to

$$c_1 < \theta_1 + \frac{(1 - b_0)\, v_h}{1 - b_1}.$$

Noting that $1 - b_1 = 0.8$ and $1 - b_0 = 0.8\left(1 - \frac{v_h + \bar{\theta}/2}{\bar{c}}\right)$, we obtain

$$c_1 < v_h + \theta_1 - \frac{v_h\,(v_h + \bar{\theta}/2)}{\bar{c}},$$

This inequality leads to the threshold condition in equation 4. ■

2.2.2 Treatment 2: SEQUENTIAL-COMPUTER

In a second treatment, SEQUENTIAL-COMPUTER, I remove P_1's ability to rely on P_2's possibility of lying by imposing $x_2 = r_2$. To do so, in SEQUENTIAL-COMPUTER, participants in the role of P_2 do not have the opportunity to report their draw, but the computer observes the draw and submits the report on their behalf. This procedure is common knowledge among participants. The payoff structure is the same as in SEQUENTIAL. Even though the computer submits the report, the human participant bears the resulting payoff consequences. Thus, this procedure ensures that P_1 faces an objective probability distribution over r_2. This feature implies that $b_1 = b_0 = 0.2$.

In SEQUENTIAL-COMPUTER, the intention-to-help indicator for P_1 is simply $I_1 = r_1$ since lying always has the potential to help P_2.

Proposition 3 (P1's Equilibrium in SEQUENTIAL-COMPUTER)**.** *P_1 lies if*

$$c_1 < v_h + \theta_1 \tag{5}$$

Proof. P_1 knows $\Pr(r_2 = 1) = 0.2$. Comparing utilities when $x_1 = 0$:

$$\mathbb{E}[U_1(r_1 = 0)] = 0.2\, v_h \tag{6}$$

$$\mathbb{E}[U_1(r_1 = 1)] = v_h - 0.8(c_1 - \theta_1) \tag{7}$$

P_1 lies if $v_h - 0.8(c_1 - \theta_1) > 0.2\, v_h$, which simplifies to $c_1 < v_h + \theta_1$. ■

Comparing the lying cost thresholds presented in (4) and (5), it follows that a greater proportion of P_1 participants will lie when they cannot rely on P_2's incentives to lie. This result holds because the utility from prosociality and the maximum lying cost are non-negative. Given this result, I will test hypothesis 1.

Hypothesis 1 (No cost avoidance)**.** *In* SEQUENTIAL-COMPUTER*, the proportion of P_1 participants who lie will be higher than in* SEQUENTIAL*.*

Hypothesis 1 posits that more participants will misreport their privately observed random draw in SEQUENTIAL-COMPUTER than in SEQUENTIAL. The logic is that, in SEQUENTIAL, P_1 anticipates that P_2–being pivotal when a mismatch is reported–has strong incentives to misreport for the group's benefit; consequently P_1 can strategically avoid the psychological lying costs by letting P_2 lie instead of themselves. In contrast, in SEQUENTIAL-COMPUTER P_2 is a truth-teller by design, and then P_1 faces the full weight of the decision, making them more prone to lie to benefit the group. If the data do not support this hypothesis, it would suggest that intrinsic lying costs are comparatively minor, and that prosocial preferences dominate. This alternative scenario, characterized by equal lying rates across treatments, can be rationalized within the model under a particular parameter configuration described below.

Theorem 1 (Convergence under Large Cost Heterogeneity)**.** *Let $c_i \sim U[0, \bar{c}]$. As the upper bound $\bar{c}$ grows large relative to the sum of the monetary upside and average prosocial benefit,*

$$\frac{\bar{c}}{v_h + \bar{\theta}/2} \longrightarrow \infty,$$

the gap between the two lying-cost thresholds

$$\hat{c}_1^{\text{SEQUENTIAL-COMPUTER}}(\theta_1) - \hat{c}_1^{\text{SEQUENTIAL}}(\theta_1) = \frac{v_h\left(v_h + \bar{\theta}/2\right)}{\bar{c}}$$

tends to zero, and hence the predicted lying rates in both treatments coincide.

Theorem 1 highlights that under a uniform distribution $U[0, \bar{c}]$, increasing $\bar{c}$ does more than raise the upper bound of potential lying costs—it broadens the support of c_i, thereby increasing heterogeneity across individuals. As $\bar{c}$ becomes large, only a vanishingly small fraction of

participants find it optimal to exploit the strategic benefit of cost avoidance, regardless of their prosociality θ_1. In this regime, the incentive to defer one's lie onto P_2 disappears, aligning the equilibrium cutoffs in SEQUENTIAL and SEQUENTIAL-COMPUTER, and resulting in identical aggregate lying rates.

Finally, my framework is based on substitute, not complementary, lies, unlike the collaborative lying games of Weisel and Shalvi (2015). In my model, when both players must lie to secure the higher payoof (v_h) as in Weisel and Shalvi (2015), each agent's prosocial motive activates precisely when the misreport is pivotal. By contrast, in SEQUENTIAL, the first mover can free-ride on the second mover's misreport, thereby reducing the first mover's incentive to lie. Requiring both players to lie eliminates this free-ride option and transforms each lie into a joint undertaking: a cooperative act with assured mutual benefit. Thus, lying by the first mover occurs more frequently in the collaborative game than in the substitute-type game used in my experiment. Appendix B presents the formal demonstration.

2.2.3 Treatment 3: SEQUENTIAL-NOEXTERNALITY

In a third treatment, I investigate the role of positive externalities in P_1's decision-making. I use the same structure as in SEQUENTIAL-COMPUTER, but modify the payoff scheme to eliminate benefits to others. As a result, lies in this treatment are no longer Pareto white lies but instead constitute purely selfish lies. I keep P_1's monetary payoffs identical to those in SEQUENTIAL-COMPUTER, but make P_2's payoffs depend solely on their own draw: specifically, in SEQUENTIAL-NOEXTERNALITY, P_2 receives a payoff of v_h if $x_2 = 1$ and v_l otherwise. This variation implies that, in the utility function, $I_1 = 0$, since lying no longer generates any benefit for others. Conceptually, this treatment approximates a setting akin to Gneezy et al. (2018), in which the value of prosocial lying is zero.

Proposition 4 (P1's Equilibrium in SEQUENTIAL-NOEXTERNALITY)**.** *P_1 lies if $c_1 < v_h$.*

Proof. With $I_1 = 0$, the prosocial benefit vanishes. A P_1 with $x_1 = 0$ compares

$$\mathbb{E}[U_1(r_1 = 0)] = 0.2\, v_h, \tag{8}$$

$$\mathbb{E}[U_1(r_1 = 1)] = v_h - 0.8\, c_1. \tag{9}$$

They lies if

$$v_h - 0.8\, c_1 > 0.2\, v_h \quad \Longrightarrow \quad c_1 < v_h.$$

■

Theorem 2 (SEQUENTIAL-COMPUTER > SEQUENTIAL-NOEXTERNALITY)**.** *The proportion of P_1 participants who lie in* SEQUENTIAL-COMPUTER *exceeds that in* SEQUENTIAL-NOEXTERNALITY.

