Biased Agents, Extreme Beliefs: Motivated Reasoning Under Competing Models*

Zhongheng Qiao†

This Version: September 17, 2026

Click here for the latest version

Abstract

People often face environments where multiple models compete to explain the same observations. This paper examines how people update beliefs in such settings and how preferences over payoff-relevant states shape model selection and belief updating. This paper first develops a framework where preference-driven bias distorts the perceived model, affecting Bayesian and best-fit updating differently. In a laboratory experiment, most participants are classified as Bayesian updaters, who average across models, while a substantial minority are classified as best-fit updaters, who select the model that best fits the observed signal. Within-participant comparisons between the symmetric payoff and asymmetric payoff conditions indicate that asymmetric payoffs shift reported beliefs toward the preferred state, particularly among participants classified as best-fit updaters. Relative to symmetric payoffs, asymmetric payoffs increase the reported belief of the preferred state by about 8 percentage points among best-fit updaters, while the estimated effect among Bayesian updaters is close to zero. These findings help us better understand model-based learning and have implications for domains such as political polarization and financial investment, where competing narratives and strong preferences often coexist.

Keywords: Belief Updating, Model Uncertainty, Motivated Reasoning, Narratives, Mental Models, Polarization, Experiments

JEL Codes: D01, D81, D83, D91

*First draft: August 31, 2026. I am indebted to Timothy Cason, David Gill, Matthew Kovach, and Colin Sullivan for invaluable guidance. I thank Cuimin Ba, James Bland, John Conlon, Tony Fan, Stanton Hudja, Andrew Olsen, Yaroslav Rosokha, Peter Schwardmann, Elchin Suleymanov, Stephanie Wang, Zitian Wang, Daniel Woods, Sevgi Yuksel, Junya Zhou, and audiences at various seminars and conferences for helpful comments. The study was reviewed by Purdue University's Institutional Review Board (IRB-2024-1786). Financial support from the Krannert Graduate Institute at Purdue University and Institute for Humane Studies is acknowledged.

†Purdue University, 403 Mitch Daniels Blvd, West Lafayette, IN 47907, qiao50@purdue.edu

1 Introduction

Individuals routinely encounter complex learning problems in which the same information can be explained in different ways. Voters are exposed to competing stories about election candidates and economic policies. Investors receive conflicting evaluations of corporate performance. Jurors hear competing narratives about the same evidence. People can draw different conclusions from the same information (Hastorf and Cantril 1954, Lord, Ross and Lepper 1979, Gaines et al. 2007), and their preferences or motivations may affect which interpretation they find convincing (Kunda 1990). These examples illustrate two common features of the environments people face. First, multiple competing models can explain the same observed information and imply different inferences about the unobserved state of the world. Second, people may have intrinsic preferences over unobserved states of the world, such as a favored candidate being the legitimate election winner or a defendant being guilty.

These features raise a central question: how do individuals use competing models to interpret the same information? Under Bayesian updating, individuals update the weight assigned to each model and then average the model-specific posterior probabilities using those posterior model weights. An alternative is to select a single model according to a decision-relevant criterion, such as how well it fits the observed information (Schwartzstein and Sunderam 2021, Aina 2025). Motivated reasoning may work differently under these behavioral rules. Individuals who average across competing models may place more weight on a model that supports their preferred state. Individuals who select a single model may switch to another model, adopting a different explanation (Angrisani, Samek and Serrano-Padial 2024). The same preferences therefore may have different effects on posterior beliefs for individuals following different behavioral rules.

This paper examines how motivated reasoning affects individuals' selection among models and belief updating in a controlled multi-model environment. Here, a model is a data-generating process that maps unobserved states of the world into distributions over observable signals.[1] Specifically, this paper focuses on three interrelated questions: (1) To what extent do individuals' preferences over states affect model selection and belief updating? (2) What behavioral rules do individuals follow to select models and update beliefs in such environments? (3) Do motivated-reasoning effects differ across individuals who follow different behavioral rules?

