Belief Elicitation as Behavioral Design: Forecasting Others' Actions Reduces Moral Wiggle Room

Tony Hua

August 2025

Abstract

This pre-registered study tests whether prompting individuals to forecast others' behavior reduces self-serving choices in ethically relevant decisions. Across two domains—strategic ignorance in a dictator game and dishonesty in a private die-roll task—I examine whether belief elicitation activates internalized social norms and reduces moral disengagement. The design varies the timing and content of belief prompts to identify their causal impact, while controlling for confounds such as interface changes or norm framing. Results will speak to the psychological mechanisms through which norms shape behavior and evaluate whether belief elicitation functions not only as a measurement tool but also as a behavioral intervention.

Keywords: information avoidance, moral wiggle-room, social norms, social appropriateness, dishonesty, experiment

JEL Codes: C72, C91, D8, D9

Introduction

People often avoid morally relevant information to maintain plausible deniability or reduce discomfort when making self-serving decisions. This phenomenon, known as strategic ignorance, is well-documented in contexts such as allocation games and private dishonesty tasks (Dana et al., 2007; Grossman, 2014; Fischbacher and Föllmi-Heusi, 2013).

While social norms influence ethical behavior, direct norm messaging can backfire—especially when it is perceived as judgmental or coercive (Schultz et al., 2007; Bicchieri et al., 2023). A subtle intervention is to elicit individuals' predictions about others' behavior. This form of belief elicitation may activate internalized expectations or make the moral dimension of a decision more salient without overt persuasion. Such a mechanism aligns with research on second-order beliefs (Charness and Dufwenberg, 2006) and norm learning through repeated exposure (Peysakhovich and Rand, 2016).

I test whether prompting individuals to forecast others' behavior reduces self-serving choices in two domains—strategic ignorance and dishonesty—by isolating the timing and content of belief prompts to causally identify the mechanism of norm activation. To evaluate this mechanism, I apply it in two well-established paradigms where moral disengagement is common: (1) a moral wiggle-room game involving strategic ignorance (Dana et al., 2007), and (2) a private die-roll task where dishonesty yields financial gain (Fischbacher and Föllmi-Heusi, 2013). If effective, belief elicitation could serve as a scalable, low-cost behavioral tool for promoting transparency and ethical decision-making across domains where self-serving decisions are sensitive to normative framing, such as charitable giving, environmental compliance, and organizational integrity.

Contributions

This study makes three key contributions:

1. **Causal Identification**: It isolates belief elicitation on norms as a causal mechanism influencing moral behavior, removing confounds related to interface design or norm feedback present in Hua (2025).

2. **Cross-Domain Generalization**: By applying the same experimental structure to both a dictator game and a private die-roll task, the study evaluates whether belief elicitation taps into a generalizable norm activation mechanism that governs both prosociality and honesty.

3. **Post Hoc Rationalization**: The design includes measures of social appropriateness and ex post beliefs to examine whether individuals selectively update their normative perceptions to justify morally questionable actions. This offers insight into how social norms evolve and stabilize over time through cognitive self-justification.

Experimental Design

The experiment employs a between-subjects design with three treatment conditions to isolate the causal effect of belief elicitation on information avoidance. All participants complete a modified version of the moral wiggle-room game (Dana et al., 2007), in which they decide whether to reveal the payoff to a passive recipient before making an allocation decision by choosing between two options. A mock-up of the moral wiggle-room game is available in Appendix A.1. In the *Pre-Belief* condition, participants first estimate the percentage of individuals in a prior session who chose to reveal the payoff before making their own decision. In the *Post-Belief* condition, the same belief elicitation occurs only after the allocation decision has been made. In the *Placebo* condition, participants are instead asked to estimate an unrelated factual statistic (e.g., daily coffee consumption rates) before proceeding to the allocation task. The experimental interface, instructions, and choice architecture are held constant across all conditions; only the content and timing of the belief prompt vary. This design allows for clean identification of whether belief elicitation on information seeking norms alone reduces strategic ignorance.

The second experiment uses a die-reporting task based on Fischbacher and Föllmi-Heusi (2013). Participants are instructed to privately roll a six-sided die using a physical die, smartphone app, or web-based dice roller. They then report the result, with payment tied to the reported number (e.g., 1 to 6 experimental credits). Because the roll is unobserved by the experimenter, participants may choose to misreport in order to increase their earnings. After making this initial roll, participants will be asked to roll the die 6 times and report all 6 values. One of the reported rolls will be randomly selected for payment. Thus, participants will be paid for a total of two dice rolls.