Proof. From the two cutoff thresholds,

$$\hat{c}_1^{\text{SEQUENTIAL-COMPUTER}}(\theta_1) = v_h + \theta_1, \quad \text{and} \quad \hat{c}_1^{\text{SEQUENTIAL-NOEXTERNALITY}} = v_h,$$

I have that for any $\theta_1 > 0$,

$$\hat{c}_1^{\text{SEQUENTIAL-COMPUTER}}(\theta_1) = v_h + \theta_1 > v_h = \hat{c}_1^{\text{SEQUENTIAL-NOEXTERNALITY}}.$$

Since $\theta_1 \sim U[0, \bar{\theta}]$ with $\bar{\theta} > 0$, a positive mass of players satisfies $\theta_1 > 0$, so aggregate lying is strictly higher in SEQUENTIAL-COMPUTER. ■

Hypothesis 2 (Positive Externality)**.** *In* SEQUENTIAL-NOEXTERNALITY*, the proportion of* P_1 *participants who lie will be lower than in* SEQUENTIAL-COMPUTER.

Treatment SEQUENTIAL-NOEXTERNALITY isolates prosocial motives. Hypothesis 2 therefore predicts a lower misreporting rate in SEQUENTIAL-NOEXTERNALITY, because this treatment removes the prosocial channel. If the data support this hypothesis, it would indicate that participants are more willing to misreport when doing so raises a counterpart's payoff. Conversely, failure to support the hypothesis would suggest that prosocial considerations exert little influence on misreporting behavior.

While I do not directly compare SEQUENTIAL and SEQUENTIAL-NOEXTERNALITY in my empirical analysis, the theoretical comparison is instructive. The prediction is ambiguous and depends on an individual's specific prosociality θ_i. The lying threshold in SEQUENTIAL is higher than in SEQUENTIAL-NOEXTERNALITY if and only if:

$$\theta_i > \frac{v_h(v_h + \bar{\theta}/2)}{\bar{c}}$$

For individuals with sufficiently strong prosocial preferences (i.e., those for whom this condition holds), lying will be more frequent in SEQUENTIAL. However, for individuals with weak prosocial preferences, the strategic incentive to free-ride in SEQUENTIAL can dominate, making them less likely to lie than in the purely selfish NOEXTERNALITY context.

Until now I have introduced the main treatments for disentangling lying aversion and prosocial motives. In SEQUENTIAL-COMPUTER, P_2's report is fixed at 0.2, whereas in SEQUENTIAL it is subjective, granting P_1 greater opportunity to avoid lying costs, while maintaining a prosocial incentive. If lying rates are identical across these two conditions, cost-avoidance effects must be negligible and prosocial motives must fully account for behavior. In SEQUENTIAL-NOEXTERNALITY, the prosocial benefit vanishes ($I_1 = 0$), isolating self-interested lying. Hence, if lying is more frequent in SEQUENTIAL than in SEQUENTIAL-NOEXTERNALITY, positive externalities must be amplifying dishonesty; if it is less frequent, strategic cost avoidance must dominate; and if the rates coincide,

the two forces exactly counterbalance. Table 1 summarizes how each motive is isolated by the experimental design.

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

	SEQUENTIAL	SEQ-COMP[a]	SEQ-NOEXT[b]
Avoid the Lying Costs $P(r_2 = 1 \mid r_1 = 0)$	$0.2 + b_0$	0.2	0.2
Prosociality	✓	✓	×

[a] Sequential-Computer; [b] Sequential-NoExternality.

Note: The row *Sequential Lying Cost* refers to how likely it is to avoid the lying cost while obtaining the high payoff. For SEQUENTIAL, it uses b_0 to represent the subjective probability that P_1 attributes to P_2 reporting 1.

2.2.4 Treatment 4: SIMULTANEOUS

The final treatment, SIMULTANEOUS, preserves the payoff structure of the SEQUENTIAL treatment while eliminating P_1's first-mover advantage. In particular, both players simultaneously choose their report r_i, unaware of the other's decision. As in SEQUENTIAL, a lie can generate a positive externality only if it changes the joint outcome, but now each player is uncertain whether their misreport will be "effective," i.e., change the joint outcome. That is, it depends on whether the partner would otherwise have reported 0. To capture this, I set

$$I_i \;=\; r_i \,\Pr(r_j = 0 \mid r_i) \;=\; r_i\,(1-b),$$

where b denotes the (symmetric) equilibrium probability that the other player reports 1.

Proposition 5 (Symmetric Equilibrium in SIMULTANEOUS)**.** *In any symmetric equilibrium in which each player lies with probability p^* when $x_i = 0$, player i lies if and only if*

$$c_i \;<\; (1-p^*)(v_h + \theta_i),$$

and the equilibrium lying probability p^* is characterized by the fixed-point equation

$$p^* \;=\; \Pr(x_i = 1) \;+\; \Pr(x_i = 0)\,\Pr(\textit{lie} \mid x_i = 0) \;=\; 0.2 \;+\; 0.8\,\mathbb{E}_\theta\!\left[\tfrac{(1-p^*)(v_h+\theta)}{\bar{c}}\right].$$

Since $\theta \sim U[0,\bar{\theta}]$, $\mathbb{E}_\theta[v_h + \theta] = v_h + \bar{\theta}/2$, giving

$$p^* = 0.2 + 0.8\,\frac{(1-p^*)(v_h + \bar{\theta}/2)}{\bar{c}}.$$

Sketch of Proof. Since actions are simultaneous, each player forms the belief $b = p^*$ about their partner's lying probability. When $x_i = 0$, reporting 0 yields expected payoff

$$\mathbb{E}[U_i \mid r_i = 0] = b \cdot v_h,$$

while reporting 1 gives

$$\mathbb{E}[U_i \mid r_i = 1] = v_h - (c_i - \theta_i I_i) = v_h - c_i + \theta_i(1-b).$$

Setting $v_h - c_i + \theta_i(1-b) > b \cdot v_h$ yields the lying cutoff

$$\hat{c}_i(\theta_i) = (1-b)(v_h + \theta_i).$$

Enforcing $b = p^*$ and averaging over $\theta \sim U[0, \bar{\theta}]$ yields the fixed-point condition above. ■

Because $v_h + \bar{\theta} < \bar{c}$ and $\bar{\theta} \leq 1$, we know $p^* < 1$ and the prosocial term $(1-p^*)(v_h + \theta_i)$ is strictly less than $v_h + \theta_i$. Intuitively, each player anticipates that in a fraction p^* of cases their partner will already provide the externality, so the marginal benefit of her own lie is attenuated.

