Promise-Keeping and the Internal Judge*

Sorravich Kingsuwankul† Chloe Tergiman‡ Marie Claire Villeval§

March 17, 2026

Abstract

While standard models recognize intrinsic costs of lying, they typically focus on individual decision-making, treating reputational concerns as a trade off between the benefits of lying and the probability of detection by a passive observer. Using a strategic communication game in a controlled experiment, we show that honesty oaths short-circuit this calculus by internalizing the audience, even when the audience is a strategic counterpart whose beliefs and actions determine payoffs. While the oath dramatically increases truth-telling, those who do break their promise systematically avoid brazen, detectable lies, retreating instead to ambiguity. Crucially, this refusal to be a "brazen renegade" is not a strategic reaction to the receiver: it persists even when the oath is private and compliance cannot be traced to the participant by the experimenter. This inelasticity to external scrutiny challenges standard reputational models: instead, our data are consistent with the oath-taker answering to an internal judge rather than an external one. Furthermore, we show that this internal audience does not strictly

*Acknowledgements: We are grateful to N. Ali, B. Corgnet, L. Gangadharan, F. Gul, N. Jacquemet, R. Kerschbamer, C. Raymond, H. Schneider, R. Shorrer, R. Siegel, V. Valero, A. Viriyavipart, T. Woutersen, seminar attendees at the University of Arizona, and participants at presentations at the Applied Ethic and Morality Research Group at the University of Prague, the Credence Goods and Expert Markets Webinar, the Innsbruck Winter summit on (Un)Ethical Behavior in Markets, the IAER Microeconomics Workshop at Dongbei University, the PIER Research Workshop, Cornell University, Michigan Ross TO, Penn State University, Stanford University, Queen's University, the University of Arizona, the Ohio State University, the East Coast Conference, ASFEE in Montpellier, MBEES-MBEPS in Maastricht, the ESA World meeting in Lyon, the TIBER symposium in Tilburg, and the European Economic Association meeting in Barcelona for useful comments. We are thankful to Q. Thevenet for his assistance in programming the experiment. This research has benefited from the financial support of IDEXLYON from Université de Lyon (INDEPTH-IDEX/SBP/2018/03) within the Programme Investissements d'Avenir (ANR-16-IDEX-0005) operated by the French National Research Agency, and of the LABEX CORTEX(ANR-11-LABX-0042) of Université Claude Bernard Lyon 1, within the program "Investissements d'Avenir" (2019-ANR-LABX-02) operated by the French National Research Agency (ANR). This study received ethical approval from GATE-Lab Review Board (# 2021-10).

†School of Business and Economics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Tinbergen Institute, 1082 MS Amsterdam, The Netherlands. Email: s.kingsuwankul@vu.nl

‡**Corresponding Author.** Pennsylvania State University, Smeal College of Business, 334 Business Building, University Park, PA 16802, United States. Email: cjt16@psu.edu

§CNRS, Université Lumière Lyon 2, Université Jean-Monnet Saint-Etienne, emlyon business school, GATE, 35 rue Raulin, F-69007, Lyon, France. Email: villeval@gate.cnrs.fr

require the cognitive amnesia or imperfect recall required by standard self-signaling models; rather, it can also be understood as a present-moment, categorical refusal to generate inescapable evidence of one's own transgression. Finally, we show that receivers intuit this mechanism, pricing in the oath's "psychological enforceability" by granting credibility only when the speaker has no room to hide.

Keywords: Promise-keeping, Honesty Oaths, Deniability, Lying, Image Concerns, Internalized Audience, Laboratory Experiment

JEL codes: C91, D83, D91

1 Introduction

Oaths are an ancient and ubiquitous feature of human relations. Witnesses swear to tell the truth in court; physicians, financial professionals, and public officials pledge to uphold ethical standards. Although these oaths differ in content, they share a common institutional logic: they are invoked precisely where behavior cannot be fully specified or continuously monitored. Even when formal sanctions exist, they often apply only to violations that can be conclusively demonstrated. However, under asymmetric information, dishonesty is often deniable: the oath-taker can attribute a potential violation to uncertainty or bad luck, thereby shielding themselves from sanction. In such environments—where oath-breaking is difficult to verify—compliance ultimately depends on truthful representation of facts, intentions, or adherence to professional norms. This poses a puzzle: if dishonesty is difficult to detect and often plausibly deniable, why do oaths appear to change behavior at all?

An oath does not enhance monitoring or increase the probability of detection. Its force lies elsewhere: in the creation of an internal moral commitment—a "psychological contract" that raises the internal cost of dishonesty. But who is the relevant judge of compliance? Our goal in this paper is to understand how honesty oaths work and, more specifically, whose judgment—that of external observers or the oath-takers themselves—ultimately disciplines behavior.

In this paper, we focus specifically on an honesty oath that explicitly binds the agent to truthfulness, and utilize the "Announcement Game" (Tergiman and Villeval, 2023) to isolate the mechanics of promise-keeping. In this strategic framework, an Advisor privately observes the quality of an asset—represented by three cards that can each independently hold a "star" (success) or be blank (failure). The Advisor sends a message about the total number of stars (0, 1, 2, or 3) to an Investor, who decides whether to invest. If the investment is made, the actual return is determined by a random draw of just one of the Advisor's three cards. Because the Advisor's payoff relies entirely on the Investor choosing to invest , while the Investor's returns depend on a successful draw , Advisors face a material incentive to exaggerate when their true number of stars is low.

This setting is ideal for studying oaths because it distinguishes between two types of lying: detectable lies (e.g., announcing 3 stars when the truth is 1, where a failure exposes the lie) and deniable lies (e.g., announcing 2 stars when the truth is 1, where a failure can be attributed to bad luck). This distinction allows us to use the chosen type of lie as a behavioral "tracer," revealing exactly which audience the oath-taker is attempting to manage: an external observer or themselves.

We employ six between-subject treatments with 757 participants to separate the layers

of image concerns, employing a stranger matching protocol to mitigate instrumental reputational effects or direct punishment. In all treatments, it is common knowledge that Advisors are explicitly instructed to report the number of stars they observe. In our experimental design, we systematically manipulate the visibility of the oath to isolate the effects of three distinct audiences. First, we consider the internal audience (the self, or the "internal judge," in our game the Advisor). Second, we decompose external observers into two distinct categories: a passive audience (the experimenter, who observes behavior but takes no action in the game) and an active audience (the Investor, who is a direct interaction partner).

- **No-Oath** (the baseline): This treatment establishes behavior in the absence of concerns due to an oath.

- **Public-Oath and Private-Oath treatments**: These treatments introduce possible fixed costs of breaking an oath as well as active audience, passive audience, and internal audience image concerns. We manipulate the visibility of a compulsory oath. In the Public-Oath treatment, the Advisor's oath (*"I swear upon my honor that during this experiment I will behave honestly and I will always tell the truth."*) is common knowledge so that the Investor is aware the pledge has been made. In the Private-Oath treatment, the Advisor swears the oath, but the Investor is unaware of it. Comparing these allows us to measure the impact of the active audience.

- **Private-Oath^{++} treatment:** The Private-Oath^{++} treatment removes the link between decisions and identity, making the experimenter unable to identify whether a given person is an oath-breaker or not. Comparing Advisor behavior in this treatment and in the Private-Oath treatment allows us to measure the extent to which Advisors aim to escape the scrutiny of the passive audience (the experimenter).

- **Necessary and Sufficient Conditions treatments**: These treatments aim to understand the impact of the oath per se on Investors. In these treatments, by pairing investors with past Advisors from the Public-Oath treatment, we exogenously vary whether Investors are aware of the oath and the level of honesty they are facing. Furthermore, by filtering the pool of past Advisors to either retain their natural high honesty or to artificially mimic baseline untruthfulness, we also vary the level of honesty the Investors face. This allows us to test if knowledge of the oath is either required for trust to form or enough to generate trust on its own.

Prominent theoretical frameworks (e.g., Dufwenberg and Dufwenberg (2018); Abeler et al. (2019); Khalmetski and Sliwka (2019); Gneezy et al. (2018)) typically model lying costs as a combination of intrinsic preferences (a fixed cost or a cost related to the size of the lie)

and social-image/reputational concerns that depend on the inferences an external observer could draw. Our data reveal that oaths work by fundamentally shifting the locus of this observer, challenging standard intuitions about the role of audiences: the oath does not simply increase the cost of lying; rather, it collapses the mechanics of external scrutiny into an internal self-image constraint that compels the agent to answer to an "internal judge."

In our experiment, although Advisors are explicitly instructed to report the number of stars they observe, the high level of misreporting (90%) in the No-Oath treatment demonstrates that mere instructions are insufficient to curb material incentives. The introduction of the oath fundamentally recharacterizes the act of misreporting, and our data provide evidence that this recharacterization operates through a dual mechanism that governs the extensive and intensive margins differently. We formalize this through a bipartite cost structure:

$$Cost = \alpha + \beta\{Detectability > 0\}.$$

First, the presence of the oath introduces a fixed moral cost α of breaking the pledge, which is incurred whenever an individual decides to misreport at all. This cost α represents the psychological barrier to breaking the oath and is contextually sensitive; it is amplified when the oath is public ($\alpha_{private} > \alpha_{public}$), reflecting the heightened solemnity of an interpersonal promise, and shifs behavior on the extensive margin (the decision on *whether* to lie). Second, the oath introduces an internal self-image constraint β on brazenness that governs the intensive margin (the decision on *how* to lie). Crucially, while α scales with external scrutiny, external scrutiny leaves the intensive margin, and thus β, untouched. Indeed, whether the oath is public or private, and regardless of whether the experimenter can link decisions to identities, the behavior of those who choose to lie is largely stable and reflects a desire to avoid actions that would lead to a categorical self-verdict—seeing oneself as a renegade. It is this "internal plausible deniability" constraint that forces agents to audit their own integrity as strictly as an outside observer would, and dictates the nature of violations—a mechanism consistent with the spirit of the conjecture of Dufwenberg and Dufwenberg (2018).

At first glance, the concept of "internal plausible deniability" may seem a contradiction in terms: how can one hide a truth from the very mind that constructed the lie? We argue that maintaining internal plausible deniability does not have to create uncertainty about the violation itself, but rather constrains the narrative the agent can tolerate. To illustrate this psychological maneuver, consider a driver who grazes a parked car and drives away without looking back at the bumper. They know they have caused a collision, even in the absence of

any external witnesses. But by refusing to look, they avoid confronting the definitive visual proof that would force them to categorize themselves as a destructive driver. Similarly, we find that the oath induces agents to avoid actions that could lead to definitive evidence of their own transgressions, preferring ambiguity that protects their self-concept from an unambiguous moral verdict: if a simple lie is a transgression of a rule, a broken oath is a transgression of the self.

Turning to the Investors, we find that they appear to intuit this internal constraint. To assess the impact of the *knowledge* of the oath on Investors, we look to first period investment rates. We find that Investors generally adopt a "benefit of the doubt" approach for deniable claims even in the absence of the oath. The oath instead functions mainly by granting immediate credibility to maximal claims where the oath leaves no room to hide. In doing so, the oath protects high-stakes announcements from the skepticism they otherwise face in the absence of a pledge. Finally, the Necessary and Sufficient Condition treatments establish the primacy of experience: mere knowledge of the oath is neither necessary nor sufficient to sustain trust in the long run; trust must be validated by honest behavior.

2 Related literature

This paper contributes to several strands of the literature studying lying and image concerns, promise-keeping, and oaths.

Lying and image concerns. The literature on lying has increasingly recognized that individuals are motivated not only by material payoffs but also by intrinsic costs of lying and reputational concerns (e.g., Gneezy et al., 2018; Khalmetski and Sliwka, 2019; Schudy et al., 2024), often modeled as increasing with the magnitude of the lie or the probability of external detection. In their meta-analysis of the Fischbacher and Föllmi-Heusi (2013) rolling-the-die paradigm, Abeler et al. (2019) argue that empirical patterns of truth-telling are most consistent with an intrinsic preference for being honest and a preference for being seen as honest by the experimenter. Theoretical models formalize these reputational concerns in different ways. For instance, the perceived cheating aversion framework by Dufwenberg and Dufwenberg (2018) posits that decision-makers experience disutility proportional to the extent an audience infers the extent of cheating. Alternatively, Fries (2024) formalizes a "character-based" image concern, where agents care about signaling their intrinsic moral type to an external audience (an honor-stigma effect).[1]

[1]For a theoretical treatment of how lying costs operate when the audience is a strategic counterpart rather than a passive experimenter, see Kartik (2009). Unlike reputational models, Kartik (2009) focuses on functional or exogenous

While these models focus on how individuals engaged in an individual decision-making task à la Fischbacher and Föllmi-Heusi (2013) manage the inferences of an external passive observer, Dufwenberg and Dufwenberg (2018) also raise the critical conjecture that the relevant audience might actually be the decision-maker's own self. The concept of an internal audience has so far received limited empirical evidence, though some studies hint at this possibility. Notably, Fries et al. (2021) found that implementing double-blind procedures (thereby removing the external audience) failed to increase maximum cheating, a pattern consistent with the relevant audience being internal (see also Fischbacher and Föllmi-Heusi, 2013). Unfortunately, as they acknowledge, severe ceiling effects in their baseline treatments produce a confound and prevent a clean empirical test of this internalized mechanism. Related evidence from advice settings suggests that individuals engage in self-deception to preserve an ethical self-image, but this ability is constrained when they are forced to form an unbiased evaluation before learning about their material incentives (Gneezy et al., 2020).

A distinct route to internal discipline appears in the identity and self-signaling literature. In these settings, individuals play an intra-personal game with an "acting self" who wishes to protect a valued self-image by managing the information their "observing self" later uses to evaluate their type. As Bénabou and Henkel (2025) emphasize in their review of self-signaling models (see also Akerlof and Kranton, 2000; Bénabou and Tirole, 2006, 2011; Bénabou et al., 2024), individuals manage their beliefs to preserve their self-concept. Standard formalizations of this self-signaling require the agent to partially forget their exact motives or situational context in order to later draw Bayesian inferences about their own moral type from their past actions.

By employing the Announcement Game to distinguish between detectable and deniable lies across varying levels of oath visibility, we show that oaths fundamentally shift the locus of reputational concerns from the external observer to the self, even in the presence of an active audience in a strategic setting. Under oath, lying costs no longer scale with the size of the lie. Instead, the oath introduces a *categorical* constraint: individuals either comply fully, or, if they break the oath, strictly avoid detectable violations. Importantly, this avoidance of brazen lies persists even when violations are completely untraceable to both the Investor and the experimenter, ruling out explanations based on strategic concealment. Ultimately, this induces agents to maintain what we term "internal plausible deniability." Our concept and findings are consistent with standard self-signaling frameworks: agents may avoid brazen lies specifically to prevent generating inescapable evidence that would later pierce their cognitive amnesia. However, our concept does not strictly require imperfect recall. Rather,

costs of misrepresentation (e.g., manipulation effort or legal penalties). In this strategic setting, because the receiver rationally deflates messages to account for expected bias, *all types of senders* are driven toward equilibrium "language inflation"—paying a cost to claim a higher type simply to ensure their true information is correctly interpreted.

the internal judge functions just as powerfully as a present-moment, categorical refusal to confront definitive proof of one's own transgression. Under an oath, Advisors engage in deniable lies not to fool external observers, nor necessarily because of imperfect recall, but to actively protect their immediate self-concept.

Promise-keeping. A large experimental literature studies promise-keeping in strategic environments, typically in trust games or social dilemmas where one player promises a future action to another (e.g., Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006; Ismayilov and Potters, 2016; Ederer and Stremitzer, 2017; Casella et al., 2018; Di Bartolomeo et al., 2019; Sengupta and Vanberg, 2023; Grubiak, 2025). In these settings, promises are often conditional on the receiver's prior cooperation or trust, and deviations may generate guilt or disappointment costs driven by the receiver's reliance or by the harm imposed on the receiver.

Our setting differs in two fundamental ways. First, the moral commitment we study is unconditional: our senders swear to report truthfully regardless of the receiver's actions. Second, the commitment concerns the truthful transmission of exogenous information rather than a costly future action. These features allow us to abstract from reciprocity, reliance, and strategic belief manipulation, and instead focus on the structure of moral costs themselves. While Vanberg (2008) shows that promise-keeping cannot be fully explained by guilt aversion, our design goes further by decomposing how moral costs related to promise-keeping operate.

Recent work has explored additional mechanisms in promise-keeping, including internal consistency (Ismayilov and Potters, 2016) and transparency of commitments (Lang and Schudy, 2023). However, these studies generally do not distinguish whether moral costs are fundamentally social or internal.

Oath-taking. Finally, we contribute to the literature on oath-taking (see Table A1 in Online Appendix A for an overview, see also Zickfeld et al., 2025). While prior studies confirm that oaths reduce lying in strategic settings (e.g., Beck, 2021; Jacquemet et al., 2024), they do not isolate the specific moral costs driving this behavior. Our objective goes beyond documenting that oaths work to understanding why they work. Our design allows us to decompose the image concerns that drive compliance, clarifying when and for whom oaths are effective.

3 Experimental design and procedures

We first describe the general structure of the session, the treatment variations, then the procedure.

3.1 General structure

We use the Announcement Game of Tergiman and Villeval (2023), where participants are assigned either the role of Advisor or Investor. In each period, Advisor and Investor start with an endowment of 30 and 100 tokens, respectively. At the beginning of the period, a random draw determines a set of three cards for the Advisor, each of which has an independent 0.5 probability of displaying a star, which indicates a successful project. This probability is known to both the Advisor and the Investor, but the number of stars (0, 1, 2, or 3) is private information to Advisor. After observing the number of stars, the Advisor sends a message to Investor. Importantly, in all treatments, participants are explicitly instructed to report the number of stars they observe. However, the message is cheap talk, and the Advisor can misreport. The Investor then decides whether or not to invest his endowment with the Advisor. Next, Nature selects one of the Advisor's three cards to determine whether the investment is a success (if the drawn card displays a star) or a failure (if it is blank).