Participants are randomly assigned to one of six between-subjects conditions in a 3 by 2 design, varying belief elicitation of others' behaviors, i.e. beliefs elicited before, after, or placebo belief, and the framing of the norm, i.e. are people misreporting versus are people truthfully reporting. In the *Pre-Belief* condition, participants are asked to estimate the percentage of others who misreported the value when submitting their own report. In the *Post-Belief* condition, this same belief elicitation occurs only after they have reported their die outcome. In the *Placebo* condition, participants are asked to estimate an unrelated factual statistic (such as daily coffee consumption) prior to reporting their roll.

Amendment 8.7.2025

The *Placebo* condition, in which participants estimated an unrelated factual statistic prior to their decision, will be dropped from the design. This change is based on pilot results indicating limited additional interpretive value and is intended to conserve budget for the main treatment arms.

This design allows us to test whether belief elicitation influences dishonest reporting in private, self-serving contexts, thereby generalizing the hypothesized mechanism of norm activation across distinct behavioral domains. It also tests to see if the framing of norms as being either antisocial (misreporting) or prosocial (truthful) can influence behavior.

To explore whether individuals engage in post hoc rationalization of their decisions, all participants will complete a social appropriateness rating questionnaire following the allocation task. This questionnaire, adapted from the method developed by Krupka and Weber (2013), asks participants to evaluate both their personal view and their perception of others' views regarding the appropriateness of revealing or not revealing the payoff information or whether or not to report the appropriate value from the dice roll. While Hua (2025) found no evidence that individuals form self-justifying beliefs in advance to exploit moral wiggle room, it remains an open question whether such justifications arise after the decision has been made, particularly as a way to align internal or social narratives with one's behavior. Measuring these ex post beliefs provides insight into whether individuals selectively revise their perceptions of social norms to preserve a positive self-image. Such dynamics may play a critical role in how descriptive norms stabilize over time, especially if repeated acts of post-decision justification contribute to the perceived legitimacy of strategic ignorance.

All participants will complete a short demographics questionnaire at the end of the study. This includes items on age, gender, education, and political affiliation, which may be used in exploratory analyses to examine heterogeneity in responsiveness to belief elicitation or norm salience. These variables are not central to the core hypotheses but may help identify subgroup-specific patterns relevant to future interventions.

After completing the survey questionnaire, subjects complete the social value orientation (SVO) task (Murphy et al., 2011), choosing a allocations between themselves and a third participant for 6 different payment scales. Subjects are told that one of the six allocations will be realized.

Pre-Trial Data Collection

Three pre-trial sessions were conducted on Prolific between June and July 2025.

- **Trial 1** involved 46 participants in an incentivized die-roll experiment using a virtual die embedded in the experimental interface. Participants reported the outcome of their roll, which determined their payout.

- **Trial 2** included 54 participants in a similarly incentivized die-roll task. However, participants were instructed to use either a physical die or an external application to generate the result.

- **Trial 3** introduced 120 participants to the moral wiggle-room game who were told that they would predict the behavior of other participants in a different study. Without playing the game, subjects were asked to predict the proportion of others who chose to reveal the recipient's payoffs.

Trials 1 and 2 successfully replicated the findings of Fischbacher and Föllmi-Heusi (2013), with only 2 out of 46 participants appearing to misreport their roll in Trial 1 and a significantly skewed distribution of reported outcomes in Trial 2. These results confirm that participants responded to the incentive structures in ways consistent with prior research, validating the experimental setup for the main study.

Given that participants appeared highly sensitive to perceived observability by the experimenter, monetary incentives for belief elicitation were removed. This design choice aimed to avoid inadvertently signaling that participants' honesty might be monitored—particularly in the die-roll task, where even subtle cues can influence reporting behavior. Furthermore, an incentivized belief elicitation would require estimating dice rolling behavior in an environment in which such dice rolls can be observed, which would be fundamentally different from the task participants would be asked to do in the main study. Instead, the belief elicitation prompt will be framed as a neutral, non-incentivized question to encourage reflection on social norms rather than concern over detection. This task would minimize potential deception and simplify the experimental design.

To ensure that belief reports remained meaningful in the absence of incentives, Trial 3 compared the distribution of elicited beliefs to that observed in Hua (2025). A Kolmogorov–Smirnov test indicated no significant difference between the two distributions ($p = 0.9$), suggesting that the removal of belief-based incentives did not compromise the validity of elicited responses.