To answer these questions, I develop a theoretical framework and design a laboratory experiment. In the framework, agents want to learn an unobserved state from information that can be explained by competing models. They have priors over states and models, and each state is associated with a payoff. Preference-driven distortion operates through agents' mental rep-

[1] Throughout the paper, "models" denote these data-generating processes. "Narratives" and "stories" refer to causal interpretations linking observed information to unobserved states.

resentation of the information structure. States associated with higher payoffs receive greater weight in this mental representation. The distortion affects both how agents interpret the observed information under each model and how well each model appears to fit that information. I apply the same distortion under the two behavioral rules. For agents following the Bayesian behavioral rule (Bayesian agents), it affects model weights and the posteriors implied by each model. For agents following the Best-fit behavioral rule (Best-fit agents), it affects the posteriors implied by each model and may also reverse model selection. An agent may switch to a model under which the same information implies a higher probability for the preferred state. Such a switch can generate a large change in posterior beliefs. The framework characterizes the distortion required for these switches and compares the belief changes across behavioral rules.

The experiment uses a $2 \times 2 \times 2$ belief-updating task: two states, two signals, and two models. In the classic ball-and-urn setup, participants observe two carts, A and B; two urns in each cart, X and Y; and purple and orange balls in each urn. The computer selects X or Y according to the given state prior and independently selects either cart with equal probability. Participants observe the color of a ball drawn from the selected urn and report their belief that Urn X was selected. The labels X and Y represent the states, ball colors represent signals, and the two carts represent models. Each participant completes a symmetric payoff (*SP*) block and an asymmetric payoff (*AP*) block of belief updating tasks in random order. In the symmetric payoff block, participants receive the same state-dependent payment whether Urn X or Urn Y is selected. This block provides a benchmark for their updating behavior without a preference for either state. In the asymmetric payoff block, the payment is higher when Urn X is selected. Participants therefore have a reason to prefer Urn X, while the priors and information structures remain unchanged. Each block contains the same nine decision problems, with each problem repeated three times. The decision problems vary in state priors and model pairs. I use reported beliefs from the symmetric payoff block to distinguish the behavioral rules participants follow. I then compare reported beliefs across blocks, controlling for the decision problem and observed signal. This within-subject design allows me to examine how preferences affect belief updating for participants following different behavioral rules. The experiment also measures participants' probabilistic reasoning ability, overconfidence, and risk attitudes, and elicits their opinions about competing economic narratives. The theoretical framework and experimental design motivate three hypotheses. First, asymmetric payoffs shift reported beliefs toward the preferred state. Second, this effect is larger on average among Best-fit participants than among Bayesian participants. Third, the average effect among Best-fit participants increases as the objective fit gap, the difference in how well the two models fit the observed signal, narrows.

The laboratory results show that asymmetric payoffs shift beliefs toward the preferred state, with the largest effects among Best-fit participants. I first apply a finite mixture model to classify participants using only their reported beliefs in the symmetric payoff block. More than

60% of participants are classified as Bayesian, and around 17% are classified as Best-fit. The remainder are classified as Non-movers. Thus, most participants weight competing models, while a substantial minority rely on a single model that fits the signal best, which provides evidence on the Best-fit behavioral rule studied in recent theoretical and experimental work (Aina 2025, Aina and Schneider 2025). Asymmetric payoffs increase the reported beliefs of the preferred state by 1.7 percentage points overall. The effect is 8.4 percentage points among Best-fit participants and 0.5 percentage points among Bayesian participants. Best-fit participants respond significantly more strongly to asymmetric payoffs than Bayesian participants. This difference remains robust under alternative sets of behavioral types. It also holds when I allow different types to have different levels of noise. In that specification, about one-third of participants are classified as Best-fit, and asymmetric payoffs still have a significantly larger effect among Best-fit participants than among Bayesian participants.

These results support Hypotheses 1 and 2. The estimates for Hypothesis 3 have the predicted sign but are not statistically significant. These findings suggest that the behavioral rules individuals follow matter for motivated reasoning. The same preferences may affect reported beliefs differently depending on how individuals use competing models to interpret information. Best-fit participants also report more extreme beliefs in the symmetric payoff block. This difference persists in the decision problem where both behavioral rules predict the same beliefs. Finally, an unincentivized survey suggests a connection between participants' updating behavior in the laboratory and their views about the world. Participants who hold more extreme views on economic narratives are more likely to be classified as Best-fit.