Irrespective of the Investor's decision, both the Advisor and the Investor learn about the outcome of Nature's draw (but we never inform Investors about the content of the three cards). The payoff of the Advisor depends on whether the Investor invested or not: the Advisor earns 230 tokens if the Investor invested, 30 otherwise. The payoff of the Investor depends on the decision to invest and the outcome of Nature's draw. If the Investor decides not to invest, he keeps his 100 tokens. If he decides to invest and a star is drawn, the investment is successful, and he earns 300 tokens (100-100+300). If instead a blank card is drawn, the investment fails, and he earns 30 tokens (100-100+30). Payoffs are realized, and the game ends.

In this game, truth-telling cannot be an equilibrium outcome. Indeed, the payoffs are such that in the absence of information, a risk-neutral Investor would want to invest. However, the Investor would never want to invest if he knew that the Advisor's portfolio has no stars in it. Given that the Advisor always prefers the Investor to invest, the Advisor strictly prefers to lie and conceal the fact that there are no stars. In this cheap talk game, the equilibria are all payoff-equivalent to the babbling equilibrium where the Advisor sends an uninformative message and Investor always invests.

In the laboratory, roles were fixed for the whole session and neutrally framed in the instructions as Player A (Advisor) or B (Investor). Participants played this game 18 times

with random and anonymous rematching between periods.[2] At the end of the session, the program randomly selected one period to count for payment for the Announcement Game.

The Announcement Game took place in the second part of the experiment. In part 1, that is prior to receiving the instructions for the Announcement Game, participants played eight practice periods of the "Truthful Announcement Game" where they all held the role of Investor, and where each participant was paired with a truth-telling computer. These periods aimed to help participants familiarize themselves with the setting and encounter all possible cases of the announcements corresponding to the probabilities used in the Announcement Game. After the eight practice periods, participants made five incentivized decisions: in period 9, participants decided whether or not to invest in each of the four possible scenarios (*i.e.,* 0, 1, 2, 3 stars), and in period 10, they made an investment decision without being informed of the number of stars. Investment decisions in periods 9 and 10 allow us to establish a benchmark of risk attitudes to study Investors' behavior in the Announcement Game. Either period 9 or 10 was randomly selected to count for payment for the Truthful Announcement Game, without feedback until the end of the session.

After completing the Announcement Game, all participants answered incentivized questions about their beliefs depending on their role. The Investors had to estimate the proportion of truthful announcements for each number of stars announced to them during the session (*i.e.,* their first-order beliefs about the Advisors' behavior). The Advisors were asked to estimate both the Investors' guesses about the proportion of truthful announcements for each number of stars (*i.e.,* their second-order belief) and the proportion of truthful announcements for each number of stars made by all Advisors in the session (*i.e.,* their first-order belief). They earned a 0.50 Euro bonus if their answer in one (Investors) or two (Advisors) randomly selected guesses was accurate within an interval of 5 percentage points.[3]

Finally, participants were told their earnings in each part and answered a questionnaire including socio-demographic information and questions on the motivation of their choices in the experiment.

3.2 Treatments

Below, we describe our six between-subject treatments. The full set of translated instructions is available in Online Appendix B.

[2]To avoid any end-game effects, participants were only informed that they would play a minimum of 10 periods and a maximum of 30 periods, though the number of periods, 18, was pre-determined before the experiment.

[3]Belief data as well as additional information on end-of-experiment beliefs are discussed in Online Appendix E.

No-Oath treatment: In the No-Oath treatment, no oaths were taken by participants in the role of Advisor. However, in this baseline —as in all our treatments— the instructions explicitly directed participants to "announce [their] total number of stars" to the Investor. Thus, lying in the baseline constitutes a violation of the task instructions, even in the absence of a sworn oath.

Public-Oath treatment: In the Public-Oath treatment, participants in the role of Advisor were required to swear the following oath: "I swear upon my honor that during this experiment I will behave honestly and I will always tell the truth."[4] The instructions for all participants stated that those who would be assigned the role of Advisor would be asked to take an oath by which they would swear upon their honor that during the experiment, they would behave honestly and always tell the truth. The oath-taking stage took place after participants discovered their role in the game. The Advisors swore the oath by typing in the text of the oath on their computers to facilitate its encoding in memory (see Naka and Naoi, 1995; Skinner et al., 1997).[5] The text for the oath and taking of the oath was common knowledge.

Private-Oath treatment: In the Private-Oath treatment, the instructions handed out to all participants were identical to those of the No-Oath treatment. However, after discovering their role in the Announcement Game, all the participants in the role of Advisors were asked to take the oath, while those in the role of Investors, who were not informed of the existence of the oath, were asked to type a neutral sentence during the oath-taking stage.[6] The Advisors were made aware that all participants in their role had to take the oath and were also informed of the asymmetry in information with respect to the Investors.

The Private-Oath treatment serves a dual purpose in our identification strategy. First, we compare behavior in the Private-Oath treatment with that in the No-Oath treatment to identify the importance of the self-image and passive audience-image concerns of breaking an oath: since the instructions for the Investors were identical across these two treatments, the impact on the reputation of the Advisors on the Investors as it relates to breaking the oath

[4] As shown in Table A1 in Online Appendix A, while most previous experiments utilize voluntary oaths, refusal rates in these studies are consistently low. We mandated the oath to better align with real-world professional contexts, where employees are typically required to sign pledges upon hiring. This mandate also ensured that the act of oath-taking was common knowledge. Participants remained free to opt out of the experiment entirely rather than sign the oath; none chose to do so.

[5] The computer program checked that the words "I swear upon my honor", "honestly", and "truth" were correctly typed. We also checked the actual oaths typed by participants after the experiment. All were correctly typed, except for minor typos.

[6] This was done to avoid placing the Investors in a situation in which they would wonder why other participants were typing while they were not.

is the same (and nil) in both those treatments, so that differences in behavior across these treatments can be assigned to the joint private and passive audience image concerns. Second, comparing the Private-Oath treatment to the Public-Oath treatment allows us to isolate the effect of the active audience. Because the only difference between these two conditions is the Investor's awareness of the pledge, this comparison reveals how the scrutiny and expectations of a direct counterpart influence the Advisor's decision to break the oath.

Private-Oath^{++} treatment: This treatment was identical to the Private-Oath treatment except for the payment procedure, which was modified so that participants knew ex-ante that the experimenter could not link any violations of the oath to their identity. In the Private-Oath^{++} treatment, participants were informed at the beginning of the session that at the end of this session, they would receive an envelope containing their payment in cash at their computer terminal. They learned that these envelopes (identified by the computer numbers) would be distributed by an independent person who would not be informed of the content of the experiment and would not have access to the experimental data. To preserve anonymity while maintaining accounting compliance, participants were informed that they would be asked to sign a "collective receipt," that is, a document including a table listing all payment amounts but without any corresponding computer numbers. Each participant would only be required to confirm that their payment amount appeared on the list without having to indicate which one it was. The experimenters did not participate in the payment procedure in any way, nor were they present in the room once the experiment was launched. Participants knew they would not interact with the experimenter and that it was impossible to link a payment or action to an identifiable person. This procedure removes identifiability, rendering actions anonymous to the experimenter.

Necessary Condition treatment: All participants in this treatment were assigned the role of Investors. They were told that the instructions they would receive would be an *excerpt* of the instructions that were given to participants from a past session. Those instructions were those from the Public-Oath treatment and were identical to them, except for the removal of the section on the oath.[7] After reading those instructions, participants learned that in each period, they were going to be randomly matched with Advisors from those past sessions and would face the announcement of the Advisor they would be matched with. In addition, their payment would be determined by Nature's draw for that particular Advisor in that

[7]In the Public-Oath treatment, the reference to the honesty oath was placed at the very end of the instructions, and we simply removed that section in the Necessary Condition treatment. In our instructions, we were very explicit that the instructions they were to receive were an "*excerpt*" of the prior ones. On the reasons for not considering withholding information about treatment manipulations as deception see Hey (1998); Hertwig and Ortmann (2008); Charness et al. (2022).

period.[8,9] To ensure that participants had enough opportunity to learn from experience, we extended the number of periods in this treatment to 27; as in the other treatments, this number was unknown to the participants. Comparing investment decisions in this treatment to those in the Public-Oath treatment allows us to study whether the Investors react to the fact that they know that an oath was sworn, or to their actual experience of the Advisors' truthfulness.

Sufficient Condition treatment: All participants in this treatment were assigned the role of Investor. The procedure of the Sufficient Condition treatment was similar to that of the Necessary Condition treatment, except that in part 2, the instructions explicitly integrated those used in the Public-Oath treatment, including the part about the oath. Participants were then told that in each period they would be randomly matched with an Advisor drawn from "*a subset of A players*" from these past sessions. Unbeknownst to the participants, this subset of Advisors was chosen such that the empirical distribution of announcements was as close as possible to that in the No-Oath treatment. In other words, we created a situation where the Investors in the Sufficient Condition treatment faced the Advisors who took a public oath but behaved on average like those who did not take any oath. Comparing investment decisions in this treatment to those in the Public-Oath and No-Oath treatments allows us to examine whether the knowledge about the existence of the oath in itself is sufficient to persuade the Investors to invest and whether this changes over time as they experience untruthful announcements. Table 1 summarizes the experimental design.

3.3 Procedures

The experiment was run at GATE-Lab, Lyon, France. We ran a total of 36 sessions: 7 sessions with 152 participants in the No-Oath treatment (76 participants in each role), 7 sessions with 154 participants in the Public-Oath treatment (77 in each role), 7 sessions with 146 participants in the Private-Oath treatment (73 in each role), 7 sessions with 150

[8] In each period and for each Investor independently, the program randomly selected one of the 1386 announcements made by the 77 Advisors who participated in the Public-Oath treatment and showed the Investor that announcement. The card (star or blank) that determined the outcome of the investment in that period was the one actually drawn by Nature on that specific announcement in the past session. Thus, the Necessary Condition treatment preserved the same empirical distribution of the Advisors' announcements and Nature's draws as in the Public-Oath treatment. Moreover, due to a lack of time trends in the announcements by Advisors in the Public-Oath treatment, our randomization procedure does not create any period-related bias in this treatment. A Kolmogorov-Smirnov test reveals no significant difference between the actual distribution of announcements in Necessary Condition treatment (*ex-post*) and that in Public-Oath treatment ($p = 1.000$).

[9] Naturally, one other difference is that the Investors' decisions in this treatment impact the payoff of no one else. We acknowledge that participants may take others' payoffs into account. However, Tergiman and Villeval (2023) showed that with the parameters used, social preferences do not play a large role in the Announcement Game.

Table 1: Experimental Design and Identification Strategy

Treatment	Design Features			Image Concerns Activated			Identification Strategy	
	Oath Sworn	Exp. Identifiable	Known to Inv.	Self Image	Passive Audience	Active Audience	**Comparison**	**Audience Effect Identified**
Panel A: Advisor Incentives								
No-Oath (Baseline)	×	×	×	×	×	×	–	*Baseline (No Audience)*
Private-Oath++	✓	×	×	✓	×	×	vs. *Private-Oath*	**Passive Audience** (Experimenter)
Private-Oath	✓	✓	×	✓	✓	×	vs. *No-Oath*	**Self+Passive Audiences**
Public-Oath	✓	✓	✓	✓	✓	✓	vs. *Private-Oath*	**Active Audience** (Investor)
Panel B: Investor Trust Formation								
Necessary Cond.	(✓)[a]	×	✓	*n/a (Investor focus)*			vs. *Public-Oath* & *Private-Oath*	Is knowledge **necessary** for trust?
Sufficient Cond.	(✓)[b]	✓	✓	*n/a (Investor focus)*			vs. *Public-Oath* & *No-Oath*	Is knowledge **sufficient** for trust?

Notes: "Exp. Identifiable" refers to whether the experimenter could link decisions to identity (removed in Private-Oath++). "Active Audience" refers to the Investor, and "Passive Audience" is the experimenter.
[a] Advisors in this treatment were drawn from the *Public-Oath* history (truthful), but Investors were unaware of the oath.
[b] Advisors in this treatment were drawn from the *Public-Oath* history but filtered to mimic *No-Oath* behavior (untruthful); Investors were told an oath was signed.

participants in the Private-Oath^{++} treatment (75 in each role), 5 sessions with 77 participants in the Necessary Condition treatment (all Investors), and 3 sessions with 78 participants in the Sufficient Condition treatment (all Investors).[10] All 757 participants were recruited via HRoot (Bock et al., 2014). Participants were primarily students from local engineering and business schools.[11] The experiment was programmed using z-Tree (Fischbacher, 2007).

Upon arrival, participants randomly drew a ticket from an opaque bag, which assigned them to terminals. Instructions were distributed at the beginning of the relevant part and read aloud. The average duration of a session was 80 minutes.

In all treatments, the instructions were delivered via audio recording. After privately answering any questions (which were rare and never related to the oath), the experimenter left the room. The computer program was then launched by a technician. In all treatments except the Private-Oath^{++} treatment, participants were informed in advance that their earnings would be paid by bank transfer rather than in cash. Payment information was entered by participants into a separate file that was sent directly to the institute's accountant and was never shared with the researchers.

Participants' earnings consisted of their payoffs from each part of the experiment: the payoff in the relevant scenario in the randomly chosen period (9 or 10) in the Truthful Announcement Game, the payoff from the randomly chosen round in the Announcement

[10] We based the number of observations on an ex-ante power calculation. Assuming a Type-I error rate of 0.05 and a medium effect size (Cohen's d = 0.60), targeting 77 observations per role and treatment would allow us to uncover the hypothesized effect of oath using a two-tailed Mann-Whitney rank-sum test and achieve a power level of 0.95.

[11] See Tables C1 and C2 in Online Appendix C for summary statistics of the socio-demographic characteristics of Advisors and Investors.

Game, and the payment from the belief elicitation. On average, participants earned 20 Euros (SD = 6.2), including a 7-Euro show-up fee.

4 Results

We begin by evaluating how honesty oaths affect Advisor behavior in the Announcement Game. Our analysis exploits the specific structure of the Announcement Game to isolate the mechanisms behind promise-keeping. By distinguishing between *detectable lies* (which risk definitive exposure) and *deniable lies* (where bad outcomes can be attributed to bad luck), the game allows us to use the nature of the lie as a "tracer" to identify the specific audience the liar is attempting to deceive—the active audience (Investor), the passive audience (experimenter), or the internal audience (oneself).

To structure this analysis, we decompose dishonest behavior along two margins. The extensive margin captures whether an Advisor deviates from truthful reporting at all, while the intensive margin captures the magnitude and detectability of the deviation, conditional on a lie occurring. This distinction is crucial because different mechanisms predict different patterns across margins. For example, if oaths primarily raise the expected cost of external detection or social judgment, we would expect them to have a large impact on the intensive margin by pushing liars toward deniability. By contrast, if oaths operate by imposing a fixed internal cost of oath-breaking, they should primarily reduce the incidence of lying itself, while leaving the conditional structure of lying largely unchanged.

To analyze the data, and unless otherwise specified, we use two-tailed rank-sum tests as our main statistical tool to evaluate differences across treatments, and signed-rank tests to evaluate differences within treatments, using one observation per subject.[12] We further support our claims via the use of econometric analyses, using mainly random-effects linear probability models with clustering at the individual level, all placed in Online Appendix C.

4.1 Overall Effects of Public Oaths

Before turning to mechanisms, we first establish that the oath has a meaningful behavioral impact in this environment. The introduction of a public oath triggers a substantial shift in truth-telling. Recall that in the baseline, Advisors are already instructed to report the number of stars they observe; yet, without the oath, 90.8% fail to comply. In the absence of an oath, honesty is the exception; under a public oath, it becomes the norm. As shown in the top panel of Table 2, the fraction of Advisors who always tell the truth surges from

[12] We average choices within a subject and use this average choice as our unit of measure.

9.2% in the No-Oath treatment to 74.0% in the Public-Oath treatment ($p < 0.001$).

Table 2: Detectable and deniable lies in the No-Oath, Public-Oath, and Private-Oath treatments

Treatments	(1) No-Oath	(2) Public-Oath	(3) Private-Oath	(1-2)	(2-3) p-value	(1-3)
All Advisors						
% Advisors who always tell the truth	9.2%	74.0%	39.7%	$p < 0.001$	$p < 0.001$	$p < 0.001$
% Untrue announcements	52.3%	10.8%	27.7%	$p < 0.001$	$p < 0.001$	$p < 0.001$
Relative frequency						
Detectable lies	25.2%	3.9%	9.6%	$p < 0.001$	$p = 0.002$	$p < 0.001$
Deniable lies	57.4%	13.8%	37.9%	$p < 0.001$	$p < 0.001$	$p = 0.001$
Absolute frequency						
Detectable lies	71.1%	11.7%	32.9%	$p < 0.001$	$p = 0.002$	$p < 0.001$
Deniable lies	86.8%	23.4%	56.2%	$p < 0.001$	$p < 0.001$	$p < 0.001$
Among Advisors who lie at least once						
% Untrue announcements	57.6%	41.4%	46.0%	$p = 0.010$	$p = 0.397$	$p = 0.004$
Relative frequency						
Detectable lies	27.8%	15.0%	15.9%	$p = 0.007$	$p = 0.593$	$p = 0.003$
Deniable lies	63.2%	53.2%	62.8%	$p = 0.302$	$p = 0.316$	$p = 0.884$
Absolute frequency						
Detectable lies	78.3%	45.0%	54.5%	$p = 0.004$	$p = 0.479$	$p = 0.008$
Deniable lies	95.7%	90.0%	93.2%	$p = 0.334$	$p = 0.660$	$p = 0.568$

Notes: Relative frequency statistics show the average frequency with which participants engage in any particular type of lie. Absolute frequency statistics show the proportion of Advisors (A players) who make any particular type of lie at least once. The relative frequency of detectable lies corresponds to how often Advisors who saw fewer than three stars announced three stars. The relative frequency of deniable lies corresponds to how frequently Advisors who saw fewer than two stars reported two stars or announced one star while seeing none. This table presents data both for all Advisors (top panel) and for the subset who make at least one lie (lower panel).