Hypotheses and Proposed Analysis

Amendment 8.7.2025

Following the pilot (N = 175), the hypothesis structure has been revised to integrate Social Value Orientation (SVO) as a key moderator variable in the moral wiggle-room game. The new hypotheses (H4A) explicitly test whether SVO types (Prosocial vs. Individualist/Competitive) moderate the effect of belief elicitation and information conditions (*Pre-Belief*, *Post-Belief*, *Self/Self*) on the decision to remain ignorant.

The numbering of hypotheses has been updated throughout to reflect the insertion of these SVO-specific tests. The naming convention for hypotheses has been updated to clearly separate those relating to the moral wiggle-room game from those relating to the dice-rolling game. The SVO task (Murphy et al., 2011) was already included in the original design for exploratory purposes; this amendment elevates its role to a planned moderator in confirmatory analyses for the moral wiggle-room game. This change ensures that heterogeneity by prosocial type is tested explicitly and aligns the analysis plan with emerging theoretical critiques and pilot findings.

0.1 Moral Wiggle Room Game Hypothesis

Hypothesis 1A: Social Activation Participants who are asked to estimate others' behavior before making their own decision (*Pre-Belief* condition) will be significantly more likely to reveal the hidden payoff information than participants in the *Post-Belief* and *Placebo* conditions. This would support the hypothesis that belief elicitation causally reduces strategic ignorance.

I will estimate the proportion of subjects who choose to reveal across treatments using logistic regression with treatment dummies. The *Pre-Belief* condition will be compared against the *Post-Belief* and *Placebo* conditions. Robustness checks will include demographic controls.

Hypothesis 2A: Placebo Participants in the *Post-Belief* condition will not differ significantly in reveal rates from those in the *Placebo* condition, suggesting that belief elicitation only affects behavior when it occurs before the decision is made.

This will be made implicit with the logistic regression with treatment dummies.

Hypothesis 3A: Self Justifying Evaluation Participants who choose not to reveal the hidden payoff information will rate non-revealing behavior as more socially appropriate—both personally and as perceived by others—than those who choose to reveal. This would suggest post hoc rationalization to justify strategic ignorance.

I will compare social appropriateness ratings (both personal and perceived others') for non-revealing behavior between participants who chose to reveal and those who did not, using two-sample t-tests and a robustness check with a linear regressions with reveal decision as the independent variable and demographic controls.

Hypothesis 4B: Social Justification If participants in the *Post-Belief* condition are more likely to reveal than in *Pre-Belief* and *Placebo*, then those who remained ignorant in the latter treatments are more likely to believe others also did not reveal, consistent with post hoc rationalization.

Conditional on a significant difference in ignorance rates, I will examine whether participants in the *Post-Belief* condition who remained ignorant estimate lower rates of revealing by others than those. I will compare distributions using a Kolmogorov–Smirnov test.

Hypothesis 5A: SVO Moderation Ignorance rates will vary across treatments (*Self/Self*, *Pre-Belief*, *Post-Belief*) as a function of Social Value Orientation (SVO) type. Specifically, we expect SVO to moderate the treatment effect such that:

1. Low-SVO participants (Individualist/Competitive) are more likely to avoid information in the social setting (*Pre-Belief* and *Post-Belief*) than in *Self/Self*, reflecting image-related avoidance.

2. High-SVO participants (Prosocial) will show little or no increase in ignorance in the social setting relative to *Self/Self*, as their preferred choice aligns with the fair option.

3. Belief elicitation (*Post-Belief*) will reduce image-related avoidance in the social setting, with this reduction most evident among Prosocial participants as classified by SVO ratings.

0.2 Dice Rolling Game Hypotheses

Hypothesis 1B: Social Activation Participants in the *Pre-Belief* condition will report significantly fewer high-value outcomes (e.g., 5 or 6s) than those in the *Post-Belief* or *Placebo* conditions, consistent with reduced dishonesty due to norm activation.

I will compare the distribution of reported die-roll outcomes across treatment conditions using chi-squared goodness-of-fit tests to assess deviations from a uniform distribution, and conduct pairwise comparisons between the *Pre-Belief* condition and both the

Post-Belief and *Placebo* conditions to test whether belief elicitation prior to reporting reduces the frequency of high-value outcomes (e.g., 6s).

Hypothesis 2B: Placebo Participants in the *Post-Belief* and *Placebo* conditions will not differ significantly in their reported outcomes, indicating that belief elicitation must precede the decision to affect behavior.