The remainder of this paper is organized as follows. Section 2 reviews the related literature and positions this paper's contributions. Section 3 presents the theoretical framework, characterizing model selection and belief updating under preference-driven distortion. Section 4 presents the experimental design and procedures. Section 5 presents the main results and further discussion. Section 6 concludes.

2 Related Literature

Learning often takes place when several models can explain the same observed information. In such settings, agents need to draw inferences within each model and apply a rule for using the models. This paper studies how preferences over payoff-relevant states affect both parts of this problem.[2] The framework has three central features. First, the same information can have opposite implications for the preferred state under different models. Second, a preference-driven distortion changes model-specific posterior probabilities and how agents perceive each

[2] Kunda (1990) and Bénabou and Tirole (2016) provide broad discussions of motivated beliefs. Battigalli and Dufwenberg (2022) review belief-dependent motivations and psychological games.

model. Third, the effect of this distortion depends on the behavioral rule agents follow to weight or select models.

This paper first contributes to the literature on motivated reasoning. The same information may imply different posterior probabilities under different models, and whether the information supports the preferred state depends on how agents weight or select the models. Preferences can therefore affect model-specific posterior probabilities, model weights, or model selection, which creates an additional channel for motivated reasoning. In most existing studies, information is unambiguously good or bad news, which leaves no room to change its valence. The closest empirical work on motivated reasoning falls into two broad groups. The first group studies motivated reasoning about self-relevant beliefs.[3] Coutts, Gerhards and Murad (2024) show that aggregate feedback jointly determined by own and teammate performance leads participants to distort beliefs about both dimensions, which supports self-serving beliefs about their own ability. Bolte and Fan (2024) find that participants judge identical ego-favorable reports as more likely to represent independent information than identical ego-unfavorable reports. They find no significant asymmetry in posterior beliefs about own test performance. Like my experiment, these designs allow participants to interpret the same information in different ways. However, whether the information is good or bad news remains fixed in these settings.

The second group studies motivated reasoning about uncertain future outcomes or other world states that are not self-relevant. Mayraz (2011) shows that randomly assigned stakes shift beliefs toward payoff-favorable outcomes. The shift is larger when subjective uncertainty is greater. Charness and Dave (2017) also induce preferences over states through payoffs within a known signal structure. Participants with a preferred state update closer to Bayes overall. Barron (2021) varies state-contingent payoffs within a known information structure and finds approximately Bayesian updating on average, without a significant good-news asymmetry.[4] Like these studies, I use state-dependent payoffs to create a preferred state and extend this setting to competing models, so preferences can also affect how agents weight or select models.[5]

Recent work studies motivated reasoning when the source or accuracy of information is uncer-

[3] For example, this literature studies intelligence and perceived ability (e.g., Eil and Rao 2011, Möbius et al. 2022), task performance (e.g., Ertac 2011, Coutts 2019, Drobner and Goerg 2024), self-serving fairness (e.g., Babcock et al. 1995), persuasion and self-persuasion (e.g., Schwardmann and van der Weele 2019, Schwardmann, Tripodi and van der Weele 2022), and motivated or biased memory (e.g., Zimmermann 2020, Huffman, Raymond and Shvets 2022, Conlon 2026).

[4] Evidence for asymmetric belief updating is mixed across settings, including self-relevant ones. Some experiments find no asymmetry, stronger responses to bad news, or effects that depend on the setting (Ertac 2011, Coutts 2019, Thaler 2026). Drobner (2022) finds optimistic updating when participants expect no immediate resolution and neutral updating under immediate resolution. Engelmann et al. (2024) find wishful thinking in the loss domain but not in the gain domain, and stronger when the evidence is more ambiguous.

[5] Related work studies politically motivated beliefs about news, events, or facts (Kimball et al. 2024, Thaler 2024, Barron, Becker and Huck 2025). Another strand studies how motivated beliefs are exchanged or supplied in social settings (Oprea and Yuksel 2022, Thaler 2021) and how senders react to wishful thinkers (Filiz-Ozbay and Liu 2026). Other work studies motivated reasoning in asset valuation or project continuation (Wang and Zheng 2026, Qiao 2025).