Crucially, this effect is not limited to "risky" detectable lies. If the oath merely increased the fear of detection by Investors, we would expect a reduction only in detectable lies. Instead, the oath successfully suppresses *deniable lies*—those shielded from definitive exposure—suggesting that the oath serves as a psychological anchor rather than a mere calculation of detection risk. The relative frequency of deniable lies collapses from 57.4% to 13.8% ($p < 0.001$), while detectable lies virtually disappear, dropping from 25.2% to 3.9% ($p < 0.001$).

The Retreat of Oath-Breakers. Even among those who break their oath (26.0% in Public-Oath), the nature of dishonesty changes fundamentally. These subjects do not lie indiscriminately; rather, they appear to filter their dishonesty. Indeed, among those who lie at least once, the oath reduces the brazen, detectable lies that could brand them as definitive

oath-breakers, leaving the ambiguity of bad luck to hide behind. Specifically, among those liars, the frequency of detectable lies drops significantly compared to the baseline (27.8% vs. 15.0%, $p = 0.007$). Crucially, this reduction does not trigger a strategic retreat into ambiguity: the impact on deniable lies is rather muted, and not only does not rise to compensate, but instead rather trends downward (63.2% in the Baseline vs. 53.2% in the Public-Oath treatment, $p = 0.302$).

4.2 Decomposing the Mechanism: The Impact of the Passive and Self Audiences

Having established that the oath changes behavior on average, we next ask through which mechanisms this effect operates. A primary concern is whether the avoidance of lies is driven by the fear that the experimenter will identify the violator, or whether it is internally driven.

Isolating the Passive Audience Scrutiny. To evaluate this, we compare the Private-Oath to the Private-Oath^{++} (anonymous compliance) treatment. In the latter, the experimenter knows the oath was signed but cannot link compliance to any specific participant's identity. If the choice of making a deniable lie were merely a strategy to maintain plausibility in the eyes of the researcher, this constraint should vanish in the Private-Oath^{++} setting, and liars should return to profitable, brazen lies. They do not. Behavior along both margins is virtually identical across these two treatments: the fraction of truth-tellers (38.7% vs 39.7%), the frequency of detectable lies (10.4% vs 9.6%) and deniable lies (38.1% vs 37.9%) are under one percentage point apart across the two treatments and statistically indistinguishable ($p = 0.895$, $p = 0.629$, and $p = 0.896$. See also Table C7 in Appendix C).

This invariance is critical. It shows that the visibility of an oath violation to the experimenter plays no role in shaping how individuals choose to lie. In the Private-Oath++ treatment, compliance with the oath cannot be linked to the identity of any particular subject, ruling out strategic concerns about being identified as an oath-breaker by the experimenter. Yet behavior along the intensive margin is unchanged. Behavior, therefore, cannot be driven by a reputational calculation vis-à-vis the experimenter. Instead, it reflects an internal constraint: subjects act as though violating the oath remains costly even when no external observer can assign culpability.

Below, we formalize this intuition by showing how an internal fixed cost of oath-breaking, combined with an internal self-image cost, can account for the observed patterns when the Investor is unaware of the pledge.

The Fixed Cost. Since experimenter scrutiny does not affect Advisor behavior, we can now isolate the pure internal cost of breaking one's word. To do so, we compare the Private-Oath treatment to the No-Oath baseline. Crucially, the external environment in these two treatments is identical: the Investor is unaware of any oath. The sole difference is that Advisors in the Private-Oath treatment have signed a pledge. Even in this minimal setting, the oath significantly constrains behavior. The fraction of truth-tellers is four times higher in the Private-Oath treatment than in the baseline (39.7% vs. 9.2%, $p < 0.001$). This suggests the existence of a binding constraint for approximately 30% of participants—a constraint that operates independently of external judgment, and strong enough to block any and all possible deviations from the oath.

The Self-Audience Image Cost. This internal constraint shapes *how* participants lie, not just *whether* they lie. If the costs incurred for breaking an oath were merely a fixed penalty for breaking the oath, we would expect liars in the Private-Oath treatment to behave like liars in the No-Oath treatment (who frequently lie maximally). Instead, we observe a systematic retreat from brazenness.

Relative to the No-Oath treatment, oath-breakers in the Private-Oath treatment significantly reduce the frequency of detectable lies (from 27.8% to 15.9%, $p = 0.003$), while leaving the frequency of deniable lies essentially unchanged (63.2% vs. 62.8%, $p = 0.884$). This shift is particularly revealing because announcing 3 stars maximizes expected monetary payoffs: Investors in the Private-Oath treatment invest more following 3-star announcements (89.5% for 3-star announcements vs. 72.6% for 2-star announcements, $p < 0.001$, signed-rank test). Thus, conditional on violating the oath, these Advisors forego the most profitable deviations and instead concentrate their announcements in forms that preserve ambiguity, even though the Investors can never know of any violations of the oath (since they are unaware of it).

This pattern is further bolstered by the fact that even within the class of detectable lies, Advisors do not appear to condition their behavior on the degree of detectability: in the Private-Oath treatment, the probability of announcing 3 stars is similar whether the true state is 0, 1, or 2 stars.[13] The relative inelasticity to detection risk indicates that Advisors are unlikely to be engaging in a marginal calculus over exposure probabilities.

Taken together, these patterns along the intensive margin suggest that oath-breakers in the Private-Oath treatment behave as if breaking the oath triggers an additional self-image constraint that penalizes unequivocal violations but is insensitive to gradations in detectability. What matters is not how likely a lie is to be exposed, but whether it rules

[13]These values are 15.5%, 14.5%, and 18.5%, respectively. The smallest p-value for pairwise signed-rank tests is $p = 0.172$.

out ex-post reinterpretation by one's self. In this sense, the self-image costs of oath-breaking seem to be governed by a preference for "internal plausible deniability."

On its face, the notion of internal plausible deniability may appear paradoxical: Advisors surely know whether they have broken their oath. Our claim is not that the oath creates uncertainty about the factual status of one's action, but that it constrains the narrative through which that action can be integrated into one's self-image. The oath transforms lying from a mere deviation from the truth into a symbolic act of self-betrayal. In this setting, deniable lies retain a critical psychological property: they allow Advisors to maintain their self-concept by only taking actions that rule out confronting themselves as definitive oath-breakers.

This mechanism reflects real-world behavior in which people knowingly violate rules but preserve ambiguity to avoid a clear condemnation of themselves. Rather than denying the violation, they avoid looking at evidence that would force them to see it as blameworthy. For example, even in the absence of any external witness, a driver who feels a collision while reversing may leave without checking the damage. Not looking back does not undo the collision; it prevents confronting visible proof that would remove ambiguity about the harm caused. The ambiguity is not about whether a norm was violated but whether the self must fully acknowledge it. In this way, ambiguity serves to shield the individual's own self-concept rather than to obscure the reality of the transgression.

In our experiment, deniable lies play this role precisely. Even when no external audience can observe or sanction the violation of the oath, Advisors disproportionately avoid brazen misreports that would force a clean break between keeping and breaking one's promise. The oath thus induces agents to subject themselves to the same evidentiary standard they would anticipate from an external judge, internalizing the demand for plausibility and making the act of lying itself psychologically non-fungible.

These findings provide empirical support for the intriguing conjecture raised by Dufwenberg and Dufwenberg (2018) that an agent might act as their own audience, auditing their behavior much as an external observer would. In addition, our concept of internal plausible deniability is consistent with, but does not require, the cognitive amnesia present in the self-signaling literature (e.g., Bénabou and Tirole, 2006, 2011). Indeed, the constraint can operate categorically in the present moment by restricting the kind of evidence an Advisor is willing to generate. By avoiding detectable lies, the agent prevents the realization of unambiguous proof—such as drawing a blank card after claiming three stars—that would force an immediate, definitive self-classification as a renegade. Thus, the internal audience does not have to rely on probabilistic inferences drawn from forgotten actions, but rather on an active psychological refusal to confront inescapable proof of one's own betrayal.

The Public-Oath treatment provides a sharp test of the "internalized audience" hypothesis. If the preference for deniability reflects a fundamental self-image constraint, it should persist even when the active audience who directly engages with the oath-taker is present—a setting in which strategic and reputational considerations might otherwise be expected to dominate or reshape lying behavior.

4.3 The Impact of the Active Audience

We now turn to the Public-Oath treatment, where the Investor is aware of the pledge. This environment allows us to study how external social pressure coming from an audience that is active in the game (the Investors) interacts with the internal mechanism identified above. We find that while making the oath public dramatically increases compliance on the extensive margin, its effect on how remaining liars behave is quantitatively small and strikingly inelastic.

Active Audience as a Commitment Multiplier. The presence of the active audience has a dramatic effect on the extensive margin. The proportion of fully truthful Advisors nearly doubles in the Public-Oath treatment compared to the level observed in the Private-Oath treatment (74.0% vs. 39.7%, $p < 0.001$).

As we will show below, this happens without a retreat towards deniable lies, leading us to interpret this surge in truth-telling at the extensive margin as being the result of an increase in the fixed cost of lying. The active audience can scale up the fixed cost of breaking the promise, making the pledge solemn enough to make truth-telling the norm. This scaling can be interpreted as responding to the nature of the transgression: while breaking a private oath would be a failure of internal consistency, breaking a promise made to another would be experienced as an act of interpersonal betrayal. Even without the threat of any external sanctions, the psychological cost of this relational breach would then appear to compound the cost of a "mere" private violation.

Inelasticity at the Intensive Margin. Crucially, this scaling appears to occur without altering the strategic calculus of those who remain unfaithful to the oath. Just as removing experimenter scrutiny (in Private-Oath^{++}) failed to unleash brazen lying, adding investor scrutiny in Public-Oath fails to deter it: if the active audience acted as a monitor in the eyes of the Advisors, we would expect the remaining liars to retreat from detectable lies, reduce detectability within detectable lies, and increase deniable ones. We find no evidence of such shifts; instead, oath-breaker behavior appears remarkably inelastic to the introduction of the active audience (see lower panel of Table 2).

The high rate of compliance in the Public-Oath treatment naturally results in a modest number of oath-breakers, limiting the statistical power of tests on the intensive margin. However, our conclusion of inelasticity primarily rests on the comparatively small magnitude of the behavioral response relative to the large structural change in audience and risk of detection. The point estimates themselves suggest little scope for meaningful monitoring effects.

- **Rigidity of Detectable Lies:** Among oath-breakers, the constraint on brazen lying proves remarkably rigid, showing an insensitivity to the audience and the risk of detection. First, the overall proportion of detectable lies is largely stable across the Public and Private-Oath treatments (15.0% vs. 15.9%, $p = 0.593$). Second, this behavior is inelastic with respect to the risk level: while the probability of detection drops from 100% to 67% and 33% across states (0, 1, or 2 stars), the probability of making a detectable lie varies by no more than 4.5 percentage points (13.2%, 13.0%, and 17.3% respectively; $p > 0.563$). Third, the probabilities of announcing 3 stars given any particular state is stable across the Private and Public-Oath treatments (with a maximum difference of 2.3 percentage points and a minimum p-value for pairwise comparisons of $p = 0.322$). Overall, regarding detectable lies at the intensive margin, advisors seem to ignore both the active audience and the objective risk of detection.

- **Absence of Strategic Substitution (Deniable Lies):** We also find no evidence that liars strategically adapt to the audience by substituting brazen lies with safer deniable ones. If the active audience induced fear of detection, we would expect a significant rise in deniable lies. Instead, compared to the near-doubling of truth-tellers on the extensive margin, the prevalence of deniable lies among liars remains not only muted, but also tends in the incorrect direction (62.8% in Private-Oath vs. 53.2% in Public-Oath; $p = 0.316$). As a complementary within-state check, we also examine escalation within deniable lies, specifically, the probability of taking the profitable deniable action (announcing 2 stars) conditional on seeing 0, and again find no meaningful change across treatments (53.7% vs. 55.3% in the Private and Public-Oath treatments, respectively; $p = 0.862$).

This invariance highlights a fundamental distinction between oath-breaking and standard strategic lying. In standard models, lying behavior is elastic with respect to the audience: as the risk of detection increases, agents substitute brazen lies with safer, deniable ones. Instead, we observe a distinct stiffness on the intensive margin, as we detect no change in strategy when the active audience is watching. This suggests that the constraint on how

to lie responds to an internal rule rather than a marginal calculation of social risk: once crossed, the moral threshold imposed by the oath renders further adjustments to audience structure largely irrelevant.[14]

This inelasticity is even more striking when we consider the behavior of the Investors. As shown in Table 4, Investors in the Public-Oath treatment are significantly more willing to invest after high announcements (i.e., 97.7% vs. 85.8% for 3 and 2 stars, respectively; $p < 0.001$, signed-rank test). If oath-breakers were playing a strategic game against an external audience, one might expect them to respond to this increased credibility by altering their strategy to exploit it (e.g., by increasing the frequency of brazen and deniable lies to capture the higher expected returns). Furthermore, if this were a strategic calculation, we might expect to see learning dynamics over time as Advisors realize that brazen lies are highly profitable and largely unchallenged. However, the data show no such dynamics: the fraction of detectable lies remains completely stable over the course of the experiment (see Table C4 and Figure F1 in the Online Appendix). The fact that they do not corroborates that their primary constraint is internal: the oath-breaker is managing their own self-image, not the Investor's beliefs.

4.4 A Dual Mechanism of Compliance.

Synthesizing the large shift in truth-telling on the extensive margin—an almost doubling of the truth-telling rates from 37% to 74%—with the relative invariance of lying on the intensive margin across all Oath treatments, we find the data regarding the impact of the oath are most parsimoniously explained by decomposing the oath's impact into two distinct components.

First, the oath introduces a fixed moral cost of violation, α, which is paid whenever an individual chooses to misreport at all. This cost is independent of the magnitude or profitability of the lie. The level of this cost is, however, contextual: $\alpha_{private} < \alpha_{public}$. This explains the shift on the extensive margin: the surge in truth-telling when moving from the No-Oath setting to the Private Oath setting occurs because the fixed cost is introduced, and the further surge in truth-telling when the oath is public occurs because the fixed cost of breaking the oath is amplified. Conceptually, this fixed cost can be interpreted as reflecting the solemnity of the pledge: the more public the oath, the more solemn it feels, and the higher the psychological barrier to breaking it.

Second, the oath introduces an internal self-image constraint on brazenness, β. This governs the intensive margin, explaining why those who do lie avoid detectable violations

[14]These conclusions are supported by regression analyses (Tables C4 and C5), which are robust to alternative clustering specifications (Table C6; all in Appendix C).

regardless of any external audience. The stability of this avoidance across treatments supports that this is a self-image constraint. It is this "internal plausible deniability" constraint that dictates the nature of the remaining transgressions under oath.

This second mechanism is fundamentally categorical rather than continuous. Conditional on violating the oath, individuals face the additional psychological cost β if the violation is detectable. Importantly, this cost does not scale with the size of the lie or the probability of external detection. Instead, it reflects the loss of internal plausible deniability: detectable violations force a definitive self-classification as an oath-breaker, whereas deniable violations preserve ambiguity about one's moral identity.

Formally, we propose that the cost of breaking a promise be written as:

$$\alpha_{public,private} + \beta \cdot \mathbb{I}_{\{Detectability>0\}}$$

This bipartite structure can rationalize why the oath simultaneously deters many individuals from lying altogether while leaving the conditional pattern of dishonesty unchanged across public, private, and anonymous settings. Once the fixed cost is paid, behavior is disciplined not by fear of exposure, but by an internal evidentiary standard.

Alternative Cost Structures. While this proposed dual mechanism parsimoniously accounts for behavior on both the extensive and intensive margins, it is important to consider whether other standard cost structures could also rationalize the data. Before concluding our analysis of Advisor behavior, we briefly address several alternative modeling approaches.

Revolving Door Hypothesis.
One might argue that the relative inelasticity with respect to the audience among oath-breakers masks a "revolving door" effect and that external evaluative pressure could operate through offsetting selection at the internal margin. For example, increased observability might induce some Advisors who would otherwise make detectable misreports to switch to deniable lies, while simultaneously inducing others who would otherwise make deniable lies to switch to full honesty. In principle, appropriate combinations of these responses could preserve the observed distribution of violations. While arithmetically possible, this class of explanations would require a specific calibration of lying costs where the brazen types are sensitive enough to the oath to change how they lie, but not sensitive enough to stop lying, while deniable types exit at a rate that exactly preserves the baseline ratio. In addition, the persistence of distinct liar profiles (pure detectable, pure deniable, and mixed) across the Public and Private treatments—and their relative invariance to external scrutiny in the

Private-Oath^{++} condition—would require a fairly finely balanced set of countervailing shifts. Such accounts would generate limited independent comparative statics, as any observed invariance could be rationalized ex post by suitable choices of offsetting flows. By contrast, the internal plausible deniability mechanism we propose delivers the observed stability without requiring fine-tuning, predicting directly that detectability—not observability—governs behavior once the oath is violated.

Fixed Cost Only hypothesis.
This is easily rejected as if the only cost of breaking the oath were fixed, the remaining liars should shift to the most profitable announcement, i.e. announcing 3 stars. Instead, in our data, oath-breakers specifically retreat from brazen lies.

Size Dependent Costs hypothesis.
Under this hypothesis, the cost of breaking the oath varies with the distance between the announcement and the truth. The data also reject this possibility. We find that a cost structure based on the magnitude of the lie cannot explain the data, even when accounting for the higher expected returns of larger lies. For instance, while announcing 3 stars when observing 1 (a detectable lie) yields a higher expected payoff than announcing 2 stars when observing 0 (a deniable lie), participants choose the detectable lie significantly less often (14.5% vs. 53.7%, $p < 0.001$, signed-rank test), despite both representing a deviation of magnitude 2. This reversal of the expected payoff gradient indicates that the binding constraint is not a marginal trade-off between return and size, but is instead consistent with an aversion to "detectable" violations.

External Plausible Deniability hypothesis.
One might wonder if the data can be explained by an "augmented" model of lying in which the active audience introduces instead an additional *external* plausible deniability constraint, so that agents face a fixed cost of breaking the oath, and both an internal and external plausible deniability constraint. Our data are inconsistent with this modeling approach as well. Indeed, as explained above, if the active audience introduced a β_{public} parameter, then we should see a drop in detectable lies (which trigger the cost), and a shift towards deniable ones (that generate no cost). We observe no such shift: the lying profile of oath-breakers in the Public and Private-Oath treatments is relatively stable, and we observe a significant shift on the extensive margin.