I will compare the distribution of reported die-roll outcomes between the *Post-Belief* and *Placebo* conditions using chi-squared goodness-of-fit tests and two-sample proportion tests to determine whether belief elicitation affects behavior only when it occurs before the reporting decision; no significant difference between these two conditions would support the hypothesis.

Hypothesis 3B: Self Justifying Evaluation Participants who report a higher-than-average die roll (e.g., a 5 or 6) will rate misreporting behavior as more socially appropriate—both personally and as perceived by others—than those who report a lower roll. This would suggest post hoc rationalization to justify dishonesty.

I will assess whether participants who report higher die-roll outcomes (e.g., a 5 or 6) rate misreporting behavior as more socially appropriate—both personally and as perceived by others—using two-sample t-tests comparing high vs. low reporters, as well as linear regressions with reported die value as a continuous independent variable, controlling for treatment condition and demographics.

Hypothesis 4B: Social Justification If participants in the *Post-Belief* condition report more high-value outcomes (e.g., 5 or 6s) than in *Pre-Belief* and *Placebo*, then those who report high rolls will also estimate a higher rate of misreporting by others, consistent with post hoc rationalization.

Conditional on a significant difference in misreports, I will examine whether participants in the *Post-Belief* condition who report high die-roll outcomes (e.g., a 5 or 6) estimate higher rates of misreporting by others than those who report lower outcomes, using linear regression with reported die value as the independent variable and elicited belief about others' misreporting as the dependent variable, controlling for demographics.

This hypothesis can be tested across both norm conditions: truthful reporting and misreporting.

Hypothesis 5B: Norm Recognition Moderation Participants exposed to truthful reporting as the norm (i.e., asked to estimate the percentage who report honestly) will report fewer high-value outcomes than those exposed to dishonest reporting as the norm (i.e., asked to estimate the percentage who misreport)

I will compare the distribution of reported die-roll outcomes between participants exposed to truthful reporting norms and those exposed to misreporting norms using chi-squared tests and two-sample proportion tests, as well as regressions with reported die outcome as the dependent variable and norm framing (truthful vs. misreporting) as the key independent variable, controlling for belief elicitation timing and demographics.

Procedures and Power Analysis

Subjects will be recruited using the Prolific recruitment platform, and the experimental interface will be programmed using the LIONESS web platform (Giamattei et al., 2020). Subjects will remain anonymous, and their Prolific ID will be scrubbed after data retrieval to ensure anonymity. All treatment arms will be fielded simultaneously with random assignment to treatment at the point of entry to minimize selection effects and ensure balanced samples across conditions.

Moral Wiggle Room Game

In Hua (2025), an exploratory comparison across separate studies suggested that belief elicitation may reduce information avoidance by as much as 30 percentage points (from 62% to 32%). Because this estimate relies on between-study variation, the present design adopts a more conservative assumption of a 20 percentage point treatment effect (e.g., 60% vs. 40% ignorance). A power analysis for a two-sided test with $\alpha = 0.05$ and power = 0.80 indicates that detecting this effect requires approximately 97 participants per group. At a power of 0.90, this raises participants to 125 per arm. A total of 3 treatment arms between experiments approximated to a target of about 100 to 125 participants per arm (300 to 375 total), depending on the power.

However, exploratory results from Hua (2025) suggest a larger effect size (roughly 30 percentage points), which would allow for reliable detection with smaller samples (e.g., 50 per condition). To allow flexibility in planning, piloting, and budgetary constraints, I specify a recruitment range of 150 to 375 participants total. If strong evidence of treatment effects emerges before the upper bound is reached, data collection may be terminated early.

Amendment 8.7.2025

After initial data collection (N = 175) under the 60/10 vs. 50/50 payoff structure, ignorance rates were 45% in the *Pre-Belief* condition and 53% in the *Post-Belief* condition. Exploratory analysis indicated that differences in behavior were driven primarily by Prosocial types as classified by SVO ratings. Among these Prosocial participants, ignorance rates were 33% (*Pre-Belief*) and 21% (*Post-Belief*), suggesting that the effect of the belief-elicitation treatment may be constrained by an upper bound in the current payoff environment.