tain. Thaler (2024) identifies motivated reasoning through trust in information sources. After reporting a guess on politicized questions or on their own performance, subjects receive a message on whether the correct answer lies above or below that guess. The message may come from one of two sources, one always correct and one always incorrect. The experiment shows that subjects overtrust messages that favor their own political party or performance. Avtar and Ballyk (2026) study how motivated reasoning affects beliefs about the source of ego-relevant information. Subjects report the probability that a message came from the more informative of two sources. They find that motivation shifts source beliefs in an ego-favorable direction, especially when knowing the source matters more for inferring the state. Hu and Papadokonstantaki (2025) instead study ego-relevant beliefs about the state when feedback accuracy is known, compound-uncertain with stated probabilities, or ambiguous. They find that asymmetric updating is strongest with known accuracy and absent under ambiguity. I use state-dependent payoffs in an abstract learning environment to isolate preferences from self-image. In this learning environment, competing models provide conflicting inferences about the uncertain states, so the valence of information depends on how subjects use the models. In addition, even when the model is held fixed, motivated reasoning may affect the model-specific posteriors. This paper therefore examines how motivated reasoning affects belief updating through both channels, and how its effect differs across the behavioral rules subjects follow.

This paper also develops a theoretical framework linking motivated reasoning to model uncertainty. Kovach (2020) axiomatizes wishful thinking by allowing an agent's endowed act to shift beliefs toward states in which that act yields higher utility. Building on this idea, I apply a preference-driven distortion separately within each competing model. The distortion raises the perceived likelihood of states that yield higher utility.[6] Within each model, it changes model-specific posterior beliefs and subjective fit. Chambers, Masatlioglu and Raymond (2023) characterize distortions that produce the same posterior whether they are applied before or after conditioning on a signal. The distortion in the present framework belongs to this class. Liu (2026) studies two sources of bias within a known information structure. Beliefs shift toward states that initially appear more likely and toward states that yield higher utility. I focus on the second effect and study how it changes inference across competing models.[7] A complementary literature gives beliefs or expectations a direct role in utility (Brunnermeier and Parker 2005, Battigalli and Dufwenberg 2009, Battigalli, Corrao and Dufwenberg 2019). Brunnermeier and Parker (2005) let subjective expectations of future utility flows enter current felicity. Optimal

[6] Ba, Bohren and Imas (2024) develop and test a two-stage model. Agents first form a mental representation of the information structure by overweighting states most representative of the observed signal. They then process this representation imprecisely. The present paper also uses a distorted mental representation of the information structure that overweights the preferred states.

[7] Another strand of the literature on non-Bayesian updating studies related distortions that depend on the agent's current beliefs over states (e.g., Rabin and Schrag 1999, Fryer Jr, Harms and Jackson 2019; see Benjamin 2019 for a systematic review). In this paper, the distortion depends on the payoff from each state.

beliefs balance anticipatory gains from optimism against losses caused by distorted decisions.[8] In the present framework, utility depends only on the realized state. State-dependent payoffs determine the preference-driven distortion of model fits and posterior beliefs.

This paper also contributes to the literature on model selection. It develops a common framework in which preference-driven distortion operates under the Bayesian and best-fit behavioral rules. The framework predicts that the distortion affects agents following these rules differently. The closest theoretical work studies fit-based model selection.[9] Schwartzstein and Sunderam (2021) study persuasion when a receiver selects the model that makes the observed data most likely, given the receiver's prior over states. Aina (2025) builds on this fit-based rule in a setting where a persuader supplies multiple models before the signal is realized. The author also notes Bayesian model averaging as an alternative, under which posterior model probabilities weight model-specific posteriors.[10] I adopt the best-fit behavioral rule from this work alongside the Bayesian rule and connect both rules to motivated reasoning.[11]

Among experiments with competing models, Aina and Schneider (2025) provide the closest benchmark in a setting without induced preferences over states: subjects update beliefs when the signal may come from two conflicting models. Subjects observe both models' predictions, which separates model weighting from errors in probabilistic reasoning. They find that the most frequent updating rule gives full weight to the model that best fits the observed data. I study a similar updating problem and ask how motivated reasoning affects participants who follow different behavioral rules. The symmetric payoff (*SP*) block in my paper provides a comparable environment, except that participants do not observe these model-specific predictions. Based on beliefs from the *SP* block, a nontrivial proportion of participants are classified as Best-fit updaters. However, the most common behavioral rule in the *SP* block is the Bayesian rule, which differs from their finding. The asymmetric payoff (*AP*) block adds the motivated reasoning component, which is the margin this paper studies.