More broadly, models that discipline dishonesty through marginal incentives—whether

reputational, probabilistic, or magnitude-based—would predict smooth substitution toward less detectable violations when scrutiny increases. The defining feature of our data is the relative absence of such substitution. Instead, behavior responds categorically, consistent with a model in which the oath transforms lying into an identity-relevant act that agents evaluate using an internal evidentiary standard rather than a marginal calculation.

4.5 The Impact of the Oath on Investors: Psychological Enforceability and the Primacy of Experience

While the oath dramatically constrains the Advisor, its reception by the Investor seems calculated rather than automatic. Our results are consistent with Investors assessing the pledge's "psychological enforceability." Investors appear to intuit that the oath is binding only when the Advisor is fully exposed as a renegade, i.e., when the Advisor cannot hide behind bad luck.

Psychological Enforceability: The Conditional Value of the Oath. We begin by assessing whether the mere knowledge of the oath leads to an immediate change in Investor behavior. To isolate this signaling effect from experience, we focus on first-period behavior. These first-period responses provide a clean window into how oaths affect credibility before learning or strategic adaptation can occur. To increase statistical power, we pool the data from treatments where investors are aware of the oath (Public-Oath and Sufficient Condition—the "Oath Aware" regime) and compare them to treatments where investors are unaware of it (No-Oath, Private-Oath, Private-Oath++ and Necessary Condition—the "Oath-Unaware" regime). Table 3 summarizes mean investment levels following different announcements across oath-knowledge regimes.

Table 3: 1st-Period investment rates across Oath-knowledge regimes for 2 and 3-star announcements

	Oath Unaware Regime ($n = 301$)	Oath Aware Regime ($n = 155$)	*p-value*
0 or 1 star announced	8.0% (100)	17.5% (40)	$p = 0.187$
2 stars announced	90.4% (135)	88.9% (72)	$p = 0.734$
3 stars announced	78.9% (66)	95.3% (43)	$p = 0.028$
$p - value$ (2 vs. 3 stars)	$p = 0.024$	$p = 0.234$	

Notes: The p-values are the results of ranksum tests. Number of observations in parentheses.

Although we do not know investors' true beliefs, the sharp contrast in the response to identical announcements across oath-knowledge regimes suggests that investors condition

on the perceived psychological cost of exaggeration. Indeed, in the Oath-Unaware regime, investment rates are not monotonic in announcement strength: they increase from lower announcements to 2 stars then decline following a 3-star announcement (90.4% vs. 78.9%; $p = 0.024$). This non-monotonicity suggests that, in the absence of an oath, Investors view extreme claims as cheap talk. They naturally discount these "too good to be true" announcements, anticipating that Advisors with poor draws will disproportionately inflate their reports to the maximum category.[15]

In contrast, in the Oath-Aware regime, first-period behavior changes markedly. Investments increase with announcement strength, and 3-star announcements elicit weakly higher investments than 2-star announcements (95.3% vs. 88.9%; $p = 0.234$). More importantly, and unlike in the oath-unaware regime, there is no evidence of reduced likelihood of investing following the strongest announcement. In short, knowledge of the oath leads to the disappearance of the non-monotonic response that occurs when no oath-taking is observed.

This contrast highlights the conditional value of the oath. When announcements are made under oath, Investors appear to treat the highest claims as more credible, as if they understand that, for Advisors, the oath imposes a psychological cost on dishonesty—specifically, the risk of being identified as an oath-breaker. Taken together, these results suggest that oaths do not simply raise average trust; they restore a monotonic relationship between announcement strength and investment. Without an oath, investors discount extreme claims that seem too good to be true; with an oath, stronger announcements carry greater persuasive force. This pattern suggests that psychological enforceability operates through Investors' beliefs about Advisors' moral or self-image costs when making strong claims. Thus, the oath does not function as a generic signal of character, but as a conditional constraint that listeners trust only when the speaker has no room to hide.

The Primacy of Experience. Having investigated the immediate impact of the knowledge of the oath *per se* on Investors, we investigate whether that knowledge is enough to sustain trust over time. For this we compare long-run behavior in the Sufficient Condition—where the oath is known but lying is highly prevalent—to the Public Oath and No-Oath treatments. We find that investment rates in the Sufficient Condition treatment not only fail to stay high, they essentially collapse: by the end of the game, behavior is in fact closer to the No-Oath baseline (see Online Appendix D). These results show that the oath does not function as a long-lasting ex-ante signal. While it may generate goodwill initially in the case of maximal announcements, it is not sustained after a string of obvious violations of the

[15]Empirically validating this skepticism, Advisors in the Oath-Unaware regime frequently default to the highest possible announcement: 17.7% of the time, compared to only 3.1% in the Oath-Aware regime ($p = 0.004$).

oath.

The Necessary Condition treatment demonstrates that knowledge of the oath is not necessary to build trust. Trust can be driven entirely by the verification of honest behavior. In this treatment, Investors faced largely truthful announcements made by Advisors from the Public-Oath treatment, but were unaware that an oath had been signed. Despite the lack of knowledge, by the end of the game and across all levels of announcements, Investors successfully calibrated their trust to the actual quality of the Advisors (see Online Appendix D). Indeed, they invested significantly more often in high announcements than in the Private-Oath treatment (where truth-telling was lower) and reached rates statistically identical to the Public-Oath treatment (where truth-telling was equivalent). The fact that Investors in the Necessary Condition reached the same high investment levels as the Public-Oath treatment—without knowing an oath was signed—demonstrates that trust is earned ex-post through interaction.

Table 4: Investment rates in the Announcement and Truthful Announcement Games

		No-Oath	Public-Oath
0 stars announced	% Announcement Game	9.5%[1] (21)	1.7% (59)
	% Truthful Ann. Game	0.0%	2.6%
1 star announced	% Announcement Game	14.5% (54)	16.8% (77)
	% Truthful Ann. Game	19.7%	16.9%
2 stars announced	% Announcement Game	70.5% (76)	85.8% (77)
	% Truthful Ann. Game	98.7%	98.7%
3 stars announced	% Announcement Game	77.8% (76)	97.7% (72)
	% Truthful Ann. Game	100.0%	98.7%
Average investment rate		65.8%	56.9%
Average earnings of B players		149.7	159.9

Notes: The table displays the fraction of times participants invested in the Announcement and Truthful Announcement Games. Numbers of participants faced with a particular announcement at least once are in parentheses. [1] While this percentage may seem relatively high, this represents a small number of participants, as only 21 B players saw a zero-star announcement in this treatment, and of these two invested.

Economic Consequences. Finally, we look at the economic consequences of a public oath. Table 4 shows investment rates for all periods in the No Oath and Public Oath treatments, both for particular announcements and on average. In the Public-Oath treatment, Investors experience largely truthful advisors, and we indeed observe that Investors treat

signals as largely informative.[16] For example, compared with No-Oath, the credibility of 2- and 3-star announcements is significantly increased. Indeed, the investment gap with the Truthful Announcement Game, where the three cards were perfectly revealed by the computer before the investment decision, narrows ($p < 0.001$ in both cases), triggering near-universal investment for the 3-star announcements (97.7% vs. 77.8%, $p < 0.001$). Truthful Advisors in the Public Oath treatment reveal their 0- and 1-star hands, leading Investors to completely avoid these projects. While this leads to missed opportunities in 1-star cases (which are profitable on average), the value of avoiding the disastrous 0-star "traps" and the increased investments after 2 and 3-star announcements are far greater. Thus, although overall investment volume drops (from 65.8% to 56.9%, $p = 0.005$), the quality of investment improves significantly, raising average earnings from 149.7 to 159.9 tokens ($p = 0.016$).

5 Conclusion

The central question in this paper is straightforward: to whom does the oath-taker answer? While standard economic models of lying typically view compliance as a calculation involving external judgment—trading off the benefits of a lie against the probability of detection by an audience—we find that oaths activate a fundamental psychological transformation. The primary contribution of this study is the finding that oaths effectively internalize the audience.

If oath-takers were primarily driven by social image concerns (what the audience thinks), we would expect their lying strategies to change when the active audience is removed or when the risk of detection by the passive audience vanishes. We find no evidence of this. Whether the oath is sworn publicly, privately, or even under conditions of increased anonymity with respect to the passive audience (the experimenter), the participants who choose to lie under oath exhibit a similar aversion to brazenness. This indicates that the oath-taker is less concerned with managing the beliefs of an external observer; they are managing their own self-image through internal plausible deniability. This internal constraint does not strictly require the cognitive amnesia required by standard models of self-signaling. Rather than forgetting their actions to make future probabilistic inferences about their own character, individuals can judge themselves in the present moment through the lens of a skeptical observer, actively refusing to generate undeniable evidence of their own transgression: the power of the oath appears to lie less in the fear of being caught by others, but more in the refusal to be caught by oneself.

[16] As a complement, Table C8 in the Online Appendix reports the estimates from a Linear Probability model of the likelihood of investing in high announcements.

This internalization mechanism forces a re-evaluation of the role of the public sphere. We find that the audience does not act primarily as a policeman monitoring compliance, but rather as a commitment multiplier. Swearing the oath publicly significantly increases the rate of full compliance compared to swearing it privately (the extensive margin), but it does not alter the nature of the lies told by those who break it (the intensive margin). This suggests that the presence of an active audience serves to heighten the solemnity of the pledge—raising the fixed cost of breaking the oath—but once that threshold is crossed, the constraint remains largely internal.

Finally, our results highlight a critical practical nuance: while the oath binds the speaker generally, it convinces the listener only conditionally. We find that knowledge of the oath *per se* is not a blanket signal of integrity; rather, investors appear to price in its "psychological enforceability." For deniable claims, where the Advisor can hide behind bad luck, knowledge of the oath does not impact Investor behavior, who generally give Advisors the benefit of the doubt. However, for the highest claims—which investors significantly penalize in the absence of an oath—knowledge of the pledge increases credibility. Thus, knowledge of the oath does not generate increased trust so much as it prevents the discount attached to "too-good-to-be-true" news, signaling credibility only when the Advisor's claim leaves no room to hide. In the long run, however, experience is the primary determinant of behavior: knowledge of the oath is neither necessary for trust to emerge nor sufficient to prevent its collapse when advisors are untruthful.

These findings offer a blueprint for institutional design. While we remain cautious of the generalizability of our findings, the results are relevant for the implementation of oaths in the real world. Since the mechanism is internal but scalable, the focus should be on visibility and maintenance. Publicizing the oath is essential not because it invites external scrutiny, but because it amplifies the internal psychological weight of reneging on the promise. Our survey with residents in the Netherlands (a country where all financial advisors take a compulsory oath) reveals that only 53.5% were either certain or thought it likely that such an oath exists, emphasizing a need for publicity (see Online Appendix G for full survey and related discussion). Furthermore, because this cost is tied to the salience of the pledge, institutions should implement reminders (which has been shown to be effective among Dutch advisors, see Weitzel and Kirchler (2023)) periodically to prevent the potential decay of this "internal audience" over time. Future research could explore how these preferences might erode in higher-stakes environments, or how these internal costs may vary across cultures, paralleling the global examination of honesty conducted by Cohn et al. (2019).

References

Abeler, J., D. Nosenzo, and C. Raymond (2019). Preferences for Truth-Telling. *Econometrica 87*(4), 1115–1153.

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *Quarterly Journal of Economics 115*(3), 715–753.

Beck, T. (2021). How the honesty oath works: Quick, intuitive truth telling under oath. *Journal of Behavioral and Experimental Economics 94*, 101728.

Bock, O., I. Baetge, and A. Nicklisch (2014). hroot: Hamburg Registration and Organization Online Tool. *European Economic Review 71*, 117–120.

Bénabou, R., A. Falk, L. Henkel, and J. Tirole (2024). Eliciting moral preferences under image concerns: Theory and experiment. *Working Paper*.

Bénabou, R. and L. Henkel (2025). Identity as self-image. *NBER Working Paper*, 34297.

Bénabou, R. and J. Tirole (2006). Incentives and prosocial behavior. *American Economic Review 96*(5), 1652–1678.

Bénabou, R. and J. Tirole (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics 126*(2), 805–855.

Casella, A., N. Kartik, L. Sanchez, and S. Turban (2018). Communication in context: Interpreting promises in an experiment on competition and trust. *Proceedings of the National Academy of Sciences of the USA 115*(5), 933–938.

Charness, G. and M. Dufwenberg (2006). Promises and partnership. *Econometrica 74*(6), 1579–1601.

Charness, G., A. Samek, and J. van de Ven (2022). What is considered deception in experimental economics? *Experimental Economics 25*, 385–412.

Cohn, A., M. A. Maréchal, D. Tannenbaum, and C. L. Zünd (2019). Civic honesty around the globe. *Science*, eaau8712.

Di Bartolomeo, G., M. Dufwenberg, S. Papa, and F. Passarelli (2019). Promises, expectations and causation. *Games and Economic Behavior 113*, 137—-146.

Dufwenberg, M. and M. A. Dufwenberg (2018). Lies in disguise – a theoretical analysis of cheating. *Journal of Economic Theory 175*, 248 – 264.

Ederer, F. and A. Stremitzer (2017). Promises and expectations. *Games and Economic Behavior 106*, 161—-178.

Ellingsen, T. and M. Johannesson (2004). Promises, threats and fairness. *The Economic Journal 114*(495), 397–420.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics 10*(2), 171–178.

Fischbacher, U. and F. Föllmi-Heusi (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association 11*(3), 525–547.

Fries, T. (2024). Signaling motives in lying games. *Games and Economic Behavior 147*, 338–376.

Fries, T., U. Gneezy, A. Kajackaite, and D. Parra (2021). Observability and lying. *Journal of Economic Behavior & Organization 189*, 132–149.

Gneezy, U., A. Kajackaite, and J. Sobel (2018). Lying Aversion and the Size of the Lie. *American Economic Review 108*(2), 419–453.

Gneezy, U., S. Saccardo, M. Serra-Garcia, and R. van Veldhuizen (2020). Bribing the Self. *Games and Economic Behavior 120*, 311–324.

Grubiak, K. P. (2025). Promises, image concerns, and excuses–An experimental investigation. *Frontiers in Behavioral Economics 4*.

Hertwig, R. and A. Ortmann (2008). Deception in experiments: Revisiting the arguments in its defense. *Ethics & Behavior 18*(1), 59–92.

Hey, J. (1998). Experimental economics and deception: A comment. *Journal of Economic Psychology 19*(3), 397–401.

Ismayilov, H. and J. Potters (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics 19*(2), 382—-393.

Jacquemet, N., S. Luchini, J. F. Shogren, and A. Zylbersztejn (2024). Commitment to the truth creates trust in market exchange: Experimental evidence. *Games and Economic Behavior 148*, 279–295.

Kartik, N. (2009). Strategic communication with lying costs. *Review of Economic Studies 76*(4), 1359–1395.

Khalmetski, K. and D. Sliwka (2019). Disguising lies—image concerns and partial lying in cheating games. *American Economic Journal: Microeconomics 11*(4), 79–110.

Lang, M. and S. Schudy (2023). (Dis)honesty and the value of transparency for campaign promises. *European Economic Review 159*, 104560.

Naka, M. and H. Naoi (1995). The effect of repeated writing on memory. *Memory & Cognition 23*(2), 201–212.

Schudy, S., S. Grundmann, and L. Spantig (2024). Individual preferences for truth-telling. *CESifo Working Paper Series 11521*.

Sengupta, A. and C. Vanberg (2023). Promise keeping and reliance damage. *European Economic Review 152*, 104344.

Skinner, C. H., T. McLaughlin, and P. Logan (1997). Cover, copy, and compare: A self-managed academic intervention effective across skills, students, and settings. *Journal of Behavioral Education 7*(3), 295–306.

Tergiman, C. and M. C. Villeval (2023). The Way People Lie in Markets: Detectable Vs. Deniable Lies. *Management Science 69*(6), 3340–3357.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations. *Econometrica 76*(6), 1467–1480.

Weitzel, U. and M. Kirchler (2023). The Banker's Oath And Financial Advice. *Journal of Banking & Finance 148*, 106750.

Zickfeld, J. H., K. A. Ścigała, C. T. Elbæk, J. Michael, M. H. Tønnesen, G. Levy, S. Ayal, I. Thielmann, L. Nockur, E. Peer, V. Capraro, R. Barkan, S. Bø, Bahník, D. Nosenzo, R. Hertwig, N. Mazar, A. Weiss, A.-K. Koessler, R. Montal-Rosenberg, S. Hafenbrädl, Y. A. Nielsen, P. Kanngiesser, S. Schindler, P. Gerlach, N. Köbis, N. Jacquemet, M. Vranka, D. Ariely, J. B. Martuza, Y. Feldman, M. Białek, J. K. Woike, Z. Rahwan, A. Seidl, E. Chou, A. Kajackaite, S. Schudy, U. Glogowsky, A. Z. Czarna, S. Pfattheicher, and P. Mitkidis (2025). Effectiveness of ex ante honesty oaths in reducing dishonesty depends on content. *Nature Human Behaviour 9*(1), 169–187.