Hypothesis 3 (Strategic uncertainty)**.** *If lying costs are high relative to prosocial benefits, the attenuation of the externality in* SIMULTANEOUS *induces more lying than in* SEQUENTIAL. *Conversely, if prosocial motives dominate, the reduced effectiveness of any single lie may suppress dishonesty relative to* SEQUENTIAL.

In the SIMULTANEOUS treatment, each participant submits a report without observing the counterpart's report, creating uncertainty about whether their lie will be pivotal in securing he higher joint payoff. Hypothesis 3 therefore tests conditional prosociality. If lying rate is higher in SIMULTANEOUS, this would indicate either (i) that participants misreport primarily for self-interest—seeking to avoid intrinsic lying costs—or (ii) that prosocial utility exhibits a warm-glow form in which intentions drive utility. Conversely, if lying is lower, it implies that participants misreport only when doing so is clearly instrumental in improving the counterpart's payoff.

Table 2 summarizes, for each of the four treatments, the decision nodes, the corresponding payoff structure, and the theoretical predictions. Finally, note that direct treatment comparisons in which only one factor changes are possible only for the following pairs: SEQUENTIAL-SEQUENTIAL-COMPUTER, SEQUENTIAL-SIMULTANEOUS, SEQUENTIAL-COMPUTER-SEQUENTIAL-NOEXTERNALITY. In the SEQUENTIAL-SEQUENTIAL-NOEXTERNALITY, SEQUENTIAL-COMPUTER-SIMULTANEOUS, and SEQUENTIAL-NOEXTERNALITY-SIMULTANEOUS comparisons, multiple factors vary simultaneously, hindering empirical identification of causal effects.

Table 2. Summary of actions, payoffs, and predictions in each treatment

Treatment	P_1	P_2	Payoffs	Prediction
SEQUENTIAL	reports r_1	learns r_1 and then reports r_2	$v_i = \begin{cases} v_l & \text{if } r_1 = r_2 = 0, \\ v_h & \text{otherwise} \end{cases}$	baseline
SEQ-COMP[a]	reports r_1	learns r_1, but the report is made by the computer	$v_i = \begin{cases} v_l & \text{if } r_1 = r_2 = 0, \\ v_h & \text{otherwise} \end{cases}$	P_1 lies more than in SEQUENTIAL
SEQ-NOEXT[b]	reports r_1	learns r_1, but the report is made by the computer	$v_1 = \begin{cases} v_l & \text{if } r_1 = r_2 = 0, \\ v_h & \text{otherwise} \end{cases}$, $v_2 = \begin{cases} v_l & \text{if } r_2 = 0, \\ v_h & \text{if } r_2 = 1 \end{cases}$	P_1 lies less than in SEQ-COMP[a]
SIMULTANEOUS	both report r_i at the same time		$v_i = \begin{cases} v_l & \text{if } r_1 = r_2 = 0, \\ v_h & \text{otherwise} \end{cases}$	ambiguous relative to SEQUENTIAL

[a] Sequential-Computer; [b] Sequential-NoExternality.

3 The mind-cheating game

The treatments described in Section 2.2 were initially implemented in an online experiment in which the experimenter observed both the random draw and the participant's report. A detailed explanation of this version of the game, along with its main results, is provided in Appendix A. A comparison between the two versions of the game can be found in Parra (2024).

3.1 Overview and design

The game used in the experiment is called the mind-cheating game[4] in which participants choose one color out of five in their minds (see colors in Figure 1). The colors were chosen such that individuals with color vision deficiencies can distinguish them. Then, they draw a color from a deck of cards displayed on their computer screens. 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 reward structure assigns payoffs of $v_h = \$2.5$ and $v_l = \$0.3$. 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, individual-level cheating cannot be directly observed. A

[4] Mind games have previously been 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).

key advantage of the mind-cheating game is that, regardless of the color a participant mentally selects, the probability of a match remains constant and follows a binomial distribution with a known parameter $p = 0.2$.

Figure 1. Colors used in mind game

In the experiment, I use the treatments presented in Table 2. Specifically, in SEQUENTIAL, 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 SEQUENTIAL-COMPUTER, 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 SEQUENTIAL-NOEXTERNALITY, decisions are identical to SEQUENTIAL-COMPUTER, 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.[5]

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 threshold, the robot guesses r_2 for P_1. If

[5]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.

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.[6] I did not run the whole experiment in one session to avoid overloading the server and minimize the probability of technical issues. Table C.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 analyze the reports at an aggregate level. Theoretically, the random draw follows a binomial distribution with a 0.2 probability of the high-paying state. Therefore, I will use this theoretical distribution as a benchmark under truth-telling. Table 3 presents the mean and standard deviation of the main outcome variables. *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 elicitation task. Finally, *Time Spent Reporting* is the time in seconds a participant took to report whether their colors match.

[6] A total of 1009 people showed up, but some left in the middle of the session

Table 3. Summary Statistics of Main Variables in the Mind Game

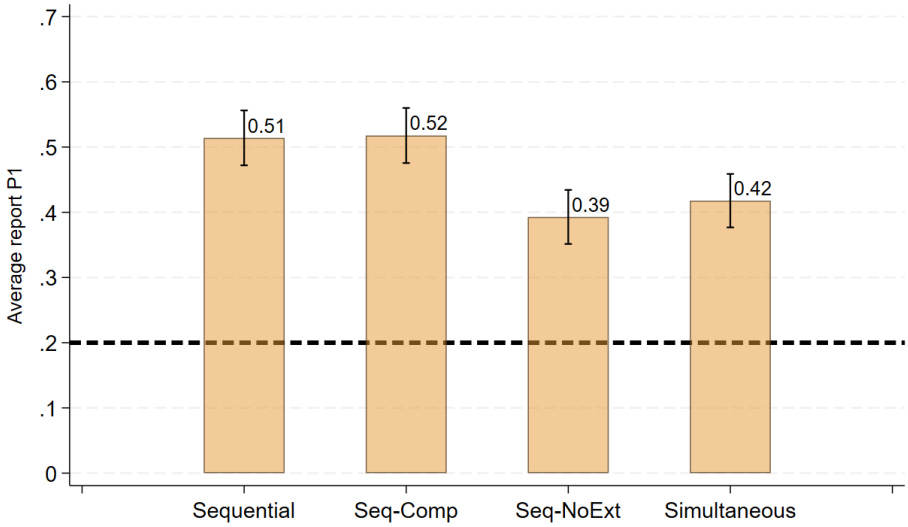
Variable	Sequential	Seq-Comp[a]	Seq-NoExt[b]	Simultaneous
P_1's report = 1	0.514	0.518	0.393	0.418
	(0.502)	(0.501)	(0.490)	(0.495)
Yes is more likely	0.451	0.326	0.250	0.527
	(0.499)	(0.471)	(0.435)	(0.501)
Belief about Pr($r_2 = 1$)	53.148	46.809	42.843	59.144
	(24.353)	(24.138)	(23.370)	(24.835)
Times clicked info	0.338	0.376	0.371	0.329
	(0.504)	(0.515)	(0.514)	(0.527)
Time spent reporting	15.796	16.823	18.957	9.473
	(11.423)	(9.821)	(25.469)	(6.952)
Observations	142	141	140	146

Notes: Standard deviations in parentheses. The table reports summary statistics for participants in the role of P_1 only. For the Sequential treatment, all participants are included since roles are not differentiated in that condition.
a Sequential-Computer; b Sequential-NoExternality.