To test the full extent of the treatment's effects, I amend the design as follows. The payoff structure for the moral wiggle room game will change to: dictators choose between (i) 65 or 50 credits for themselves and 0 or 50 credits for the recipient. The higher payoff for the dictator and lower payoff for the recipient are intended to increase the proportion of Prosocial types who may exploit ignorance. Two additional conditions will be added:

1. A **Full-Information** condition in which the state is revealed before choice.
2. A **Self/Self** condition Exley and Kessler (2023) to help disentangle social-image and self-image effects.

To concentrate statistical power on the primary contrast of interest (*Pre-Belief* vs. *Post-Belief*), the *Placebo* condition from the original design will be dropped. The initial analysis indicated that there was virtually no difference between the *Placebo* and *Pre-Belief* treatment. The new design will therefore include three arms: *Pre-Belief*, *Post-Belief*, and *Full-Information*, plus the additional *Self/Self* condition for mechanism testing.

Given the observed effect sizes in the pilot and the aim of detecting a minimum detectable effect of 15 percentage points (e.g., 0.65 vs. 0.50 reveal rates) with $\alpha = 0.05$, two-tailed, and power = 0.80, the required sample size for the main confirmatory analysis (whole sample) is approximately 169 participants per arm. Accordingly, the primary recruitment target will be at least 169 participants per main treatment arm, with the possibility of collecting up to 180 per arm depending on recruitment feasibility and budget constraints.

Pilot data indicate that approximately 60% of participants are classified as Prosocial by the SVO measure, and exploratory results suggest that the treatment gap may be concentrated in this subgroup. A fully powered analysis of the Prosocial subgroup alone would require approximately:

$$N_{\text{per arm}} = \frac{169}{0.60} \approx 282$$

participants per arm. While subgroup analyses will be treated as exploratory under the primary recruitment target, if funding permits, recruitment will be extended toward the higher target to allow adequate power for confirmatory tests within the Prosocial subgroup.

The *Full-Information* and *Self/Self* conditions are included as secondary diagnostic treatments to help interpret the main treatment effect, rather than to serve as primary confirmatory contrasts. As such, these arms will be recruited at a smaller size, targeting approximately 60-100 participants per arm. This sample size is expected to be sufficient for obtaining stable descriptive estimates and for exploratory hypothesis testing, but will not provide the same statistical power as the main *Pre-Belief* vs. *Post-Belief* comparison. The reduced sample allocation allows for concentrating statistical power and resources on the primary research question while still enabling meaningful comparison to these diagnostic benchmarks.

The original 60/10 vs. 50/50 data will be retained for exploratory purposes or robustness checks in the appendix. All confirmatory analyses will be conducted on data collected after this amendment.

Dice Rolling Game

In the dice rolling game, the goal is to test whether the distribution of reported outcomes in each group differs from the uniform distribution, and whether treatment affects that distribution. Based on Fischbacher and Föllmi-Heusi (2013) who recruited 265 subjects, a Chi-square test with a medium effect size between 0.20 to 0.25 requires a minimum sample size of 143 to 205 participants per treatment arm. Thus, 200 subjects per treatment arm is a reasonable upper bound. With the addition of a truthful reporting versus misreporting framing, this translates into a 3 by 2 design. With 6 treatment arms containing 150-200 participants each, the experiment will total 900-1,200 participants. As with the moral wiggle room game, data collection may be terminated early if the effect sizes are larger than anticipated.

Amendment 8.13.2025

To maintain consistency with the moral wiggle room game, the framing has been changed from “truthfully report” to “accurately report” and from “misreport” to “inaccurately report.” This avoids any overt moral imposition. Furthermore, given that the *Pre-Belief* condition effectively serves a neutral framing condition, the *Placebo* conditions will be dropped. With this amendment, we intend to recruit between 150 to 225 participants for each of the 4 treatment arms, totaling between 600 to 900 subjects.

Amendment 8.20.2025

Amendment 8.20.2025 — Binary Dictator Game (CIG Only)

Rationale

The main MWRG examines belief elicitation when information avoidance is possible. To test whether the effect of belief elicitation generalizes to clear-cut fairness tradeoffs without ambiguity, I add a binary dictator game with the CIG payoff structure. This isolates whether belief elicitation influences prosocial choice when all information is transparent. Although belief elicitation is unincentivized, subjects will be predicting the behavior of the *Full Info* condition from the moral wiggle room game from the first experiment.