ONLINE APPENDIX

A Online Appendix: Literature

Table A1: Overview of previous experimental studies on oath-taking on various behaviors

Reference	Task	Common knowledge	Compulsory	Findings
Strategic settings				
Jacquemet et al. (2018)	Coordination game	No	No	Positive impact on coordination rates.
Jacquemet et al. (2019)	Deception game	No	No	For selfish lies in loaded frame, the fraction of liars reduced from 35% in No Oath to 16.7% in (voluntary) Oath. Neutral frame does not reduce lies.
Beck (2021)	Deception game	Yes	Yes	Fraction of liars reduced from 42% in No Oath to 20% in (compulsory) Oath. No impact on the size of lies.
Jacquemet et al. (2021)	Deception game	No	No	For selfish lies, fraction of liars goes from 32.8% in No Oath to 14.3% in (voluntary) Oath. No impact on Pareto lies.
Koessler et al. (2021)	Public good game	Yes	Both	Positive impact of both mandatory and voluntary oaths on contributions, but the effect of the voluntary oath fades away more quickly than when it is mandatory.
Davis and Jaber-Lopez (2022)	Binary social dilemma game	No	Both	Impact of both mandatory and voluntary oaths, but crowding-out effect by non oath takers when it is voluntary.
Hergueux et al. (2022)	Public good game	No	No	Positive impact on contributions, according to social types.
Koessler (2022)	Public good game	Yes	No	Positive impact on contributions. More pledges when the majority of the group does so.
Weitzel and Kirchler (2023)	Audit study	Yes	Yes	Positive impact of reminders of the bankers' oath on real financial advisors' loan recommendations.
Jacquemet et al. (2024)	Trust game	No	No	Oath increases cooperation only with preplay communication. The effect of oath is equivalent to that of a mild fine.
Individual decision-making				
Carlsson et al. (2013)	Contingent valuation	-	No	Decrease of the shares of null and very high willingness-to-pay.
Jacquemet et al. (2013)	Second-price auctions	-	No	Decrease of the shares of null and very high willingness-to-pay.
Krüger (2016)	Coin cheating task	-	Yes	Both positive and negative effects depending on gender
Jacquemet et al. (2017)	Voting referenda	-	No	Decrease of the hypothetical bias.
Heinicke et al. (2019)	Number cheating task	-	Yes	Negative impact on extreme lies, no impact on partial lies, in both gain and loss frames.
Koessler et al. (2019)	Tax evasion game	-	No	No impact on compliance except when combined with non-financial rewards.
Schild et al. (2019)	Mind game	-	Yes	Negative impact when lies are unobservable, no impact when lies are observable.
Beck et al. (2020)	Die cheating task	-	Yes	Negative impact on both extreme and partial lies.
Jacquemet et al. (2020)	Tax evasion game	-	No	No impact on extreme lies, negative impact on partial lies.
Kemper et al. (2020)	Discrete choices	-	Yes	Decrease of hypothetical bias in the estimation of willingness-to-pay.
Prima et al. (2020)	Asset reporting	-	Yes	No impact on lies.
Jacquemet et al. (2021)	Coin cheating task	-	No	Negative impact on extreme lies, no impact on partial lies.
Peer and Feldman (2021)	Reporting perform.	-	Yes	Long-term negative impact on lies.
Akin (2022)	Coin cheating task	-	Yes	Negative effects on undeserved applications.
Babin et al. (2022)	Reporting eye color	-	No	Negative impact on extreme lies, no impact on partial lies.
Babin and Chauhan (2023)	Coin cheating task	-	No	Negative impact on both plausible and implausible misreporting.
Cagala et al. (2024)	Exam-taking	-	Yes	Cheating doubles.
Cagala et al. (2024)	Chip cheating task	-	Yes	No impact of commitment requests on lies in both offline and online settings.
Zickfeld et al. (2025)	Tax evasion game	-	No	Positive impact on tax compliance depending on context.

B Online Appendix: Instructions [Translated from French]

NO-OATH, PRIVATE-OATH, AND PRIVATE-OATH^{++} TREATMENTS

[The printed instructions in Private-Oath treatment and Private-Oath^{++} treatment are the same as in the No-Oath treatment. However, the A players in the Private-Oath treatments discovered about the oath on their computer screens at the beginning of part 2. During the oath-taking stage, A players had to type a sentence by which they swore on their honour that, during this experiment, they would be honest and always tell the truth. Meanwhile, the B players were asked to type a neutral sentence: "The Nobel Prize in Economics for the year 2021 has been awarded to three American economists."]

Preliminary instructions *[Only for the Private-Oath^{++} treatment]*

We are testing a new payment method for participants in experiments at GATE-Lab.

At the end of the session, you will receive an envelope containing your earnings in cash. This envelope will be given to you by an independent person who will never be informed of the content of the experiment, will not know the contents of your envelope, and will not have access to the experimental data.

After taking your cash from the envelope, you can leave the experimental room one by one. Please leave the instructions, the computer tag, and the envelope at your seat.

Before you leave GATE-Lab, you will see a table in the reception room showing only the earnings distributed during the session, presented in random order (i.e., with no mention of the computer numbers or other information). All you will have to do is sign a statement that one of the amounts shown on this list corresponds to the earnings you have received, but you do not have to indicate which one. While guaranteeing compliance with accounting rules, this procedure ensures that no one (neither the experimenters nor the person handing out the envelopes) will ever be able to link your decisions during the experiment, your earnings and your identity.

Welcome *[Common to all treatments]*

Hello and welcome to an experiment on decision-making. Please turn off your phone and put it away. You are not allowed to communicate with other participants during the whole experiment, otherwise you will be excluded from the session and any potential earnings.

During this session you can earn money. The amount of money you will earn depends on your decisions and the decisions of other participants in the session. Please read the instructions carefully. All your decisions and responses will be anonymous.

This session consists of two successive parts. The amount you earn at the end of this session is the sum of your earnings in the different parts plus your participation fee of 5 Euros. During the session we will not talk in Euros but in tokens. The conversion rate from tokens to Euros is as follows:

100 tokens = 4 Euros

You will be paid via bank transfer by the CNRS. To do this, at the end of the session we will ask for your IBAN number. We will also ask you to send us a bank statement (containing the IBAN) in pdf format with the title "your first name-your last name" to the following e-mail address: gatelab[at]gate.cnrs.fr. Due to administrative and banking delays, the transfer may take up to approximately two weeks. We promise that your bank details will only be used to make the transfer. This information will be kept separate from the files containing your decisions and the

researchers who process your decisions will not have access to your bank details. *[This paragraph was omitted in the Private-Oath^{++} treatment.]*

You will now find out the instructions for Part 1. The instructions for Part 2 will be explained at the end of Part 1.

Part 1 *[Common to all treatments]*

This part consists of 10 periods. The first 8 periods are training rounds and nothing you decide during these 8 periods counts towards determining your actual earnings. For these 8 training periods the earnings stated are therefore hypothetical.
The only periods that can count towards your earnings in this part are the 9th and 10th periods. At the end of the session, the program will draw period 9 or period 10 and the earnings in the drawn period will constitute your earnings for this part. Each of these two periods has 50 chances out of 100 to be drawn.

Description of the task
In each period, you receive an initial endowment of 100 tokens and you must decide whether to keep these tokens as your earnings for the period or to invest them all in an investment project. If this project is successful, you earn 3 times the number of tokens invested, that is 300 tokens. If the project is not successful, you earn 30 tokens.

Description of the investment project
In each period, 3 cards appear on your screen, face down. Each card can have a star ($\star$) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card.
Thus, your three cards can have a total of zero stars (which happens with 12.5 chances out of 100), a total of 1 star (which happens with 37.5 chances out of 100), a total of 2 stars (which happens with 37.5 chances out of 100), or a total of 3 stars (which happens with 12.5 chances out of 100).

You must press the "Reveal" button to reveal your three cards. In each case, the cards appear in the following format:

After revealing the cards, you must choose whether or not to invest your 100 tokens.
After your choice, the program draws one of your three cards, each card having the same chance of being drawn (so each card has 1 chance out of 3 to be selected).

- If the card drawn has a star and you have invested, the project is a success and you earn 300 tokens (*i.e.,* endowment of 100 – investment of 100 + earnings of 300).
- If the card drawn is blank and you have invested, the project is not a success and you earn 30 tokens (*i.e.,* endowment of 100 – investment of 100 + earnings of 30).
- If you did not invest your tokens, you keep your initial endowment of 100 tokens and therefore earn 100 tokens.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program from the three cards has a star or not.

Remember that in these 8 training periods, these earnings are hypothetical and nothing you decide in these periods counts towards determining your actual earnings.

To make sure that you have encountered all possible cases, we have in advance chosen cases that correspond to the probabilities announced in the task description.

Period 9
If drawn at the end of the session, this period determines your actual earnings for this part. The rules and task are the same as in the previous 8 periods. The only difference is in the way you must make your investment choice.

Your screen will show 3 cards, face down. Each card has 50 chances out of 100 to have a star. In this period, you have to make a decision in each of the following 4 scenarios. Would you invest in the project or not if the program announced that among the three cards there are :

Scenario a) 0 stars?

Scenario b) 1 star?

Scenario c) 2 stars?

Scenario d) 3 stars?

Once you have answered these questions, the program will inform you of the total number of stars among your three cards.

Your earnings
If this period 9 is drawn, your earnings are determined by the answer to the scenario that applies. That is, the one that corresponds to the total number of stars among your three cards. For example, suppose that the three cards hide a total of two stars; in this case, your decision in scenario (c) applies. Another example, suppose the three cards hide a total of three stars; in this case, your decision in scenario (d) applies.

The program then draws one of your three cards at random.

- If you have invested in the project and the card drawn has a star, then you earn 300 tokens (endowment of 100 - investment of 100 + earnings of 300).
- If you have invested in the project and the card drawn is blank, then you win 30 tokens (endowment of 100 - investment of 100 + earnings of 30).
- If you have not invested, you earn the 100 tokens of your initial endowment.

As you can see, the principle is the same as in the 8 training periods, but here you make a decision in each possible scenario.
Since only one of your responses will count towards your earnings if this period is drawn, when you make your decision in each scenario it is in your best interest to treat each scenario as if it were the one that actually counts towards your earnings for this part.

Period 10
If drawn at the end of the session, this period determines your actual earnings for this part. As in the previous periods your screen will show 3 cards, face down. Each card has 50 chances out of 100 to have a star. You must again decide whether or not to invest in the project. However, unlike the previous periods, you only have to make one decision without being informed of the number of cards with a star. Only at the end of the session will you be informed by the programme of the total number of stars among your three cards if this period is selected for payment.
Once you have made your decision, the program will then draw one of the three cards at random.

Your earnings
If period 10 is drawn for payment, your earnings are determined as follows:

- If you have invested and the card drawn has a star, then you earn 300 tokens.
- If you have invested and the card drawn is blank, then you earn 30 tokens.
- If you did not invest, you earn the 100 tokens from your initial endowment.

You will be informed of the period drawn (9 or 10), the card drawn and your earnings in that round at the end of the session.

Please read these instructions again. If you have any questions, please raise your hand or press the red button on the side of your desk. We will come and answer your questions in private immediately.

Part 2 *(Instructions distributed after completion of part 1)*

In this part, each of you will be given a role, either 'A' or 'B'. Half of the participants have a role of A and the other half have a role of B. Your role remains the same for the entirety of part 2: you will never change it.
Part 2 has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period, each of you is paired such that there is one participant A and one participant B in each pair. You will never know the identity of the participant you are paired with. At the beginning of each period, you are randomly re-matched with a new participant. It is unlikely that you will be paired with the same participant two periods in a row.

Your task in each period
Participant A: Participant A sees three cards on his/her screen, face down. Each card can have a star ($\star$) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card. Thus, participant A can have *a total of* 0 stars, 1 star, 2 stars or 3 stars. Participant A can see how many stars he/she has by pressing the "Reveal" button.
Participant A's task is to then announce his/her total number of stars to participant B with whom he/she is matched with. Participant B cannot see Participant A's cards at any time.

Participant B: Participant B receives an initial endowment of 100 tokens and he/she is only informed of the announcement of Participant A with whom he/she is paired about his/her number of stars. Participant B has to decide whether he/she wants to keep his/her tokens or invest them in Participant A's project.

Determination of earnings

If you are Participant A: You earn a fixed amount of 30 tokens, plus 200 tokens if Participant B has invested in your project.
If you are Participant B: Once you have made your investment decision, the program draws one of Participant A's three cards at random.

- If you have invested in the project and the card drawn has a star, the project is a success and you earn 300 tokens (*i.e.,* endowment of 100 - investment of 100 + earnings of 300).
- If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens (*i.e.,* endowment of 100 - investment of 100 + earnings of 30).
- If you did not invest, you earn the 100 tokens of your initial endowment.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

Your screen
At the end of each period, your screen will show a history of what happened in the previous periods. Specifically, you will see four types of information:

1. your announcements in previous periods, if you are Participant A; or the announcements of the different Participants A you were paired with, if you are Participant B;
2. whether the card drawn in previous periods had a star or not, regardless of your role and decision;
3. whether you invested in the previous periods, if you are Participant B; or whether the Participants B you were matched with invested or not, if you are Participant A.
4. your potential earnings in each previous period.

For each of you, **only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part**. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.
Remember that **the pairs are re-matched in each period**.

[***ADDITIONAL PARAGRAPH IN PUBLIC-OATH TREATMENT ONLY***]

Oath
At the beginning of part 2, **Participants A will have to take an oath about the truthfulness of their announcements in part 2**. By this oath, they will swear on their honour that during this experiment they will be honest and always tell the truth.

[After the instructions and role assignment, the A players in Public-Oath treatment swore the oath by typing it on their computer. The B players did not type anything during this stage. After all A players took the oath, every participant was informed that the oath had been taken.]

- - -

Please read these instructions again. If you have any questions, please raise your hand or press the red button. We will come and answer your questions in private immediately.

- - -

End of the instructions in part 2

[At the end of part 2 but before informing participants of their final gains, we elicited beliefs (displayed on computer screen only), depending on their role in the Announcement Game. For the B players, we asked the following question about their beliefs concerning the honesty of the A players in their session.]
Please estimate below the percentage of true announcements among all the announcements you received in part 2. Enter integer values between 0 and 100. At the end of the session the program will draw one of these questions. If your answer is correct, plus or minus 5%, you will earn an extra 0.50€.

In your opinion, what percentage of announcements were true when Participant A announced:

- 0 stars: _%
- 1 star: _%
- 2 stars:_%
- 3 stars:_%

[For the A players, we asked two questions (in randomized order at an individual level). One elicited their beliefs about the honesty of the A players in their session, which reads as follows:]
Please estimate below the percentage of true announcements among all announcements made in part 2 by all participants A in this session. Enter integer values between 0 and 100. At the end of the session the program will draw one of these questions. If your answer is correct, plus or minus 5%, you will earn an extra 0.50€.

In your opinion, what percentage of announcements were true when participants A in this session announced:

- 0 stars: _%
- 1 star: _%
- 2 stars:_%
- 3 stars:_%

[The other question asked about their beliefs about what the B players think about their honesty, which reads as follows:]
Your participant B is answering the following question: "In your opinion, what percentage of announcements were true when participant A told you?".
What do you think your participant B's answer was to these questions?
Enter integer values between 0 and 100. At the end of the session the program will draw one of these questions. If your answer is correct, plus or minus 5%, you win an extra 0.50€.

- 0 stars: _%
- 1 star: _%
- 2 stars:_%
- 3 stars:_%

[After the belief elicitations, participants were informed about their final gains (flat payment and additional payoffs from part 1, part 2, belief questions). They then completed a final questionnaire consisting of the following information:]

- Age
- Gender
- Status
- School
- *[Player A only]* In part 2, if sometimes you did not announce the correct number of stars or if you announce the correct number of stars all the time, can you explain why you did so? If you have different reasons, can you rank them?
- *[Player B only]* In part 2, if sometimes you have invested or not in Participant A's project, can you explain why? If you have different reasons, can you rank them.
- What do you think this experiment is testing?
- If you had to leave a piece of advice for a successor in your role, what would it be?

- - -

NECESSARY CONDITION TREATMENT

[The written instructions for part 1 of the experiment are the same as in the other treatments. The following instruction was given to participants for part 2.]

Part 2

Some weeks ago, participants from the same subjects pool as you took part in an experimental session. These participants performed the same task as you in part 1, under exactly the same conditions as you did. Please read below an extract from the instructions they were given for part 2 (text in italics). The supplementary instructions that only concern you will be given after.

Extract from the instructions given to past participants

In this part, each of you will be given a role, either 'A' or 'B'. Half of the participants have a role of A and the other half have a role of B. Your role remains the same for the entirety of part 2: you will never change it.
Part 2 has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period, each of you is paired such that there is one participant A and one participant B in each pair. You will never know the identity of the participant you are paired with. At the beginning of each period, you are randomly re-matched with a new participant. It is unlikely that you will be paired with the same participant two periods in a row.

Your task in each period
Participant A: Participant A sees three cards on his/her screen, face down. Each card can have a star (⋆) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card. Thus, participant A can have a total of *0 stars, 1 star, 2 stars or 3 stars. Participant A can see how many stars he/she has by pressing the "Reveal" button.*

Participant A's task is to then announce his/her total number of stars to participant B with whom he/she is matched with. Participant B cannot see Participant A's cards at any time.

Participant B: Participant B receives an initial endowment of 100 tokens and he/she is only informed of the announcement of Participant A with whom he/she is paired about his/her number of stars. Participant B has to decide whether he/she wants to keep his/her tokens or invest them in Participant A's project.

Determination of earnings
If you are Participant A: You earn a fixed amount of 30 tokens, plus 200 tokens if Participant B has invested in your project.

If you are Participant B: Once you have made your investment decision, the program draws one of Participant A's three cards at random.

- *If you have invested in the project and the card drawn has a star, the project is a success and you earn 300 tokens (i.e., endowment of 100 - investment of 100 + earnings of 300).*
- *If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens (i.e., endowment of 100 - investment of 100 + earnings of 30).*

- *If you did not invest, you earn the 100 tokens of your initial endowment.*

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

Your screen
At the end of each period, your screen will show a history of what happened in the previous periods. Specifically, you will see four types of information:

1. *Your announcements in previous periods, if you are Participant A; or the announcements of the different Participants A you were paired with, if you are Participant B;*
2. *Whether the card drawn in previous periods had a star or not, regardless of your role and decision;*
3. *Whether you invested in the previous periods, if you are Participant B; or whether the Participants B you were matched with invested or not, if you are Participant A.*
4. *Your potential earnings in each previous period.*

For each of you, ***only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part****. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.*

Remember that ***the pairs are re-matched in each period****.*

End of the extract

Supplementary instructions for participants in today's session

The previous instructions apply to you in this part as well but there are some differences from the previous extract:

1. **All participants in today's session are assigned to <u>the role B</u>**; you will keep this role throughout the part.
2. Participants in **role A are not present in this session**: they have already participated in a previous session with other participants in role B.
3. At the beginning of each period, you will be matched with a participant A from a previous session randomly selected by the program.
4. Before you decide whether to invest, you will be informed of the announcement made by this participant A from a previous session after the initial draw of his/her three cards.
5. After deciding whether or not to invest, you will be informed of the card that was randomly drawn by the program from the three cards of this participant A from a previous session. This card drawn will determine your earnings for the period if you have decided to invest.
6. Your investment decisions will only affect your earnings and not the earnings of the participants A from previous sessions with whom you are matched.