A visual representation of the P_1's report is presented in Figure 2. 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 2 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 SEQUENTIAL, 40% in SEQUENTIAL-COMPUTER, 23.75% in SEQUENTIAL-NOEXTERNALITY, and 27.50% in SIMULTANEOUS. The pairwise comparisons using one-sided Fisher Exact test show that the difference between SEQUENTIAL and SEQUENTIAL-COMPUTER is not significant ($p = 0.523$), the difference between SEQUENTIAL-COMPUTER and SEQUENTIAL-NOEXTERNALITY is significant ($p = 0.024$), and the difference between SEQUENTIAL and SIMULTANEOUS is not significant ($p = 0.064$).

I use a Linear Probability Model 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, including gender, age, number of experiments in which they have participated, and their student status. Table 4 also reports the mean of r_1 for the SEQUENTIAL treatment in the row labeled SEQUENTIAL *mean*. This information makes it easier to interpret the coefficients of the treatment variables. Additionally, in the row SEQUENTIAL-COMPUTER *vs.* SEQUENTIAL-NOEXTERNALITY, Table

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



Note: The dashed horizontal line displays the underlying theoretical proportion of *Yes* under truth-telling. Error bars represent ±1 standard error of the mean.

4 reports the p-value from a Chi-square test evaluating the equality of coefficients of SEQUENTIAL-COMPUTER and SEQUENTIAL-NOEXTERNALITY. This test was performed post-estimation because, as explained before, a direct comparison between SEQUENTIAL and SEQUENTIAL-NOEXTERNALITY is not clean. The regressions *Report* P_1 *1* and *Report* P_1 *2* confirm that there is no difference in lying between SEQUENTIAL and SEQUENTIAL-COMPUTER, as shown in Figure 2, leading to Result 1.

Result 1 (Related to Hypothesis 1)**.** P_1 *lying behavior is not different in* SEQUENTIAL *than in* SEQUENTIAL-COMPUTER.

I hypothesized that most of the participants assigned the role of P_1 in the treatment SEQUENTIAL would try to avoid incurring their lying costs and pass the burden to P_2. However, Result 1 showed similar lying rates in SEQUENTIAL and SEQUENTIAL-COMPUTER. This finding was contrary to our hypothesis and suggested that either P_1 in SEQUENTIAL expected P_2 to be honest and secured the highest payoff, or they valued the prosocial lie enough to lie.

One might ask whether the similar misreporting rates in the SEQUENTIAL and SEQUENTIAL-COMPUTER treatments point to social motives more complex than the baseline model captures. For instance, if P_1 chooses not to report a match, might this provoke negative reciprocity from P_2, who could reason, "Why should I misreport on our behalf if you will not?" While such retaliation is plausible, three pieces of evidence suggest it is not the dominant force at play. First, the belief-elicitation data reveal no systematic downward revision of P_1's expected honesty when they forgo the misreport, indicating that P_2s were not anticipating punitive behavior. Second, P_2's own misreporting frequency actually increases—rather than decreases—after an honest report, which

Table 4. Regressions testing the differences across treatments

	Report P_1 1	Report P_1 2	Report P_2 SEQ[a]
SEQ-COMP[b]	0.004	0.001	
	(0.058)	(0.059)	
SEQ-NOEXT[c]	-0.121**	-0.127**	
	(0.059)	(0.058)	
SIMULTANEOUS	-0.096	-0.102*	
	(0.059)	(0.058)	
P_1's report=1			-0.242***
			(0.083)
SEQUENTIAL mean	0.514	0.511	
SEQ-COMP vs. SEQ-NOEXT	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 SEQUENTIAL uses data from P_2 in SEQUENTIAL. Bootstrap standard errors are presented in parentheses.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
[a] SEQUENTIAL; [b] SEQUENTIAL-COMPUTER; [c] SEQUENTIAL-NOEXTERNALITY.

is inconsistent with simple negative reciprocity. Third, as shown in Theorem 1, the baseline model already accounts for the absence of a treatment difference when prosocial benefits vary sufficiently, obviating the need for additional parameters.

The row SEQUENTIAL-COMPUTER *vs.* SEQUENTIAL-NOEXTERNALITY reports the p-value from a post-estimation test comparing these two conditions. I use this post-estimation test because SEQUENTIAL-NOEXTERNALITY is only directly comparable with SEQUENTIAL-COMPUTER. These p-values reveal that P_1 lying is lower in SEQUENTIAL-NOEXTERNALITY than in SEQUENTIAL-COMPUTER, which is consistent with hypothesis 2. Specifically, participants are approximately 12% less likely to report a color match when reporting 1 yields no benefit to a third party. These results support prosociality as a primary driver of increased lying in the presence of external beneficiaries.

Result 2 (Related to Hypothesis 2)**.** *P_1 lies less in* SEQUENTIAL-NOEXTERNALITY *than in* SEQUENTIAL-COMPUTER.

A third finding from Table 4 is that the effect of the SIMULTANEOUS treatment on lying is statistically significant only at the 10% level. In regression *Report P_1 1*, the coefficient on SIMULTANEOUS is not statistically significant when individual fixed effects are excluded. However, once demographic controls are included, P_1 participants are less likely to lie in the SIMULTANEOUS treatment than

in SEQUENTIAL. The point estimate suggests that participants reported matching colors approximately 10 percentage points less often in SIMULTANEOUS. Although the effect is only marginally significant, its negative sign aligns with the prosocial-dominant interpretation of Hypothesis 3, which predicts fewer lies in SIMULTANEOUS when the reduced pivotality of a lie outweighs the incentive to avoid lying costs.

Result 3 (Related to Hypothesis 3)**.** P_1 *does not lie less in* SEQUENTIAL *than in* SIMULTANEOUS.