Design

- **Task:** One-shot binary dictator choice with full information and no reveal option.
- **Payoffs (CIG):**
  - Option A: Dictator 65, Recipient 0
  - Option B: Dictator 50, Recipient 50
- **Belief timing (between-subjects):**
  - Belief First (BF): Forecast elicited before choice.
  - Belief After (BA): Forecast elicited after choice.
- **Belief framing (between-subjects within BF/BA):** Randomize to one of two forecast frames, paralleling the DRG:
  - Frame A: "What proportion of participants in another study will choose Option A (65/0)?"
  - Frame B: "What proportion of participants in another study will choose Option B (50/50)?"
- **Interface:** Identical layout to the full-information MWRG; the only change is the absence of reveal mechanics.

Additional Measures

- **Appropriateness ratings:** After choice, participants rate how appropriate they find Option A and Option B (mirroring measures used in MWRG and DRG).

- **Social Value Orientation (SVO):** Standard SVO slider measure.
- **Questionnaire:** Demographics, political orientation, moral foundations, and other items as in the main preregistration.

Primary Outcome

Prosocial choice: share choosing Option B (50/50).

Hypotheses (Confirmatory; competing scope conditions)

- **H1a (Generalization):** BF > BA in prosocial choice. If true, belief elicitation activates fairness norms even without information avoidance.
- **H1b (Boundary condition):** BF = BA (within a pre-registered equivalence margin). If true, effects of belief elicitation are specific to contexts with moral wiggle room.
- **H2 (Belief rationalization):** In BA, selfish choosers predict lower prosociality among others than prosocial choosers.

Sample Size & Power

- **Baseline:** Prior full-information data: ∼60% prosocial.
- **Target effect (superiority):** Detect a 10 pp increase (0.60 → 0.70).
- **Power (two-sided, $\alpha = .05$):**
  - 80% power ≈ 356 per timing arm (≈ 712 total; ≈ 178 per frame cell).
  - With 150 per frame cell (≈ 600 total), power is ≈ 73% for a 10 pp effect.
- **Planned maximum sample:** $N_{\max} = 720$ total (180 per frame cell), powering the pooled BF vs. BA test.

Stopping rules (superiority *and* equivalence). All tests for superiority use two-sided $\alpha = .05$ on the pooled BF vs. BA contrast. Equivalence uses a two one sided tests with margin $\delta = 5$ percentage points (pp) on the BF–BA difference and a 90% CI. (*Rationale:* $\delta = 5$ pp is half the preregistered superiority target of 10 pp.)

1. **Interim 1** at $N = 320$ total (80 per frame cell; 160 per timing arm).
   Stop for efficacy (superiority): If $\widehat{\Delta}$_BF-BA ≥ 15 pp and two-sample proportion test $p < .05$.
   Stop for equivalence (precise null): If the **90% CI** for Δ lies entirely within $[-7\%, +7\%]$ **and** conditional power to detect a 10 pp effect at N_max is $< 20\%$.
2. **Interim 2** at $N = 480$ total (120 per frame cell; 240 per timing arm).
   Stop for efficacy (superiority): If $\widehat{\Delta}$_BF-BA ≥ 12 pp and $p < .05$.
   Stop for equivalence (precise null): If the **90% CI** for Δ lies entirely within $[-6\%, +6\%]$ **and** conditional power for a 10 pp effect at N_max is $< 20\%$.
3. **Final** at N_max $= 720$ total (180 per frame cell; 360 per timing arm).
   Superiority: Test 10 pp target (two-sided $\alpha = .05$).
   Equivalence (primary if superiority not shown): Declare equivalence if the **90% CI** for Δ lies within $[-5\%, +5\%]$.

Randomization & Implementation

- Step 1: Random assignment to BF vs. BA (50/50).
- Step 2: Independent randomization to Frame A vs. Frame B (50/50).
- Wording: Matches DRG style (frame-specific phrasing). Belief accuracy not incentivized.

Exclusions & Attention

Same preregistered exclusion criteria as the main study (failed comprehension checks, duplicate IDs, extreme outlier completion times).

Analysis Plan

- **Prosocial choice (primary):** Logit of choice on BF vs. BA, pooled across frames; report marginal effects and 95% CIs.
- **Equivalence (primary if superiority not shown):** two one sided tests for Δ with $\delta = 5$ pp using the 90% CI; also report the 95% CI.

- **Belief rationalization (H2):** In BA, regress belief forecasts on own choice (selfish vs. prosocial); compare distributions via K–S.
- **Exploratory:** Frame-specific BF–BA contrasts; SVO moderation; appropriateness ratings.