Summary
This part has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period you will be informed about the number of stars (0, 1, 2 or 3) announced by a participant A from a previous session in a period drawn at random. You will then have to decide whether or not to invest in the project of participant A from the previous session.

- If you have invested in the project and the card drawn is a star, the project is a success and you earn 300 tokens.
- If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens.
- If you have not invested, you earn the 100 tokens of your initial endowment.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

For each of you, **only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part**. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.

Remember that **in each period the program randomly selects a decision** (*i.e.,* an announcement made after an initial draw of three cards) **from all the decisions made by participants A in previous sessions. You are re-matched with a new participant A in each period** and it is very unlikely that the program will select the same participant A for you two periods in a row.

- - -

Please read these instructions again. If you have any questions, please raise your hand or press the red button. We will come and answer your questions immediately in private.

- - -

End of the instructions in part 2

[In the Necessary Condition treatment, at the end of part 2 but before informing participants of their final gains, we elicited beliefs (displayed on computer screen only), depending on their role in the Announcement Game like in the other treatments. However, after the belief elicitation about the proportion of honest announcements, participants in this treatment, who all had the role of B, were asked to compare their responses to their initial expectations about Player A's honesty before starting the Announcement Game.]

We remind you below of your estimates of the proportions of true announcements received in part 2. Please let us know how well these estimates match your initial expectations before you began part 2. Specifically, indicate for each scenario whether your estimates are:

- More or less equal to your initial expectations

- Higher than your initial expectations (before the part started, you expected a lower proportion of true announcements)
- Lower than your initial expectations (before the part started, you expected a higher proportion of true announcements)

Your estimate of the proportion of truthful announcements are ...

- 0/1/2/3 stars (corresponding estimate shown here):
  - More or less equal to your initial expectations
  - Higher than your initial expectations ...
  - Lower than your initial expectations ...

- - -

SUFFICIENT CONDITION TREATMENT

[The written instructions for part 1 of the experiment are the same as in the other treatments. In part 2, like in the Necessary Condition treatment, participants were first given the extract of the instructions used in Public-Oath treatment. Note that the extract part for the Sufficient Condition treatment included information about the oath, as in the Public-Oath treatment.]

Part 2

Some weeks ago, participants from the same subjects pool as you took part in an experimental session. These participants performed the same task as you in part 1, under exactly the same conditions as you did. Please read below an extract from the instructions they were given for part 2 (text in italics). The supplementary instructions that only concern you will be given after.

Extract from the instructions given to past participants

In this part, each of you will be given a role, either 'A' or 'B'. Half of the participants have a role of A and the other half have a role of B. Your role remains the same for the entirety of part 2: you will never change it.
Part 2 has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period, each of you is paired such that there is one participant A and one participant B in each pair. You will never know the identity of the participant you are paired with. At the beginning of each period, you are randomly re-matched with a new participant. It is unlikely that you will be paired with the same participant two periods in a row.

Your task in each period
Participant A: Participant A sees three cards on his/her screen, face down. Each card can have a star (⋆) or be blank. Each card has 50 chances out of 100 to have a star and 50 chances out of 100 to be blank. These chances are independent for each card. Thus, participant A can have a total of *0 stars, 1 star, 2 stars or 3 stars. Participant A can see how many stars he/she has by pressing the "Reveal" button.*

Participant A's task is to then announce his/her total number of stars to participant B with whom he/she is matched with. Participant B cannot see Participant A's cards at any time.

Participant B: Participant B receives an initial endowment of 100 tokens and he/she is only informed of the announcement of Participant A with whom he/she is paired about his/her number of stars. Participant B has to decide whether he/she wants to keep his/her tokens or invest them in Participant A's project.

Determination of earnings
If you are Participant A: You earn a fixed amount of 30 tokens, plus 200 tokens if Participant B has invested in your project.

If you are Participant B: Once you have made your investment decision, the program draws one of Participant A's three cards at random.

- *If you have invested in the project and the card drawn has a star, the project is a success and you earn 300 tokens (i.e., endowment of 100 - investment of 100 + earnings of 300).*

- *If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens (i.e., endowment of 100 - investment of 100 + earnings of 30).*

- *If you did not invest, you earn the 100 tokens of your initial endowment.*

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

Your screen

At the end of each period, your screen will show a history of what happened in the previous periods. Specifically, you will see four types of information:

1. *Your announcements in previous periods, if you are Participant A; or the announcements of the different Participants A you were paired with, if you are Participant B;*

2. *Whether the card drawn in previous periods had a star or not, regardless of your role and decision;*

3. *Whether you invested in the previous periods, if you are Participant B; or whether the Participants B you were matched with invested or not, if you are Participant A.*

4. *Your potential earnings in each previous period.*

For each of you, ***only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part****. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.*

Remember that ***the pairs are re-matched in each period****.*

Oath

At the beginning of part 2, ***Participants A will have to take an oath about the truthfulness of their announcements in part 2****. By this oath, they will swear on their honour that during this experiment they will be honest and always tell the truth.*

End of the extract

Supplementary instructions for participants in today's session

The previous instructions apply to you in this part as well but there are some differences from the previous extract:

1. **All participants in today's session are assigned to <u>the role B</u>**; you will keep this role throughout the part.

2. Participants in **role A are not present in this session**: they have already participated in a previous session with other participants in role B. The excerpt of the instructions you read above was for these past participants. Before starting the part, **all participants in role A had to take an oath on honor** that during the experiment they would behave honestly and would always tell the truth. These participants in role A had to type the text of the oath on their computer and this was made known to the participants in role B.

3. **At the beginning of each period, you will be matched with a participant A, randomly selected from a subset of these participants A.**

4. Before you decide whether to invest, you will be informed of the announcement made by this participant A from a previous session after the initial draw of his/her three cards.

5. After deciding whether or not to invest, you will be informed of the card that was randomly drawn by the program from the three cards of this participant A from a previous session. This card drawn will determine your earnings for the period if you have decided to invest.

6. Your investment decisions will only affect your earnings and not the earnings of the participants A from previous sessions with whom you are matched.

Summary
This part has a minimum of 10 periods and a maximum of 30 periods. The exact number of periods has been decided before the start of the session.

In each period you will be informed about the number of stars (0, 1, 2 or 3) announced by a participant A from a previous session in a period drawn at random. You will then have to decide whether or not to invest in the project of participant A from the previous session.

- If you have invested in the project and the card drawn is a star, the project is a success and you earn 300 tokens.

- If you have invested in the project and the card drawn is blank, the project is not a success and you earn 30 tokens.

- If you have not invested, you earn the 100 tokens of your initial endowment.

Whatever your choice is, you are informed at the end of the period whether the card drawn by the program among the three cards has a star or not.

For each of you, **only one period has already been drawn by the program for payment and it is your decision in that period that will determine your earnings in this part**. Thus, it is in your best interest to make your decision in each period as if it were the period that counts towards your earnings in this part.

Remember that **in each period you will face a decision** (i.e., an announcement made after an initial draw of three cards) **made by one participant A, chosen at random from a subset of the participants A from previous sessions. You are rematched to a new participant A each period** and it is unlikely that the program will select the same participant A for you two periods in a row.

- - -

Please read these instructions again. If you have any questions, please raise your hand or press the red button. We will come and answer your questions immediately in private.

- - -

End of the instructions in part 2

[In the Sufficient Condition treatment, at the end of part 2 but before informing participants of their final gains, we elicited beliefs (displayed on computer screen only), depending on their role in the Announcement Game like in the other treatments. After the belief elicitation about the proportion of honest announcements, participants in this treatment were asked to compare their responses to their initial expectations about the A players' honesty before starting part 2. Note that the wording of the question was slightly adapted between Necessary Condition and Sufficient Condition treatments.]

We will ask you questions about the possible differences between your initial expectations (after reading the instructions but before starting the part) and your experience during the experiment regarding the reliability of the announcements received from participants A.

- When the announcement was 0 stars:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 0-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 0 announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations
- When the announcement was 1 star:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 1-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 1-star announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations
- When the announcement was 2 stars:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 2-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 2-star announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations
- When the announcement was 3 stars:
  - My experience has shown me that these announcements were LESS RELIABLE than my expectations (the 3-star announcements were less truthful than expected)
  - My experience has shown me that these announcements were MORE RELIABLE than I expected (the 3-star announcements were more truthful than expected)
  - My experience has shown me that the reliability of these announcements was NEARLY EQUAL to my expectations

- - -

C Online Appendix: Tables

Table C1: Summary statistics of Advisors

	(1) No-Oath		(2) Public-Oath		(3) Private-Oath		(4) Private-Oath^{++}		(1-2)	(1-3)	(2-3)	(3-4)
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	*p-value*			
Age (Years)	21.34	3.62	20.90	1.84	20.92	2.13	21.19	2.32	0.626	0.656	0.968	0.188
Male (%)	0.66	0.48	0.44	0.50	0.44	0.50	0.55	0.50	0.007**	0.007**	0.969	0.188
Business school (%)	0.50	0.50	0.61	0.49	0.40	0.49	0.45	0.50	0.170	0.208	0.009**	0.490
Number of observations	76		77		73		75					

Notes: This table summarizes the socio-demographic characteristics of participants in the role of Advisor (Player A) for each treatment. The *p-value* reported using ranksum test for age and tests of proportion for gender (coded 1 as male, 0 otherwise) and business school (coded 1, 0 otherwise). Regression analyses controlled for any effects of these characteristics. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table C2: Summary statistics of Investors

	(1) No-Oath		(2) Public-Oath		(3) Private-Oath		(4) Private-Oath^{++}		(5) Necessary Cond.		(6) Sufficient Cond.	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Age (Years)	20.82	2.03	20.60	1.32	21.42	2.11	21.63	2.48	21.71	1.85	20.99	1.84
Male (%)	0.61	0.49	0.38	0.49	0.49	0.50	0.44	0.50	0.44	0.50	0.64	0.48
Business school (%)	0.54	0.50	0.66	0.48	0.48	0.50	0.49	0.50	0.48	0.50	0.36	0.48
Number of observations	76		77		73		75		77		78	

Notes: This table summarizes the socio-demographic characteristics of participants in the role of Investors (Player B) for each treatment. The *p-values* are reported in the next table.

Table C3: Between-treatment comparisons of socio-demographics of Investors

	No-Oath *vs.* Public-Oath	No-Oath *vs.* Private-Oath	Public-Oath *vs.* Private-Oath	Public-Oath *vs.* Necessary Cond.	Private-Oath *vs.* Necessary Cond.	Public-Oath *vs.* Sufficient Cond.	No-Oath *vs.* Sufficient Cond.	Private-Oath *vs.* Private-Oath^{++}
Age (Years)	$p = 0.935$	$p = 0.024$	$p = 0.007$	$p < 0.001$	$p = 0.126$	$p = 0.212$	$p = 0.330$	$p = 0.610$
Male (%)	$p = 0.005$	$p = 0.169$	$p = 0.150$	$p = 0.413$	$p = 0.527$	$p = 0.001$	$p = 0.647$	$p = 0.517$
Business school (%)	$p = 0.121$	$p = 0.464$	$p = 0.024$	$p = 0.023$	$p = 0.990$	$p < 0.001$	$p = 0.024$	$p = 0.866$

Note: This table reports the p-values from ranksum tests (for age) and tests of proportions (for gender and school) comparing the summary statistics of Investors (Player B) across treatments. Regression analyses controlled for any effects of these characteristics.

Tables C4 and C5 present the results of random-effects Linear Probability models (GLS) in which the dependent variable is the decision to make a detectable or a deniable lie, respectively. Models (1) and (2) consider all Advisors. Models (3) and (4) are restricted to Advisors who lied at least once in part 2.

Alternatively to models (3) and (4), we estimated two-step Heckman models, estimating in the first step the probability of making a lie and, in the second step, the probability of making a detectable or a deniable lie, using gender as the identifying variable. However, since the IMR was not significant, meaning that there is no need to correct for a selection bias, we omit these regressions and only report the GLS models.

Given random matching design, Table C6 reports p-values with alternative clustering at the session level (with and without bootstrapping), which confirms the robustness of our results with clustering at the individual level.

Table C4: Probability of making detectable lies

Dependent variable:	All Advisors		Lied at least once	
Detectable lie	(1)	(2)	(3)	(4)
No-Oath	*Ref.*	*Ref.*	*Ref.*	*Ref.*
Public-Oath	-0.201***	-0.210***	-0.121	-0.166*
	(0.038)	(0.049)	(0.071)	(0.080)
Private-Oath	-0.147***	-0.153**	-0.112*	-0.125*
	(0.040)	(0.051)	(0.052)	(0.062)
Period	0.001	0.001	0.002	0.001
	(0.001)	(0.004)	(0.002)	(0.004)
Male	0.064*	0.064*	0.058	0.058
	(0.030)	(0.030)	(0.048)	(0.048)
Age	-0.008*	-0.008*	-0.010*	-0.010*
	(0.003)	(0.003)	(0.004)	(0.004)
Business School	-0.015	-0.015	-0.008	-0.008
	(0.030)	(0.030)	(0.047)	(0.047)
Public-Oath X Period		0.001		0.005
		(0.004)		(0.005)
Private-Oath X Period		0.001		0.001
		(0.004)		(0.005)
Constant	0.382***	0.387***	0.430***	0.441***
	(0.088)	(0.094)	(0.108)	(0.113)
Number of observations	3551	3551	2096	2096
Number of clusters	226	226	133	133

Notes: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to make a detectable lie (*i.e.*, announcing three stars, conditional on observing less than three stars), for all Advisors (models (1) and (2)) and those who lied at least once (models (3) and (4)). Independent variables in model (1) include treatment dummies (with No-Oath treatment as the reference category), period, a male dummy, age (in years), and a dummy for being a student at the business school. Model (2) includes interaction terms between the treatment dummies and the period variable. Models (3) and (4) correspond to models (1) and (2) respectively for the subset of Advisors who lied at least once. Standard errors are clustered at the individual level. This analysis excludes the Private-Oath^{++} treatment. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table C5: Probability of making deniable lies

Dependent variable:	All Advisors		Lied at least once	
Deniable lie	(1)	(2)	(3)	(4)
No-Oath	*Ref.*	*Ref.*	*Ref.*	*Ref.*
Public-Oath	-0.406***	-0.424***	-0.094	-0.252*
	(0.055)	(0.065)	(0.086)	(0.121)
Private-Oath	-0.171**	-0.242***	0.013	-0.135
	(0.063)	(0.072)	(0.061)	(0.086)
Period	0.007***	0.004	0.012***	0.004
	(0.002)	(0.004)	(0.003)	(0.005)
Male	0.109*	0.109*	0.034	0.036
	(0.049)	(0.049)	(0.058)	(0.057)
Age	0.008	0.008	0.007	0.007
	(0.007)	(0.007)	(0.006)	(0.006)
Business School	-0.007	-0.007	0.031	0.034
	(0.047)	(0.047)	(0.056)	(0.056)
Public-Oath X Period		0.002		0.017
		(0.005)		(0.009)
Private-Oath X Period		0.007		0.016*
		(0.005)		(0.007)
Constant	0.277	0.307	0.325*	0.401*
	(0.160)	(0.163)	(0.153)	(0.157)
Number of observations	2039	2039	1179	1179
Number of clusters	226	226	133	133

Notes: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to make a deniable lie (*i.e.*, announcing one or two stars conditional on observing zero or one star, respectively) for all Advisors (models (1) and (2)) and those who lied at least once (models (3) and (4)). Independent variables in model (1) include treatment dummies (with No-Oath treatment as the reference category), period, a male dummy, age (in years), and a dummy for being a student at the business school. Model (2) includes interaction terms between the treatment dummies and the period variable. Models (3) and (4) correspond to models (1) and (2), respectively, for the subset of Advisors who lied at least once. Standard errors are clustered at the individual level. This analysis excludes the Private-Oath^{++} treatment. * p<0.05; ** p<0.01; *** p<0.001.

Table C6: Robustness checks: p-values with alternative clustering.

	Detectable lies				Deniable lies			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
Session level (bootstrap)								
Public-Oath	0.000	0.000	0.152	0.073	0.000	0.000	0.353	0.123
Private-Oath	0.000	0.000	0.009	0.007	0.009	0.003	0.857	0.163
Public vs. Private	0.142	0.158	0.913	0.658	0.000	0.001	0.215	0.480
Session level								
Public-Oath	0.000	0.000	0.120	0.051	0.000	0.000	0.315	0.095
Private-Oath	0.000	0.000	0.005	0.005	0.007	0.002	0.853	0.156
Public vs. Private	0.121	0.126	0.907	0.634	0.000	0.001	0.179	0.441
Individual level								
Public-Oath	0.000	0.000	0.091	0.039	0.000	0.000	0.273	0.037
Private-Oath	0.000	0.003	0.030	0.042	0.006	0.001	0.835	0.115
Public vs. Private	0.098	0.078	0.905	0.583	0.000	0.001	0.226	0.355

Notes: Columns (1)–(4) under *Detectable lies* correspond to models (1)–(4) in Table C4, and columns (1)–(4) under *Deniable lies* correspond to models (1)–(4) in Table C5. Reported p-values are from alternative clustering at the session level with and without bootstraping. Bootstrapped p-values are obtained using 1,000 repetitions.