Thus, the paper's key insight is that prosociality can offset lying costs for lying-averse individuals if their lies benefit both themselves and others. As a result, they are more likely to lie in such situations. This finding extends prior results by Wiltermuth (2011), Hurkens and Kartik (2009), Levine and Schweitzer (2015), Biziou-van Pol et al. (2015), and Kerschbamer et al. (2019) by showing that individuals are more willing to lie when doing so creates a positive externality, and that this motive is sufficiently strong to induce dishonesty even among those with high lying aversion. The question then becomes how the SIMULTANEOUS treatment contributes to this pattern. It arguably serves two roles in supporting the paper's central claim. First, it aligns with the assumption embedded in the utility framework: that participants derive utility from the consequences of their actions, but only when those actions are instrumental in producing outcomes. These conditional prosocial preferences reflect a form of instrumental consequentialism, which emphasizes the effectiveness of one's actions in benefiting others—distinct from both deontological ethics and unconditional altruistic preferences. Although the SIMULTANEOUS treatment was not explicitly designed to test this assumption, Result 3 indicates that participants are more likely to lie when their misreport is pivotal in generating benefits for others, providing empirical support for the model of conditional prosociality. In other words, having similar lying rates in SEQUENTIAL and SEQUENTIAL-COMPUTER but not less lying in SIMULTANEOUS than in SEQUENTIAL would have been contradictory in an standard model of lying with warm glow altruistic preferences.[7]

Besides the results concerning P_1's reports, Table 4 also presents important evidence regarding P_2's reports. Column *Report* P_2 presents the relationship between r_1 and r_2. This regression includes only data from the SEQUENTIAL treatment, as it is the only condition in which P_2 can lie after observing P_1's report. The coefficient on P_1*'s report = 1* indicates that P_2 was significantly more likely to report *Yes* when P_1 also reported *Yes*, relative to when P_1 reported *No*. The mean outcome when P_1*'s report = 0* is 0.373, which contextualizes the magnitude of the estimated effect. Using the Bayesian method proposed by Hugh-Jones (2019), I estimate P_2's lying rates conditional on r_1. The estimated lying rate when P_2 observed $r_1 =$ *Yes* is 9.39%, compared to 36.63% when $r_1 =$ *No*.

[7] Setting $I_i = 1$ for every lie collapses the model to a standard *warm-glow* specification with a fixed lying cost: $U_i = v(\mathbf{r}) - c_i \, 1_{\{x_i \neq r_i\}} + \theta_i \Delta v_j$. The resulting thresholds are identical in SEQUENTIAL-COMPUTER and SEQUENTIAL ($\hat{c}_1 = v_h + \theta_1 v_h$) because a single lie deterministically benefits both players in both treatments. For SIMULTANEOUS the threshold becomes $\hat{c}_1 = v_h + \theta_1(1-b)v_h$, so the model predicts *more* lying than in SEQUENTIAL—contrary to Result 3. The pivotality-adjusted cost term $c_i - \theta_i I_i$ restores the correct ranking because $I_i = (1-b) < 1$ only in SIMULTANEOUS.

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

The results suggest that participants face very high intrinsic lying costs. This strong intrinsic aversion to dishonesty eliminates strategic free-riding behavior, rendering the SEQUENTIAL and SEQUENTIAL-COMPUTER treatments behaviorally equivalent for P_1. However, the prosocial motive—even if infrequently activated—nonetheless increases the attractiveness of misreporting relative to purely selfish incentives, explaining why P_1's lying rates in these two treatments exceed those in the SEQUENTIAL-NOEXTERNALITY treatment. Finally, the uncertainty inherent in the SIMULTANEOUS treatment yields a weaker incentive to lie than the certainty present in SEQUENTIAL.

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 SEQUENTIAL, 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 at least 24.65% of the groups, no one lied.[8] Additionally, one can see that despite SIMULTANEOUS creating the highest number of groups with both participants reporting *Yes*, the proportion of groups with two reports of *No* was similar as in SEQUENTIAL-COMPUTER.

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

$r_1 + r_2$	SEQUENTIAL		SEQ-COMP[a]		SEQ-NOEXT[b]		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

[a] Sequential-Computer; [b] Sequential-NoExternality.

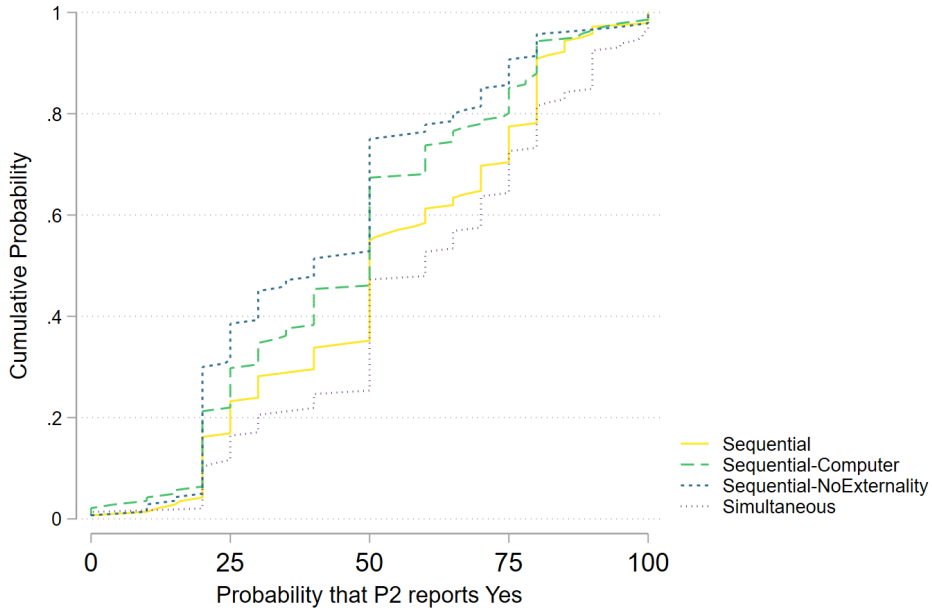
[8] When both players report *No*, we can be confident that the dyad is honest, as downward lying is a dominated strategy. However, if a dyad includes one or two *Yes* reports, we cannot determine whether those reports are truthful. Therefore, the 24.65% of dyads in the SEQUENTIAL treatment with two *No* reports should be interpreted as a lower bound on the proportion of truthful dyads.

4.1 Secondary outcomes

Another outcome of the experiment was the elicited beliefs about P_2's report. Figure 3 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 SEQUENTIAL-SIMULTANEOUS and SEQUENTIAL-COMPUTER-SEQUENTIAL-NOEXTERNALITY, there is no statistically significant difference, but in the pair SEQUENTIAL-SEQUENTIAL-COMPUTER, there is a weakly significant difference. This test implies that there are lower belief levels in SEQUENTIAL-COMPUTER than in SEQUENTIAL ($p = 0.084$), but this result is mechanical from the experimental design.