Conclusion

This study aims to causally identify whether belief elicitation on how others choose to acquire information alone can reduce individuals' willingness to avoid morally relevant information. By isolating the timing and content of belief prompts in a modified moral wiggle-room game, the design offers a clean test of whether prompting individuals to forecast others' behavior activates internalized norms or alters cognitive framing in ethically significant decisions. In addition to examining behavior, the study explores whether individuals engage in post hoc rationalization by selectively adjusting their social appropriateness ratings or beliefs after making a self-serving choice. Together, these findings will clarify the psychological mechanisms through which belief elicitation operates and inform the design of low-cost behavioral interventions that promote ethical transparency without relying on external enforcement or social pressure.

Beyond its direct implications for norm activation and transparency interventions, this study also speaks to a broader methodological concern: belief elicitation is often treated as a non-reactive measurement tool in experimental economics. If belief prompts themselves shift behavior as this study aims to test, then their inclusion, particularly before morally sensitive decisions, may unintentionally alter outcomes. This has implications for how belief elicitation is timed and interpreted in other experiments, especially those involving ethical trade-offs, fairness, or social image concerns.

References

Bicchieri, C., Dimant, E., and Sonderegger, S. (2023). It's not a lie if you believe the norm does not apply: Conditional norm-following and belief distortion. *Games and Economic Behavior*, 138:321–354.

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

Dana, J., Weber, R. A., and Kuang, J. X. (2007). Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80.

Exley, C. L. and Kessler, J. B. (2023). Information Avoidance and Image Concerns. *The Economic Journal*, 133(656):3153–3168.

Fischbacher, U. and Föllmi-Heusi, F. (2013). Lies in Disguise—An Experimental Study on Cheating. *Journal of the European Economic Association*, 11(3):525–547.

Giamattei, M., Yahosseini, K. S., Gächter, S., and Molleman, L. (2020). Lioness lab: a free web-based platform for conducting interactive experiments online. *Journal of the Economic Science Association*, 6:95–111.

Grossman, Z. (2014). Strategic ignorance and the robustness of social preferences. *Management Science*, 60(11):2659–2665.

Hua, T. (2025). I didn't know either: How beliefs about norms shape strategic ignorance. MPRA Working Paper No. 124363.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11(3):495–524.

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

Peysakhovich, A. and Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science*, 62(3):631–647.

Schultz, P. W., Nolan, J. M., Cialdini, R. B., Goldstein, N. J., and Griskevicius, V. (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science*, 18(5):429–434.

A Appendix

A.1 Moral wiggle-room Game

The moral wiggle-room game (MWRG) is a binary dictator game in which a dictator chooses between two possible allocations, A or B, between themselves and a receiver. There are two possible states of the world, the Conflicting Interest Game (CIG) and the Aligned Interest Game (AIG). In the full information condition, dictators know which state of the world they are in.

Conflicting Interest Game (CIG)

		Player 1 Gets	Player 2 Gets
Player 1 Chooses	A	6	1
	B	5	5

Aligned Interest Game (AIG)

		Player 1 Gets	Player 2 Gets
Player 1 Chooses	A	6	5
	B	5	1

In the hidden information condition, dictators are again assigned to either the CIG or AIG, but the payoffs are hidden. Thus, dictators do not know which state of the world they are in. However, dictators may reveal the state of the world by clicking on a “REVEAL” button.

Hidden Payoffs Game

		Player 1 Gets	Player 2 Gets
Player 1 Chooses	A	6	?
	B	5	?

REVEAL

The MWRG captures strategic ignorance on the part of dictators. In the canonical game, dictators are told that no one will observe their decision on whether or not to reveal the state of the world. Thus, strategic ignorance is applied against one's self, sometimes interpreted as self-image concerns with regards to an internalized impartial spectator.

Amendment 8.7.2025

Following initial data collection (N = 175) under the 60/10 vs. 50/50 payoff structure, we observed ignorance rates of 45% in the *Pre-Belief* condition and 53% in the *Post-Belief* condition. Exploratory analysis indicated that differences were concentrated among participants classified as Prosocial via the SVO measure, suggesting a possible ceiling effect in the current payoff environment.

To test whether increasing the self-serving incentive and the potential harm to the recipient alters baseline ignorance rates and treatment effects, the payoff structure in the **Conflicting Interest Game (CIG)** and **Aligned Interest Game (AIG)** has been modified as follows:

	Player 1 Gets	**Player 2 Gets**
Conflicting Interest Game (CIG)		
Option A	65	0
Option B	50	50
Aligned Interest Game (AIG)		
Option A	65	50
Option B	50	0

A.2 Belief Elicitation (Moral Wiggle-Room)

Question: Subjects in another session participated in a similar experiment to the one we had just described. What proportion or percentage of people in that session do you think **decided to reveal**?