Table C7: Detectable and deniable lies in the Private-Oath and Private-Oath^{++} treatments

Treatments	Private-Oath	Private-Oath^{++}	p-value
All Advisors			
% Advisors who always tell the truth	39.7%	38.7%	$p = 0.895$
% Untrue announcements	27.7%	29.2%	$p = 0.734$
Relative frequency			
Detectable lies	9.6%	10.4%	$p = 0.629$
Deniable lies	37.9%	38.1%	$p = 0.896$
Absolute frequency			
Detectable lies	32.9%	36.0%	$p = 0.689$
Deniable lies	56.2%	58.7%	$p = 0.758$
Among Advisors who lie at least once			
% Untrue announcements	46.0%	47.6%	$p = 0.645$
Relative frequency			
Detectable lies	15.9%	16.9%	$p = 0.593$
Deniable lies	62.8%	62.2%	$p = 0.934$
Absolute frequency			
Detectable lies	54.5%	58.7%	$p = 0.691$
Deniable lies	93.2%	95.7%	$p = 0.609$

Notes: Relative frequency statistics show the average frequency with which participants engage in any particular type of lie. Absolute frequency statistics show the proportion of Advisors who make any particular type of lie at least once. The relative frequency of detectable lies corresponds to how often Advisors who saw fewer than three stars announced three stars. The relative frequency of deniable lies corresponds to how frequently Advisors who saw fewer than two stars reported two stars or announced one star while seeing none. This table presents data both for all Advisors (top panel) and for the subset who make at least one lie (lower panel).

Table C8: Determinants of the investment decision (two- and three-star announcements)

Dependent variable: Investment decision	(1)	(2)
No-Oath	*Ref.*	*Ref.*
Public-Oath	0.171***	0.096*
	(0.035)	(0.038)
Private-Oath	0.049	-0.001
	(0.040)	(0.041)
Necessary Cond.	0.166***	0.101*
	(0.039)	(0.041)
Sufficient Cond.	0.072	0.066
	(0.039)	(0.040)
Period	-0.005***	-0.004***
	(0.001)	(0.001)
Male	0.027	0.031
	(0.024)	(0.025)
Age	0.000	0.001
	(0.006)	(0.006)
Business	-0.010	-0.017
	(0.025)	(0.025)
Index of lies detected		-0.685***
		(0.107)
% of blank cards on 2 Stars		-0.064*
		(0.027)
Constants	0.754***	0.834***
	(0.131)	(0.132)
Number of observations	6146	5552
Number of clusters	381	381
p-values from Wald tests		
Necessary Cond. *vs.* Public-Oath	0.895	0.892
Sufficient Cond. *vs.* Public-Oath	0.006	0.411

Notes: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to invest, conditioned on receiving a two- or three-star announcement (excluding Private-Oath^{++} treatment). Independent variables in model (1) include treatment dummies (with No-Oath treatment as the reference category), period, a male dummy, age (in years), and a dummy for being a student at the business school. Model (2) includes controls for the percentage of the time a lie has been detected up to the period ('Index of detection') and the percentage of blank cards on two-star announcements up to the period ('% of blank cards on 2 Stars'). Note that interaction terms between the treatments and these two control variables are not significantly different from zero; thus, we do not report this additional model. Standard errors are clustered at the individual level. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

Table C9: Robustness checks for Table C8: p-values under alternative clustering

	Two- and three-star investment	
	(1)	(2)
Session level (bootstrap)		
Public-Oath	0.000	0.018
Private-Oath	0.197	0.991
Necessary Cond.	0.000	0.004
Sufficient Cond.	0.148	0.221
Necessary Cond. *vs.* Public-Oath	0.889	0.868
Sufficient Cond. *vs.* Public-Oath	0.038	0.493
Session level		
Public-Oath	0.000	0.011
Private-Oath	0.156	0.990
Necessary Cond.	0.000	0.002
Sufficient Cond.	0.111	0.136
Necessary Cond. *vs.* Public-Oath	0.879	0.860
Sufficient Cond. *vs.* Public-Oath	0.022	0.382
Individual level		
Public-Oath	0.000	0.012
Private-Oath	0.223	0.993
Necessary Cond.	0.000	0.014
Sufficient Cond.	0.066	0.099
Necessary Cond. *vs.* Public-Oath	0.895	0.892
Sufficient Cond. *vs.* Public-Oath	0.006	0.411

Notes: Columns (1)–(2) correspond to models (1)–(2) in Table C8. Reported p-values are based on alternative clustering at the session level, with and without bootstrapping. Bootstrapped p-values are obtained using 1,000 repetitions.

D Online Appendix: Primacy of Experience

We report investment rates across treatments to examine whether knowledge of the oath is sufficient to sustain trust over time, or necessary to generate it. Given our focus on the role of experience, we restrict Investors' behavior to the last nine periods for the Sufficient Condition and Necessary Condition treatments and compare them with overall behavior in Public-Oath, No-Oath and Private-Oath treatments. Our conclusion is similar if we instead restrict to the first 18 periods for all treatments (see Tables D3 and D4).

A sufficient condition?: Recall that Investors in the Sufficient Condition treatment were informed about the oath but faced largely untruthful announcements similar to No-Oath treatment. Table D1 compares the average investment rates in the last nine periods of the Sufficient Condition treatment with average behavior in Public-Oath and No-Oath. Investors' behavior in the Sufficient Condition treatment differs significantly from that in the Public-Oath treatment but matches that of the No-Oath treatment. This is the case if we look at the overall average investment rates, but also if we look at investment rates after particular announcements. These patterns show that knowledge of the oath is not sufficient to lead Investors to trust Advisors. Instead, Investors are influenced by their own experience: when they know about the existence of an oath but realize that Advisors have perjured themselves, they become less likely to invest.[1]

Table D1: Investment rates in the Sufficient Condition (Experienced Players), Public-Oath and No-Oath treatments

	Sufficient Cond. (1)	Public-Oath (2)	No-Oath (3)	*p-value* *(1) vs. (2)*	*(1) vs. (3)*
0 stars announced	0.0%	1.7%	9.5%	$p = 1.000$	$p = 1.000$
1 star announced	6.6%	16.8%	14.5%	$p = 0.001$	$p = 0.201$
2 stars announced	64.1%	85.8%	70.5%	$p = 0.002$	$p = 0.736$
3 stars announced	80.2%	97.7%	77.8%	$p < 0.001$	$p = 0.085$
Average investment rate	65.1%	56.9%	65.8%	$p = 0.007$	$p = 0.575$
Average earnings of Investors	146.7	159.9	149.7	$p = 0.010$	$p = 0.446$

Notes: The table shows the fraction of times Investors invested in the Announcement Game, looking at the last nine periods in Sufficient Condition treatment to allow participants to gain experience vs. all 18 periods in Public-Oath and No-Oath treatments.

A necessary condition?: We turn to the Necessary Condition treatment where Investors were unaware of the oath, but faced largely truthful announcements similar to the Public-Oath treatment.

[1] A complementary timetrends analysis using random-effects Linear Probability models of the probability of investing (see Table D5) confirms this. While investment is stable in Public-Oath treatment ($p = 0.843$), the coefficient of period is negative and significant for No-Oath and Sufficient Condition treatments ($p = 0.022$ and $p < 0.001$, respectively). See also Figure **??** which displays the evolution of investment rates in high announcements.

Table D2 compares the average investment rates in the last nine periods of the Necessary Condition treatment with average behavior in Public-Oath and Private-Oath. Despite the lack of knowledge about the oath, Investors in the Necessary Condition treatment behave similarly to those in the Public-Oath treatment. The gap in investment rates after two- or three-star announcements across treatments is always small in magnitude and statistically not significant. In addition, compared to the Private-Oath, where oath is also unknown but Advisors are less truthful, Investors in the Necessary Condition treatment invest significantly more following high announcements. Knowledge about the oath in itself is not necessary for trust to develop.

Table D2: Investment rates in the Necessary Condition (Experienced Players), Public-Oath, and Private-Oath treatments

	Necessary Cond. (1)	Public-Oath (2)	Private-Oath (3)	*p-value* *(1) vs. (2)*	*(1) vs. (3)*
0 stars announced	1.1%	1.7%	4.0%	$p = 1.000$	$p = 0.546$
1 star announced	11.3%	16.8%	13.0%	$p = 0.027$	$p = 0.712$
2 stars announced	84.7%	85.8%	72.6%	$p = 0.207$	$p < 0.001$
3 stars announced	95.2%	97.7%	89.5%	$p = 0.481$	$p = 0.006$
Average investment rate	56.1%	56.9%	57.5%	$p = 0.864$	$p = 0.480$
Average earnings of Investors	160.4	159.9	150.1	$p = 0.789$	$p = 0.059$

Notes: The table shows the fraction of times Investors invested in the Announcement Game, looking at the last nine periods in Necessary Condition treatment to allow participants to gain experience vs. all 18 periods in Public-Oath and Private-Oath.

Table D3: Investment rates in the Sufficient Condition, Public-Oath and No-Oath treatments (first 18 periods)

	Sufficient Cond. (1)	Public-Oath (2)	No-Oath (3)	*p-value* *(1) vs. (2)*	*(1) vs. (3)*
0 stars announced	0.0%	1.7%	9.5%	$p = 1.000$	$p = 0.749$
1 star announced	13.9%	16.8%	14.5%	$p = 0.036$	$p = 0.969$
2 stars announced	74.9%	85.8%	70.5%	$p = 0.021$	$p = 0.134$
3 stars announced	92.0%	97.7%	77.8%	$p = 0.001$	$p < 0.001$
Average investment rate	74.6%	56.9%	65.8%	$p < 0.001$	$p = 0.008$
Average earnings of Investors	156.2	159.9	149.7	$p = 0.268$	$p = 0.304$

Notes: The table shows the fraction of times Investors invested in the Announcement Game, looking at the first 18 periods.

Table D4: Investment rates in the Necessary Condition, Public-Oath, and Private-Oath treatments (first 18 periods)

	Necessary Cond. (1)	Public-Oath (2)	Private-Oath (3)	*p-value* *(1) vs. (2)*	*(1) vs. (3)*
0 stars announced	0.5%	1.7%	4.0%	$p = 0.922$	$p = 0.143$
1 star announced	14.2%	16.8%	13.0%	$p = 0.684$	$p = 0.156$
2 stars announced	82.2%	85.8%	72.6%	$p = 0.695$	$p = 0.003$
3 stars announced	92.1%	97.7%	89.5%	$p = 0.076$	$p = 0.135$
Average investment rate	54.8%	56.9%	57.5%	$p = 0.853$	$p = 0.289$
Average earnings of Investors	157.5	159.9	150.1	$p = 0.714$	$p = 0.110$

Notes: The table shows the fraction of times Investors invested in the Announcement Game, looking at the first 18 periods.

Table D5: Time trends of investment decision by treatment (all announcements)

Dependent variable	(1)	(2)	(3)	(4)	(5)	(6)
Investment decision	No-Oath	Public-Oath	Private-Oath	Private-Oath^{++}	Necessary Cond.	Sufficient Cond.
Period	-0.006*	-0.000	-0.001	-0.002	0.001	-0.007***
	(0.003)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
Constants	0.718***	0.574***	0.583***	0.645***	0.540***	0.818***
	(0.032)	(0.027)	(0.034)	(0.034)	(0.028)	(0.027)
Number of observations	1368	1386	1314	1350	2079	2106
Number of clusters	76	77	73	75	77	78

Notes: This table presents the coefficients from random-effects Linear Probability models (GLS) in which the dependent variable is the decision to invest. The only independent variable is the period. Standard errors are clustered at the individual level. * $p<0.05$; ** $p<0.01$; *** $p<0.001$.

E Online Appendix: Beliefs

We elicited the beliefs of participants at the end of the sessions. The following tables report descriptive statistics and non-parametric tests on these beliefs.

The Investors were asked to estimate the proportion of truthful announcements for each number of stars made to them in part 2 (*i.e.,* their first-order beliefs about Advisors' truthfulness). Belief data is notoriously noisy, and this portion of our data is no exception. We note that our belief data are nonetheless often qualitatively consistent with the actions that we observe among our subjects. For example, even though the point estimates do not match actual investment rates, we see in Table E1 that in the No-Oath, Private-Oath, and Sufficient Condition treatments Investors are more distrustful when they are faced with a higher announcement than when they are faced with a lower announcement (and indeed in these treatments Advisors are in large part not truthful). In contrast, in the Public-Oath and Necessary Condition treatments, which are the treatments in which Investors generally experience truthful announcements, the credibility of announcements is more stable across announcements than in the treatments in which the Advisors are not truthful. The data in Table E2 also shows that the cross-treatment comparisons often match the patterns we observe in actual investment rates, for example, with subjects trusting three-star announcements significantly more in Public-Oath than in No-Oath (first column), or that there are no significant differences in how much they trust a three-star announcement across the No-Oath and Sufficient Condition treatments (second last column).

The within-treatment data on the first-order beliefs of Advisors on the truthfulness of other Advisors' announcements are also very noisy (see Table E3). However, the cross-treatment comparisons of these beliefs align with what would be expected given behavior. For example, as Table E4 shows, beliefs on truthfulness levels of zero-star announcements are no different across treatments. In that same table, we also see that there are large and significant cross-treatment differences in beliefs about the truthfulness of three-star announcements: it is the highest in Public-Oath, followed by Private-Oath and then No-Oath.

Finally, the second-order beliefs of Advisors regarding what they think Investors believe the proportion of truthful announcements by number of stars is (see Tables E5 and E6) generally follow the same patterns as the first order beliefs of Advisors: for example, Advisors believe that Investors will believe a three-star announcement more in the Public-Oath than in the Private-Oath, and more in the Private-Oath than in the No-Oath treatment.

Table E1: First-order beliefs of Investors

Announcements	No Oath	Public-Oath	Private-Oath	Private-Oath^{++}	Necessary Cond.	Sufficient Cond.
0 Stars	81.00	86.68	81.52	44.79	79.25	77.84
1 Star	82.69	88.30	78.29	68.69	75.64	80.66
2 Stars	53.83	80.82	62.05	58.56	71.51	60.78
3 Stars	44.26	82.53	60.64	56.48	74.12	49.65

Note: This table reports Investor's first-order beliefs about the proportion of truthful announcements, by number of stars announced and by treatment.

Table E2: Between-treatment comparisons of first-order beliefs of Investors

	No-Oath *vs.* Public-Oath	No-Oath *vs.* Private-Oath	No-Oath *vs.* Necessary Cond.	No-Oath *vs.* Sufficient Cond.	Public-Oath *vs.* Private-Oath	Public-Oath *vs.* Necessary Cond.	Private-Oath *vs.* Necessary Cond.	Public-Oath *vs.* Sufficient Cond.	Private-Oath *vs.* Private-Oath^{++}
0 Stars	$p = 0.716$	$p = 0.684$	$p = 0.979$	$p = 0.998$	$p = 0.917$	$p = 0.614$	$p = 0.532$	$p = 0.773$	$p < 0.001$
1 Star	$p = 0.239$	$p = 0.040$	$p = 0.002$	$p = 0.187$	$p = 0.071$	$p = 0.011$	$p = 0.294$	$p = 0.730$	$p = 0.097$
2 Stars	$p < 0.001$	$p = 0.034$	$p < 0.001$	$p = 0.038$	$p < 0.001$	$p = 0.009$	$p = 0.015$	$p < 0.001$	$p = 0.441$
3 Stars	$p < 0.001$	$p = 0.002$	$p < 0.001$	$p = 0.175$	$p < 0.001$	$p = 0.023$	$p = 0.010$	$p < 0.001$	$p = 0.420$

Note: This table reports the p-values from rank-sum tests comparing the 1st-order beliefs of Players B across treatments.

Table E3: First-order beliefs of Advisors

Announcements	No-Oath	Public-Oath	Private-Oath	Private-Oath^{++}
0 Stars	63.62	75.90	69.77	67.21
1 Star	63.16	79.30	70.16	68.81
2 Stars	53.72	84.55	69.12	69.60
3 Stars	50.99	87.91	72.77	71.72

Note: This table displays the first-order beliefs of Players A about the proportion of truthful announcements by the other As, by the number of stars announced, and by treatment.

Table E4: Between-treatment comparisons of first-order beliefs of Advisors

	No-Oath *vs.* Public-Oath	No-Oath *vs.* Private-Oath	Public-Oath *vs.* Private-Oath	Private-Oath *vs.* Private-Oath^{++}
0 Stars	$p = 0.124$	$p = 0.281$	$p = 0.690$	$p = 0.864$
1 Star	$p = 0.011$	$p = 0.160$	$p = 0.314$	$p = 0.791$
2 Stars	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p = 0.999$
3 Stars	$p < 0.001$	$p < 0.001$	$p = 0.009$	$p = 0.783$

Notes: This table reports the p-values from rank-sum tests comparing the first-order beliefs of Players A across treatments.

Table E5: Second-order beliefs of Advisors

Announcements	No-Oath	Public-Oath	Private-Oath	Private-Oath^{++}
0 Stars	56.07	65.19	53.07	50.68
1 Star	56.36	68.21	55.45	54.52
2 Stars	57.32	79.71	59.14	66.21
3 Stars	48.88	88.16	67.71	68.36

Note: This table displays the second-order beliefs of Players A, that is, what they think Investor believes about the proportion of truthful announcements, by number of stars announced, and by treatment.

Table E6: Between-treatment comparisons of the second-order beliefs of Advisors

	No-Oath *vs.* Public-Oath	No-Oath *vs.* Private-Oath	Public-Oath *vs.* Private-Oath	Private-Oath *vs.* Private-Oath^{++}
0 Stars	$p = 0.226$	$p = 0.939$	$p = 0.162$	$p = 0.838$
1 Star	$p = 0.095$	$p = 0.784$	$p = 0.046$	$p = 0.944$
2 Stars	$p < 0.001$	$p = 0.603$	$p < 0.001$	$p = 0.065$
3 Stars	$p < 0.001$	$p = 0.001$	$p < 0.001$	$p = 0.822$

Note: This table reports the p-values from rank-sum tests comparing the second-order beliefs of Players A across treatments.