Another outcome of the experiment concerns participants' elicited beliefs about P_2's report. Figure 3 presents the cumulative distribution function of the implied probabilities that P_2 reports *Yes* across treatments. I use the Kolmogorov-Smirnov test to assess whether these distributions differ. Pairwise comparisons show no statistically significant difference between the distributions in the SEQUENTIAL-SIMULTANEOUS and SEQUENTIAL-COMPUTERVSSEQUENTIAL-NOEXTERNALITY pairs. However, the comparison between SEQUENTIAL and SEQUENTIAL-COMPUTER yields a marginally significant difference ($p = 0.084$). This finding suggests that belief levels are lower in SEQUENTIAL-COMPUTER than in SEQUENTIAL, although this difference is largely mechanical due to the experimental design.

Figure 3. Cumulative Density Function of P_1's subjective probability that P_2 reports *Yes*



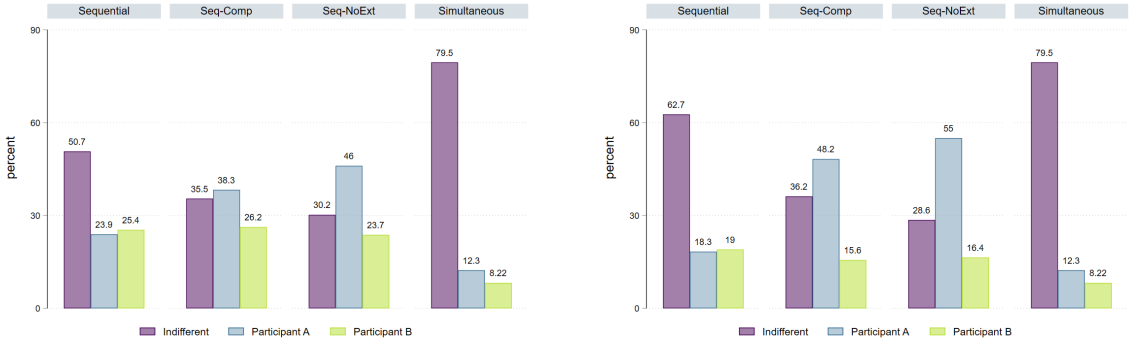
To deepen the insights from Figure 3, I use OLS regressions to study the differences in beliefs

across treatments. Table C.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 C.3 shows that those participants in SIMULTANEOUS who reported *No* believed that their partner was more likely to report *Yes* than *No*. Table C.3 also shows that beliefs were not self-serving in the mind-cheating game.[9]

Result 5 (Beliefs in mind game)**.** *P_1's subjective probability of $r_2 = 1$ is positively correlated with r_1. 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 participants: "*Imagine you were to play the same game again and had a choice, would you rather be Participant A or Participant B?*[10]" Their responses, distinguished by the role they had, are presented in Figure 4. Figure 4(a) shows that P_1 did not, in general, interpret being the first mover as an advantage in SEQUENTIAL or SIMULTANEOUS. Figure 4(b) shows that this is also true for P_2 with even more people being completely indifferent between the two roles. Figure 4(b) also shows that P_2 did not like the second mover position when they could not report and preferred being P_1.

Figure 4. Which role would participants choose if they play again?



(a) P_1 (b) P_2

5 Conclusions

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. Gneezy et al. (2018), Abeler et al. (2019), and Khalmetski and

[9] Using the Kolmogorov-Smirnov test to assess that there are no differences in the distributions gives the same results.

[10] This was the exact wording I used in both experiments to refer to P_1 and P_2.

Sliwka (2019) suggested that psychological lying cost explains why people do not lie to maximize their monetary payoff. However, it is unclear whether this effect holds in group settings where prosocial lying enters the picture. For instance, lying to get an individual bonus from a CEO is different from lying to reach a threshold that gives the bonus to a team. A similar dilemma arises when people use intermediaries for tasks such as completing tax declarations, selling a car, or selling a house. In these contexts, the intermediary has a higher utility if they lie. Therefore, when individuals lie to benefit both others and themselves, there are two competing motivators: lying aversion and prosociality.

This paper introduced a unified framework that embeds intrinsic lying costs and prosocial motivations into a strategic reporting environment. Building on a model that predicts how strategic substitution and positive externalities alter equilibrium lying, I tested three core hypotheses across four treatments. My experimental evidence shows that prosocial incentives can substantially offset psychological lying costs: subjects lie significantly more when their dishonesty benefits others, even when they could avoid lying by free-riding on a partner's lie.

An additional finding was that prosocial lying appears to align more closely with *instrumental consequentialism* than with deontological ethics. Specifically, in the theoretical model, prosociality can only outweigh lying aversion if individuals care about the outcomes their actions bring about, and only when their actions are pivotal to those outcomes. This suggests that moral justification for lying may depend on whether the lie is effective in helping others. However, the scope of this study was limited in evaluating whether instrumental consequentialism alone accounts for the observed behavior, and the experiment did not directly test this assumption. 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. Testing whether participants would be more prone to avoid lying costs in SEQUENTIAL when beliefs were elicited before reporting was beyond this paper's scope.

In spite of the mentioned limitations, the study certainly adds to the understanding of the role of prosocial lies 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 anti-fraud measures should reduce reliance on intermediaries or collective reporting procedures. Second, organizations designing team-based incentives must account for the fact that prosocial motives may encourage dishonesty when group gains are salient. Individual reporting should always be preferred over creating dependencies between people's reports.

These findings yield novel insights into dishonesty and prosociality; yet several open questions remain. First, it is essential to directly assess whether prosocial lying is driven by intentions or evaluated by consequences. While this paper offers evidence consistent with instrumental

consequentialist motivations, further research is needed to confirm and replicate these findings. Second, future research should examine fairness-based retaliation in sequential reporting settings. In our SEQUENTIAL treatment, the first mover can free-ride on the second mover's willingness to misreport, effectively behaving as a *moral free rider*. However, such free-riding may itself be perceived as a norm violation: the second mover might interpret the first mover's reluctance to bear lying costs as opportunistic unfairness and retaliate by withholding cooperation or even punishing the free rider. Isolating this dynamic–where reciprocity and fairness concerns trigger retaliatory responses–could illuminate the interplay between intrinsic lying costs, social preferences, and strategic punishment in dyadic settings. Third, future work might explore how the timing of belief elicitation influences strategic lying avoidance.

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.

Bénabou, R. and Tirole, J. (2006). Incentives and Prosocial Behavior. *The Amercian Economic Review*, 96(5):1652–1678.

Biziou-van Pol, L., Haenen, J., Novaro, A., Liberman, A. O., and Capraro, V. (2015). Does telling white lies signal pro-social preferences? *Judgment and Decision Making*, 10(6):538–548.

Bolton, G. E. and Ockenfels, A. (2000). Erc: A theory of equity, reciprocity, and competition. *American economic review*, 91(1):166–193.

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.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The quarterly journal of economics*, 114(3):817–868.

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.

Garbarino, E., Slonim, R., and Villeval, M. C. (2019). Loss aversion and lying behavior. *Journal of Economic Behavior & Organization*, 158:379–393.