Enter a value between 0 to 100, corresponding to the percentage of subjects who you think **reveal**.

Your answer _____.

A.3 Treatment

You had predicted that X% of subjects would reveal.

Hidden Payoffs Game

		Player 1 Gets	Player 2 Gets
Player 1 Chooses	A	6	?
	B	5	?

REVEAL

A.4 Social Appropriateness Rating

Directions: How socially appropriate do you think other people in this study believed it was to **reveal** the payoff table, in order to show Person 2's payoffs? If your response matches the most common response of the other participants in this study, you will receive 1 extra ECU.

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate

Directions: How socially appropriate do you think other people believed it was to **not reveal** the payoff table, in order to show Person 2's payoffs? If your response matches the most common response of the other participants in this study, you will receive 1 extra ECU.

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate

Directions: How socially appropriate do you personally believe it is to **reveal** the payoff table, in order to show Person 2's payoffs?

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate

Directions: How socially appropriate do you personally believe it is to **not reveal** the payoff table, in order to show Person 2's payoffs?

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate

Amendment 8.7.2025

In the initial pre-registered design, participants completed a *social appropriateness matching task* in which they rated the perceived appropriateness of each possible choice in the game. These ratings were intended to serve as a secondary measure of norm perception, following the approach in (Krupka and Weber, 2013).
Following the Phase 1 pilot (N = 175), I determined that the primary research question could be addressed without this additional measure. To streamline the experiment, reduce cognitive load on participants, and allocate more time to the main decision tasks, I have **dropped the social appropriateness matching ratings** from the procedure. All other instructions and task order remain unchanged except for the omission of this step.
Data from the original pilot that include social appropriateness ratings will be retained for descriptive and exploratory purposes.

A.5 Dice Rolling Game

The dice-reporting task is based on Fischbacher and Föllmi-Heusi (2013) and is designed to measure dishonest behavior under conditions of plausible deniability. Participants are instructed to privately roll a six-sided die using either a physical die, a phone app, or a browser-based dice roller. The roll is not observed by the experimenter, and participants are asked to report the outcome they obtained.

The payoff is determined solely by the number reported. Because the actual roll is unobservable, participants may choose to misreport in order to earn more. This creates an opportunity for dishonest behavior that cannot be individually detected but can be inferred statistically across the sample.

Participants are instructed: "Please roll a six-sided die using a physical die, smartphone app, or web-based dice roller. Enter the number you rolled below."

They are then presented with a belief elicitation question (Pre or Post condition):

"What percentage of participants do you think intentionally misreported (truthfully reported) their die roll?"

Amendment 8.13.2025

To maintain consistency with the moral wiggle room game, the framing has been changed from "truthfully report" to "accurately report" and from "misreport" to "inaccurately report." This avoids any overt moral imposition. Furthermore, given that the *Pre-Belief* condition effectively serves a neutral framing condition, the *Placebo* conditions will be dropped.

A.6 Belief Elicitation (Dice Rolling)

Question: Subjects in another session participated in a similar experiment to the one we had just described. What proportion or percentage of people in that session do you think **intentionally misreported (truthfully reported)** their number?

Enter a value between 0 to 100, corresponding to the percentage of subjects who you think **misreported (truthfully reported)** their number.

Your answer _____.

Amendment 8.13.2025

To maintain consistency with the moral wiggle room game, the framing has been changed from “truthfully report” to “accurately report” and from “misreport” to “inaccurately report.” This avoids any overt moral imposition. Furthermore, given that the *Pre-Belief* condition effectively serves a neutral framing condition, the Placebo conditions will be dropped.

A.7 Social Appropriateness Rating

Directions: How socially appropriate do you think other people in this study believed it was to **misreport the number rolled on the die**? If your response matches the most common response of the other participants in this study, you will receive 1 extra ECU.

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate

Directions: How socially appropriate do you think other people in this study believed it was to **truthfully report the number rolled on the die**? If your response matches the most common response of the other participants in this study, you will receive 1 extra ECU.

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate

Directions: How socially appropriate do you personally believe it is to **misreport the number rolled on the die**?

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate

Directions: How socially appropriate do you personally believe it is to **truthfully report the number rolled on the die**?

Socially Inappropriate ○ ○ ○ ○ Socially Appropriate