Related experiments explore how people use other forms of models. Musolff and Zimmermann (2025) study uncertainty between two models with varying complexity. They find that greater complexity increases neglect of model uncertainty in participants' actions, while participants' elicited beliefs still reflect model uncertainty. Ambuehl and Thysen (2025) focus on how people choose between conflicting causal models. They find that subjects tend to focus on worst-

[8] Battigalli and Dufwenberg (2009) allow utility to depend on updated beliefs and beliefs about other players' beliefs. Battigalli, Corrao and Dufwenberg (2019) embed belief-dependent motivations into dynamic games.

[9] The best-fit, or maximum likelihood, rule is also used in other settings. Gilboa and Schmeidler (1993) axiomatize the maximum likelihood update rule for multiple priors, in which the agent keeps the priors that make the observed event most likely. Later work extends the rule (e.g., Epstein and Schneider 2007, Kovach 2024, Suleymanov 2025). In this paper, the priors over states and models are explicitly given to the participants.

[10] Related theory also focuses on criteria other than fit (Yang 2023), strategic supply of models (Jain 2023), and model sharing (Schwartzstein and Sunderam 2022). Other theory studies belief updating on different grounds, such as minimal movement from priors (Dominiak, Kovach and Tserenjigmid 2023, 2025).

[11] A separate literature studies learning when the true model may be absent from the learner's set (Bohren and Hauser 2021, Montiel Olea et al. 2022, Fudenberg and Lanzani 2023, Ba 2026).

case outcomes and choose cautiously when they cannot identify the correct model. Farina and Herman (2026) focus on strategic choice of data-generating processes and find that receivers often neglect how an informed sender's private type shapes the sender's unobserved choice of data-generating process. Other experiments examine different forms of model uncertainty. Epstein and Halevy (2024) focus on settings where signals have multiple interpretations. They compare belief updating from noisy signals with updating from ambiguous signals, and find that ambiguous signals increase deviations from Bayes' rule. Liang (2025) studies compound and ambiguous uncertainty about source accuracy. Participants respond less to reports from uncertain sources than to reports with known accuracy, especially after good news.[12] The present paper introduces an additional component, *motivated reasoning*, that shapes model selection and belief updating. It documents that motivated reasoning has substantially different impacts on people who follow different behavioral rules, and the effects are especially large for best-fit agents.

More broadly, the paper also connects to research on mental models and narratives.[13] Eliaz and Spiegler (2020) represent a narrative as a causal model that maps actions into consequences. Agents adopt the narrative that maximizes anticipatory utility among those that correctly predict the observed consequences. Eliaz and Spiegler (2026) study news media that supply information and narratives to maximize consumers' anticipatory utility. Fan and Fries (2026) show that constructing or receiving a well-fitting narrative makes beliefs more extreme and volatile than learning directly from the same data. In the present paper, Bayesian model weights respond continuously to the distortion. Best-fit selection can switch the chosen model and move posterior beliefs discretely. The distinction between updating rules identifies an individual-level channel for discrete changes in posterior beliefs. Its implications for extremeness depend on the model-specific posteriors and the information structure.[14]

3 Theoretical Framework

This section proposes a framework that links motivated reasoning to belief updating with competing models. A preference-driven distortion changes the perceived information structure and

[12] A broader literature studies belief updating under ambiguity and compound uncertainty. For example: Ngangoué (2021), Shishkin and Ortoleva (2023) study belief updating under ambiguity, Moreno and Rosokha (2016), Bland and Rosokha (2021) compare belief updating under compound risk with ambiguity, and Kovach, Martin and Tserenjigmid (2026) study belief updating with qualitative AI recommendations.

[13] Broad discussions and reviews include Shiller (2017), Roos and Reccius (2024), Barron and Fries (2024b). Related experiments study responses to supplied causal models (Charles and Kendall 2025, Ambuehl and Thysen 2025, He et al. 2026), mental-model formation from data (Esponda, Vespa and Yuksel 2024, Fréchette, Vespa and Yuksel 2024, Kendall and Oprea 2024, Fan 2024), and narrative construction and persuasion (Barron and Fries 2024a, Andre et al. 2026, Liu and Zhang 2026). Related field evidence documents the adoption of competing narratives (Angrisani, Samek and Serrano-Padial 2024).