Gillen, B., Snowberg, E., and Yariv, L. (2019). Experimenting with measurement error: Techniques with applications to the caltech cohort study. *Journal of Political Economy*, 127(4):1826–1863.

Gneezy, U. and Kajackaite, A. (2020). Externalities, stakes, and lying. *Journal of Economic Behavior & Organization*, 178:629–643.

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*.

Hurkens, S. and Kartik, N. (2009). Would i lie to you? on social preferences and lying aversion. *Experimental Economics*, 12(2):180–192.

Janezic, K. A. (2020). Heterogeneity in lies and lying preferences. Technical report, Working paper.

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.

Kerschbamer, R., Neururer, D., and Gruber, A. (2019). Do altruists lie less? *Journal of Economic Behavior & Organization*, 157:560–579.

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.

Parra, D. (2024). Eliciting dishonesty in online experiments: The observed vs. mind cheating game. *Journal of Economic Psychology*, 102:102715.

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.

Sobel, J. (2020). Lying and deception in games. *Journal of Political Economy*, 128(3):907–947.

Van Veldhuizen, R. (2022). Gender differences in tournament choices: Risk preferences, overconfidence, or competitiveness? *Journal of the European Economic Association*, 20(4):1595–1618.

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.[11] 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**
SEQUENTIAL	22	23	22	23	35	125
SEQUENTIAL-COMPUTER	22	21	21	22	35	121
SEQUENTIAL-NOEXTERNALITY	23	24	22	21	37	127
SIMULTANEOUS	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.

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

[11] A total of 899 people showed up, but some left in the middle of the session.

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 A.2 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.[12]

Table A.2. Summary statistics Observed Game

	Sequential	Seq-Comp[a]	Seq-NoExt[b]	Simultaneous
Lied=1	0.080	0.056	0.126	0.038
	(0.272)	(0.230)	(0.333)	(0.192)
Student Status	0.472	0.437	0.512	0.455
	(0.501)	(0.498)	(0.502)	(0.500)
Age	26.512	26.024	25.063	26.811
	(8.395)	(8.235)	(7.433)	(9.414)
Gender	0.504	0.492	0.583	0.500
	(0.548)	(0.562)	(0.635)	(0.586)
N	125	126	127	132

Standard deviations in parenthesis.

[a] Sequential-Computer; [b] Sequential-NoExternality.

To confirm the result implied by Table A.2, 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 SEQUENTIAL, SEQUENTIAL-COMPUTER, and SIMULTANEOUS, with SEQUENTIAL as the reference treatment. Lied 2 uses data from treatments SEQUENTIAL, SEQUENTIAL-COMPUTER, and SEQUENTIAL-NOEXTERNALITY, with SEQUENTIAL-COMPUTER as the reference treatment. I use two regressions because SEQUENTIAL is not directly comparable with SEQUENTIAL-NOEXTERNALITY which make it impossible to include SEQUENTIAL-NOEXTERNALITY in Lied 1. The same intuition applies to Lied 2 because between SEQUENTIAL-COMPUTER 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.

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

[12] To ensure that only one component changes, I only compare treatments in the following pairs: SEQUENTIAL-SEQUENTIAL-COMPUTER, SEQUENTIAL-SIMULTANEOUS, and SEQUENTIAL-COMPUTER-SEQUENTIAL-NOEXTERNALITY.

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

	Lied 1	Lied 2	Beliefs 1	Beliefs 2
SEQUENTIAL	*Reference*	0.013	*Reference*	7.875***
		(0.041)		(2.616)
SEQUENTIAL-COMPUTER	-0.018	*Reference*	-7.607***	*Reference*
	(0.037)		(2.759)	
SEQUENTIAL-NOEXTERNALITY		0.061		3.183
		(0.044)		(2.651)
SIMULTANEOUS	-0.030		-1.191	
	(0.033)		(3.159)	
Time Spent Reporting	0.009**	0.012***		
	(0.004)	(0.004)		
Lied=1			24.105***	9.199**
			(5.414)	(4.522)
Constant	-0.017	-0.035	33.189***	26.030***
	(0.061)	(0.082)	(3.922)	(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 SEQUENTIAL, SEQUENTIAL-COMPUTER, and SIMULTANEOUS. Lied 2 and Beliefs 2 use data from treatments SEQUENTIAL, SEQUENTIAL-COMPUTER, and SEQUENTIAL-NOEXTERNALITY. 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.
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Squares to assess whether the reported probability of reporting $Orange$ varies across treatments. Belief 1 uses data from treatments SEQUENTIAL, SEQUENTIAL-COMPUTER, and SIMULTANEOUS with SEQUENTIAL as the reference treatment. Beliefs 2 uses data from treatments SEQUENTIAL, SEQUENTIAL-COMPUTER, and SEQUENTIAL-NOEXTERNALITY with SEQUENTIAL-COMPUTER as the reference treatment. Beliefs 1 show that the reported probability in SEQUENTIAL-COMPUTER is lower than in SEQUENTIAL while the difference between SEQUENTIAL and SIMULTANEOUS is non-significant. Column Beliefs 2 shows that the difference between SEQUENTIAL-COMPUTER and SEQUENTIAL-NOEXTERNALITY is non-significant neither.

Appendix B Comparison with collaborative lying games

Finally, I would like to compare the results of this paper theoretically with collaborative lying games (Leib et al., 2021). In particular, I want to use the theoretical model of section 2 to compare

my results in SEQUENTIAL with a game with a similar structure to the classical dyadic game of Weisel and Shalvi (2015). Imagine the same game as in SEQUENTIAL; however, instead of being paid v_h if at least one group member reports $r_i = 1$, participants only get the monetary reward if both members of the group report $r_i = 1$. Formally, the monetary payoff function becomes

$$v(r_i, r_j) = \begin{cases} v_h, & r_i + r_j = 2, \\ v_l, & \text{otherwise}, \end{cases}$$

and, as in Section 2, we normalize $v_l = 0$. The intention-to-help indicator is

$$I_i \;=\; r_i \, r_j,$$

so that the full prosocial offset θ_i applies only when one's lie is pivotal in securing the joint reward.