F Online Appendix: Dynamics

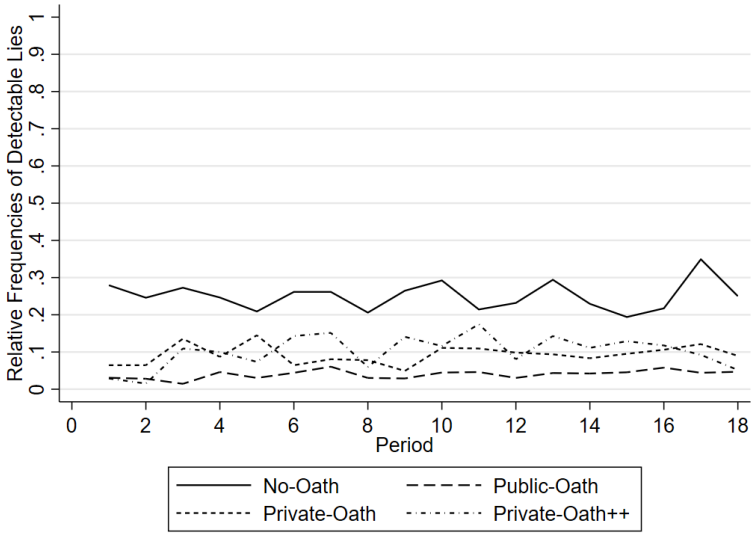

Figure F1: Evolution of the relative frequencies of detectable lies, by period and treatment

Note: The figure displays the evolution of the relative frequency of detectable lies over time, by treatment, including only the observations in which these lies were possible, that is, excluding the observations in which three stars were observed.

G Online Appendix: Survey on Prolific

We report in subsection G.1 the questionnaire used in the survey conducted in December 2022 via Prolific with 395 individuals residing in the Netherlands (198) and France (197). The survey was collected via Google Forms in Dutch and French respectively, and translated back to English. We analyze the responses in subsection G.2.

G.1 Questionnaire

Consent
This study includes a short survey conducted by Chloe Tergiman (Pennsylvania State University), Marie Claire Villeval (CNRS, University of Lyon) and Sorravich Kingsuwankul (University of Lyon). It takes about 4 minutes to complete the survey.
Participation is voluntary. You can withdraw from the survey at any time without giving a reason. This will not affect the relationship between you, the researchers, or Prolific in any way. All aspects of the study are confidential and anonymous. We will not ask for your name or any information that may identify you during the study. A report of this survey may be submitted for publication, but all information will be used in an aggregate form.
Upon completion of this survey, you will receive a fixed fee of 1€. If you have specific questions about the survey, you can contact us by e-mail at kingsuwankul[at]gate.cnrs.fr
If you agree with the above terms, please select "I agree" to continue with the survey.

- I agree *(Continue)*
- I do not agree *(Exit survey)*

Please enter your Prolific ID:
(Input box here)

Do you live in the Netherlands/France?

- Yes *(Continue)*
- No *(Exit survey)*

Q1. Could you say in general that most people can be trusted or that you have to be very careful when dealing with people?

- Most people can be trusted.
- You have to be very careful.

Q2. Please indicate on a 5-point scale from 1 "I don't trust at all" to 5 "I trust completely", how much you trust each of these groups:

- Doctors
- Members of the parliament
- Judges

- Firefighters
- Bankers

Q3A. *(If reported trust in bankers in Q2 is low: 1-3)* In your response to the previous question, you have indicated that **bankers cannot be completely trusted**. Can you explain the reason(s)? You can choose up to three reasons.

- Banking culture is not trustworthy.
- I/someone I know had a bad experience with bankers/banks.
- The banking industry attracts more dishonest employees than other sectors.
- Bankers' unethical behavior is difficult to detect.
- Unethical behavior by employees is usually not punished by the banks and can sometimes even be rewarded by bonuses.
- Bankers do not prioritize the customers' interests over those of the bank.

Is there another reason, not mentioned above, why you think bankers cannot be completely trusted? If not, you can skip this question.

Q3B. *(If reported trust in bankers in Q2 is high: 4 or 5)* In your response to the previous question, you have indicated that **bankers can generally be trusted**. Can you explain the reason(s)? You can choose up to three reasons.

- Scandals are rare and don't reflect the behavior of most bankers.
- I have had a good experience with bankers/banks so far.
- There are enough safeguards to protect clients.
- Unethical behavior by employees is usually punished by the banks.
- Bankers generally have to place the customers' interests over those of the bank.

Is there another reason, not mentioned above, why you think bankers can generally be trusted? If not, you can skip this question.

Q4. How often do you seek financial advice from your banker or financial advisor?

- I do not have a financial advisor
- Less than once a year
- Once a year
- A few times a year
- Every month or more often

Q5. Do you know whether bankers in the Netherlands/France are required to swear a professional oath of good conduct?

- I am sure that bankers do not have to swear an oath
- I am not sure, but it is unlikely that bankers have to swear an oath
- I am not sure, but it is likely that bankers have to swear an oath
- I am sure that bankers have to swear an oath

Q6. Do you think a bankers' oath should be compulsory in all European countries?

- Yes *(Go to Q7A.)*
- No *(Go to Q7B.)*

Q7A. You responded 'Yes' to the previous question. Why do you think a bankers' oath should be compulsory in all European countries?

- Because it disciplines the bankers
- Because it makes customers' interest more salient to the bankers
- For other reasons

You have responded 'For other reasons' to the previous question. Please explain why you think a bankers' oath should be compulsory in all European countries.

Q7B. You responded 'No' to the previous question. Why do you a banker's oath should not be made compulsory in all European countries?

- Because people can rely on their own experience to trust bankers or not
- Because an oath does not do anything against dishonest bankers
- For other reasons

You have responded 'For other reasons' to the previous question. Please explain why you think a bankers' oath should not be compulsory in all European countries.

Q8. What is your gender?

- Male
- Female
- Other

Q9. In what year were you born? (*e.g.* 1970)

Q10. What is the highest level of education you have obtained? If you are currently enrolled, please select the highest level of education you have attained.

- High school or below

- Bachelor
- Master or above

Q11. What is your current status?

- Student
- Employee
- Self-employed
- Retired
- Unemployed

Q12. What is your annual personal net income (your salary after income tax, social security contributions, and pension contributions have been deducted)?

- $<$ €15000
- €15000 - €29999
- €30000 - €44999
- €45000 - €59999
- €60000 and more

Q13. The next question is about the following problem. In questionnaires like ours, some participants sometimes do not read the questions carefully and just click quickly on the questionnaire. This means that many random answers jeopardize the results of surveys. To show that you have read our questions carefully, give "bankers" as your answer to the following question. What is the profession indicated?

- Firefighters
- Bankers
- Doctors
- Teachers
- Police officers

End of the survey

You have completed the survey. We thank you for your time. Important: Please complete the following 3 steps to register your survey response and receive your payment.

1. Write down this survey code: *(Survey completion code here)*
2. Click on 'Submit' on this page to register your responses. If you do not complete this step, we will not receive your information and cannot reward you.
3. Enter the survey code into your Prolific account to register your submission.

G.2 Results of the survey and discussion

Low trust in bankers Using the survey from the Netherlands as our main case, we find low average levels of trust when it comes to the banking sector: when asked to rate their trust on a scale of 1 ('I don't trust at all') to 5 ('I trust completely'), the average rating for bankers is 2.55 (which is barely higher than in the French sample - 2.40 -, although there is no such compulsory oath in France). Bankers rank at the bottom compared with members of parliament (who barely do better with an average of 2.61), judges (3.87), doctors (3.96), and firefighters (4.62). Our survey points to several reasons for this lack of trust: respondents believe bankers do not prioritize clients' interests, that the banking culture is not trustworthy, and that unethical behavior is hard to detect (and therefore to punish). All summary statistics are provided in Table G1.

Low awareness of the oath Is the Dutch bankers' oath then of no impact? Our survey shows that despite an oath being compulsory, the awareness of it is limited. In fact, only 53.5% of our respondents were either sure or thought it is likely that such an oath exists. Whether these responses are the result of knowledge of the actual oath or simply the result of the respondents thinking that bankers "obviously" have to swear an oath given their positions is uncertain. To be able to speak to this point, we conducted a similar survey in France, where such an oath does not exist. In that survey about 28% of respondents still believe it is the case. Extrapolating this to the Dutch survey hints that perhaps the answers of a non-negligible proportion of the 53% mentioned above do not stem from the "actual" knowledge of the oath. Responses to the other questions are much more in line with the surveys.

A correlation between oath awareness and trust in bankers The results of the survey complement our laboratory findings. While generally trust is low, there is, however, a positive link between being aware of the oath and trust in bankers. Respondents in the Netherlands who are aware of the oath hold a higher level of trust in bankers (Spearman's correlation coefficient $= 0.15$, $p = 0.035$). Our laboratory results show that being aware of the oath is not a necessary condition for the trust since participants in the role of Player B (who can be thought of as investors) learned from their experience. This element is also reflected in our survey, as about 60% of the respondents with high trust in bankers indicated that it was because of the good experiences they have had with their bankers and banks.

In all, there are two takeaways from the field survey. First, the awareness of bankers' oath is far from being universal, even when such a practice has been legally implemented country-wide and many discussions have been conducted both in the media and at the parliament. Second, the banking profession suffers from trust issues, but they may be alleviated as a large majority of people in the survey support the implementation of a bankers' oath. Indeed, Table G1 shows that 83% of the respondents are in favor of a compulsory bankers' oath in the European Union (this is also the case for 80% of the respondents in the French sample). 60% of those who are in favor of such an oath believe that it will help in making their interests more salient, and one-third think that it would discipline bankers. The field experiment of Weitzel and Kirchler (2023), using reminders of the oath, shows that these beliefs are in fact well grounded.

Table G1: Overview of responses from the Prolific survey in the Netherlands and France

Country	Netherlands N=198	France N=197
Trust attitudes and level of awareness		
Most people can be trusted	69%	45%
Mean trust in different groups (St.Dev.)		
Doctors	3.96 (0.77)	3.96 (0.75)
Members of the parliament	2.61 (0.87)	2.22 (0.81)
Judges	3.87 (0.83)	3.37 (0.95)
Firefighters	4.62 (0.56)	4.46 (0.68)
Bankers	2.55 (0.96)	2.40 (0.95)
Three most selected reasons for ...		
Low trust in bankers	n=162	n=174
Bankers do not prioritize customers' interest	78%	84%
Banking culture is not trustworthy	60%	65%
Unethical behaviors are hard to detect	49%	40%
High trust in bankers	n=36	n=23
I have had good experiences	64%	83%
There are enough safeguards	67%	39%
Scandals are rare	53%	61%
Awareness		
I am sure that bankers do not take an oath	2.53%	20.81%
I am not sure but it is unlikely that bankers take an oath	43.94%	51.27%
I am not sure but bankers likely take an oath	36.36%	22.84%
I am sure that bankers take an oath	17.17%	5.08%
Spearman's correlation between awareness and trust in bankers	0.15 ($p = 0.036$)	0.02 ($p = 0.782$)
Should bankers' oath should be compulsory in EU?		
Yes because ...	n=165	n=158
It disciplines bankers	34%	37%
It makes customers' interest salient	60%	57%
Other reasons	6%	6%
No because ...	n=33	n=39
People can rely on experience	3%	13%
Oath does not do anything	94%	79%
Other reasons	3%	8%
Socio-economic characteristics		
Male	49%	52%
Mean age (St.Dev.)	36 (9.2)	37 (10.8)
Education		
High school or below	12.63%	1.02%
Bachelor	49.49%	10.66%
Master or above	37.88%	88.32%
Status		
Student	7.58%	11.17%
Employee	76.77%	55.33%
Self-employed	8.59%	22.34%
Retired	0.51%	6.60%
Unemployed	6.57%	4.57%
Frequency of financial advice		
No financial advisor	59.60%	25.38%
Less than once a year	23.74%	50.76%
Once a year	11.62%	18.27%
A few times a year	5.05%	5.58%
Every month or more often	-	-
Annual personal net income		
< €15000	17.17%	27.92%
€15000 - €29999	22.73%	36.55%
€30000 - €44999	33.33%	21.32%
€45000 - €59999	17.68%	9.14%
€60000 or more	9.09%	5.08%

Notes: The table reports the descriptive statistics of the responses to the Prolific survey conducted with residents in the Netherlands and France. For awareness of the oath, it is coded as 1 if the respondent indicated that he or she is sure or it is likely that Dutch (or French) bankers take an oath of conduct.

References for Online Appendix

References

Akin, Z. (2022). Playing the victim behavior: An experimental study. Technical report, American University in Dubai.

Babin, J. J. and H. S. Chauhan (2023). Show no quarter: combating plausible lies with ex-ante honesty oaths. *Journal of the Economic Science Association 9*, 66–76.

Babin, J. J., H. S. Chauhan, and F. Liu (2022). You Can't Hide Your Lying Eyes: Honesty Oaths and Misrepresentation. *Journal of Behavioral and Experimental Economics 98*, 101880.

Beck, T. (2021). How the honesty oath works: Quick, intuitive truth telling under oath. *Journal of Behavioral and Experimental Economics 94*, 101728.

Beck, T., C. Buhren, B. Frank, and E. Khachatryan (2020). Can honesty oaths, peer interaction, or monitoring mitigate lying? *Journal of Business Ethics 163*(3), 467–484.

Cagala, T., U. Glogowsky, and J. Rincke (2024). Detecting and Preventing Cheating in Exams: Evidence from a Field Experiment. *Journal of Human Resources 59*(1), 210–241.

Cagala, T., U. Glogowsky, J. Rincke, and S. Schudy (2024). Commitment Requests Do Not Affect Truth-Telling in Laboratory and Online Experiments. *Games and Economic Behavior 143*, 179–190.

Carlsson, F., M. Kataria, A. Krupnick, E. Lampi, Löfgren, P. Qin, T. Sterner, and S. Chung (2013). The truth, the whole truth, and nothing but the truth–a multiple country test of an oath script. *Journal of Economic Behavior & Organization 89*, 105–121.

Davis, B. J. and T. Jaber-Lopez (2022). Do voluntary commitment mechanisms improve welfare? The effect of mandatory and voluntary oaths in a social dilemma. *Bulletin of Economic Research 75*(2), 525–540.

Heinicke, F., S. Rosenkranz, and U. Weitzel (2019). The effect of pledges on the distribution of lying behavior: An online experiment. *Journal of Economic Psychology 73*, 136–151.

Hergueux, J., N. Jacquemet, S. Luchini, and J. F. Shogren (2022). Leveraging the Honor Code: Public Goods Contributions under Oath. *Environmental and Resource Economics 81*(3), 591–616.

Jacquemet, N., A. James, S. Luchini, J. Murphy, and J. Shogren (2021). Do truth-telling oaths improve honesty in crowd-working? *PLoS ONE 16*.

Jacquemet, N., A. James, S. Luchini, and J. F. Shogren (2017). Referenda Under Oath. *Environmental and Resource Economics 67*(3), 479–504.

Jacquemet, N., R.-V. Joule, S. Luchini, and J. F. Shogren (2013). Preference elicitation under oath. *Journal of Environmental Economics and Management 65*, 110–132.

Jacquemet, N., S. Luchini, A. Malézieux, and J. F. Shogren (2020). Who'll stop lying under oath? Empirical evidence from tax evasion games. *European Economic Review 124*, 103369.

Jacquemet, N., S. Luchini, J. Rosaz, and J. F. Shogren (2019). Truth Telling Under Oath. *Management Science 65*(1), 426–438.

Jacquemet, N., S. Luchini, J. Rosaz, and J. F. Shogren (2021). Can we commit future managers to honesty? *Frontiers in Psychology*, 2785.

Jacquemet, N., S. Luchini, J. F. Shogren, and A. Zylbersztejn (2018). Coordination with communication under oath. *Experimental Economics 21*(3), 627–649.

Jacquemet, N., S. Luchini, J. F. Shogren, and A. Zylbersztejn (2024). Commitment to the truth creates trust in market exchange: Experimental evidence. *Games and Economic Behavior 148*, 279–295.

Kemper, N., J. Popp, and R. Nayga (2020). A query theory account of a discrete choice experiment under oath. *European Review of Agricultural Economics 47*(3), 1133–1172.

Koessler, A.-K. (2022). Pledges and how social influence shapes their effectiveness. *Journal of Behavioral and Experimental Economics 98*, 101848.

Koessler, A.-K., L. Page, and U. Dulleck (2021). Public cooperation statements. *Journal of Economic Interaction and Coordination 16*(4), 747–767.

Koessler, A.-K., B. Torgler, L. P. Feld, and B. S. Frey (2019). Commitment to pay taxes: Results from field and laboratory experiments. *European Economic Review 115*, 78–98.

Krüger, M. P. (2016). Can an oath reduce dishonesty and overconfidence for bankers? Technical report, Mimeo.

Peer, E. and Y. Feldman (2021). Honesty pledges for the behaviorally-based regulation of dishonesty. *Journal of European Public Policy*, 1–21.

Prima, R. A., S. Feeny, R. Hoffmann, and E. Satriawan (2020). Threats, Pledges, and Asset Misreporting: A Framed Field Experiment in Indonesia. *TNP2K Working Paper 57-e/2020*.

Schild, C., D. W. Heck, K. A. Ścigała, and I. Zettler (2019). Revisiting REVISE: (Re)Testing unique and combined effects of REminding, VIsibility, and SElf-engagement manipulations on cheating behavior. *Journal of Economic Psychology 75*, 102161.

Weitzel, U. and M. Kirchler (2023). The Banker's Oath And Financial Advice. *Journal of Banking & Finance 148*, 106750.

Zickfeld, J. H., K. A. Ścigała, C. T. Elbæk, J. Michael, M. H. Tønnesen, G. Levy, S. Ayal, I. Thielmann, L. Nockur, E. Peer, V. Capraro, R. Barkan, S. Bø, Bahník, D. Nosenzo, R. Hertwig, N. Mazar, A. Weiss, A.-K. Koessler, R. Montal-Rosenberg, S. Hafenbrädl, Y. A. Nielsen, P. Kanngiesser, S. Schindler, P. Gerlach, N. Köbis, N. Jacquemet, M. Vranka, D. Ariely, J. B. Martuza, Y. Feldman, M. Białek, J. K. Woike, Z. Rahwan, A. Seidl, E. Chou, A. Kajackaite, S. Schudy, U. Glogowsky, A. Z. Czarna, S. Pfattheicher, and P. Mitkidis (2025). Effectiveness of ex ante honesty oaths in reducing dishonesty depends on content. *Nature Human Behaviour 9*(1), 169–187.