Proposition 6 (P_2's Best Response in COLLABORATIVE LYING)**.** *After observing r_1, Player 2's lying-cost cutoff is*

$$\hat{c}_2(r_1) = \begin{cases} v_h + \theta_2, & r_1 = 1, \\ 0, & r_1 = 0, \end{cases}$$

and hence

$$\Pr\big(P_2 \text{ lies} \mid r_1 = 1\big) = 0.8\,\frac{v_h + \bar{\theta}/2}{\bar{c}}, \qquad \Pr\big(P_2 \text{ lies} \mid r_1 = 0\big) = 0.$$

Proposition 7 (P_1's Equilibrium Strategy in COLLABORATIVE LYING)**.** *Player 1 lies (reports $r_1 = 1$ when $x_1 = 0$) if and only if*

$$c_1 < \hat{c}_1^{\,\text{COLLABORATIVE}}(\theta_1),$$

where

$$\hat{c}_1^{\,\text{COLLABORATIVE}}(\theta_1) = 0.8\,\frac{v_h + \bar{\theta}/2}{\bar{c}}\,\big(v_h + \theta_1\big).$$

Proof. When $x_1 = 0$, Player 1's expected utility from truth-telling is

$$U_1(r_1 = 0) = 0,$$

whereas from lying it is

$$U_1(r_1 = 1) = \Pr(P_2 \text{ lies} \mid r_1 = 1)\, v_h \;-\; \big(c_1 - \theta_1 \Pr(P_2 \text{ lies} \mid r_1 = 1)\big).$$

Substituting $\Pr(P_2 \text{ lies} \mid r_1 = 1) = 0.8(v_h + \bar{\theta}/2)/\bar{c}$ and setting $U_1(1) > 0$ yields

$$0.8\,\frac{v_h + \bar{\theta}/2}{\bar{c}}\,(v_h + \theta_1) - c_1 > 0 \quad \Longleftrightarrow \quad c_1 < 0.8\,\frac{v_h + \bar{\theta}/2}{\bar{c}}\,(v_h + \theta_1),$$

which establishes the stated cutoff. ■

Comparing this to the original sequential threshold $\hat{c}_1^{\,\text{SEQUENTIAL}}(\theta_1) = v_h + \theta_1 - \frac{v_h(v_h + \bar{\theta}/2)}{\bar{c}}$, one sees at once that

$$\hat{c}_1^{\,\text{COLLABORATIVE}}(\theta_1) > \hat{c}_1^{\,\text{SEQUENTIAL}}(\theta_1) \quad \forall\, \theta_1 > 0,$$

Simply put, when players must coordinate to obtain the joint payoff, each individual's prosocial

motive—the desire to help their partner—activates the instant their lie is pivotal. By contrast, in a sequential-move game, the first mover can free-ride on the partner's misreport and thus reduce the incentive to misreport. Requiring both players to misreport eliminates the free-ride option and transforms each potential lie into a joint undertaking: prosocial agents perceive deception not as a personal violation but as a cooperative investment with assured returns. Consequently, additional cost-types choose to misreport—and thus aggregate misreporting rates increase under collaboration. This theoretical result accounts for laboratory findings that collaborative misreporting experiments often document high rates of dishonesty: coordination effectively serves as the moral sanction.

Appendix C Further Results

Figure C.1. Time P_1 spent reporting whether their colors match.

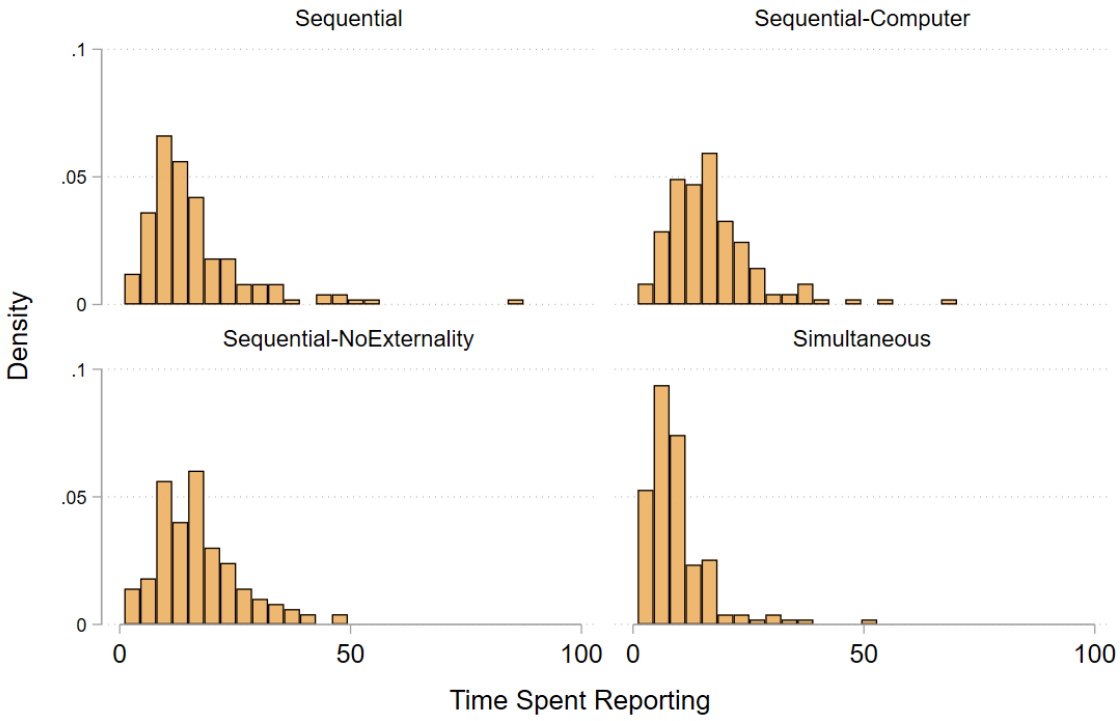


Table C.1. Participants with the role of P_1 in the Mind Game

	Session 1	Session 2	Session 3	Session 4	**Total**
Sequential	29	38	38	37	142
Sequential-Computer	29	36	39	37	141
Sequential-NoExternality	28	37	39	36	140
Simultaneous	30	38	40	38	146

Table C.2. Treatment comparisons using Fisher Exact Tests

	Sequential-Computer	**Sequential-NoExternality**	**Simultaneous**
Sequential	$p = 0.523$	-	$p = 0.064$
Sequential-Computer	-	$p = 0.024$	-

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

Table C.3. Regressions testing the differences in Beliefs across treatments

	Beliefs 1	Beliefs 2
Sequential	*Reference*	3.711
		(4.042)
Sequential-Computer	-3.719	*Reference*
	(4.025)	
Simultaneous	7.895**	
	(3.691)	
Sequential-NoExternality		-3.379
		(3.721)
P_1's report=1	6.397	2.297
	(4.090)	(4.102)
Sequential × P_1's report=1		4.129
		(5.791)
Sequential-Computer × P_1's report=1	-4.901	
	(5.785)	
Simultaneous × P_1's report=1	-2.963	
	(6.192)	
Sequential-NoExternality × P_1's report=1		-0.439
		(5.834)
Constant	49.835***	55.810***
	(5.220)	(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 D Instructions of the Mind Game

I present here the instructions used in the Mind Game. I provide the complete instructions for the treatment SEQUENTIAL 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.

D.1 SEQUENTIAL

[*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!

D.2 SEQUENTIAL-COMPUTER

[*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.

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	£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!

D.3 SEQUENTIAL-NOEXTERNALITY

[*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!

D.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!