[14] Related work also links political values or rationalization to belief divergence (Barron, Becker and Huck 2025, Le Yaouanq, Schwardmann and van der Weele 2025).

therefore affects posterior beliefs. The impact of this distortion depends on the behavioral rule people follow. Best-fit agents choose the model that best fits the observed information, whereas Bayesian agents update model weights across competing models.

3.1 Environment

An agent wants to learn an unknown state ω from the finite set $\Omega \equiv \{\omega_1, \ldots, \omega_N\}$, where $N > 1$. The state prior is $\mu_0 \in \Delta(\Omega)$, and $\mu_0(\omega)$ denotes the prior probability of state ω.

The finite signal space is $S \equiv \{s_1, \ldots, s_K\}$, where $K > 1$. A model m maps unobserved states of the world into distributions over observable signals, which is often called an information structure in the belief-updating literature. $\pi_m(s \mid \omega)$ denotes the probability of signal s in state ω under model m. For every m and ω, $\sum_{s \in S} \pi_m(s \mid \omega) = 1$. Let $\mathbb{M}$ denote the set of all such maps from Ω to $\Delta(S)$.

The agent is exposed to a finite set of competing models $M \subseteq \mathbb{M}$ and has a model prior $\rho_0 \in \Delta(M)$. The state and model are independent under the prior: $\Pr(\omega, m) = \mu_0(\omega)\rho_0(m)$.[15]

Conditional on model m and observed signal s, Bayes' rule gives the model-specific posterior $\mu_m(\omega \mid s)$. The objective fit $P_m(s)$ is the likelihood of observing signal s under model m, given the state prior:

$$\mu_m(\omega \mid s) = \frac{\mu_0(\omega)\pi_m(s \mid \omega)}{\sum_{\omega' \in \Omega} \mu_0(\omega')\pi_m(s \mid \omega')}, \tag{1a}$$

$$P_m(s) = \sum_{\omega \in \Omega} \mu_0(\omega)\pi_m(s \mid \omega). \tag{1b}$$

This section focuses on two behavioral rules for updating beliefs with competing models: the best-fit rule and the Bayesian rule. Best-fit agents select the model with the highest objective fit and update only under that model (Schwartzstein and Sunderam 2021, Aina 2025). m_s^* denotes the model with the highest objective fit to signal s, and $\mu_F(\omega \mid s)$ denotes the posterior belief of a best-fit agent:

$$m_s^* \in \arg\max_{m \in M} P_m(s), \qquad \mu_F(\omega \mid s) = \mu_{m_s^*}(\omega \mid s). \tag{2}$$

Bayesian agents retain uncertainty over the competing models. After observing s, they update each model's weight according to its prior probability and objective fit. $\rho(m \mid s)$ denotes the weight assigned to model m after observing signal s:

$$\rho(m \mid s) = \frac{\rho_0(m)P_m(s)}{\sum_{m' \in M} \rho_0(m')P_{m'}(s)}. \tag{3}$$

[15] This assumption makes the state prior common across models, as in Aina (2025).

Bayesian agents then average the model-specific posterior probabilities using these updated model weights:

$$\mu_B(\omega \mid s) = \sum_{m \in M} \rho(m \mid s)\mu_m(\omega \mid s). \tag{4}$$

Here, $\mu_B(\omega \mid s)$ denotes the posterior belief of a Bayesian agent.

I impose the following regularity conditions. The state and model priors have full support. Every conditional signal distribution also has full support:

$$\mu_0(\omega) > 0, \qquad \rho_0(m) > 0, \qquad \pi_m(s \mid \omega) > 0$$

for every $\omega \in \Omega$, $m \in M$, and $s \in S$. These conditions ensure that all posterior probabilities are strictly positive and that the ratios used below are well-defined.

Models in M are distinct, so any two models differ for at least one state-signal pair. A fixed tie-breaking rule resolves all fit ties and applies uniformly across signals and distortions.

3.2 Preferences over States

To introduce preference-driven distortions, suppose each state is associated with a payoff to the agent. Let $U(\omega)$ denote the agent's utility from the payoff associated with state ω. Without loss of generality, I can order the states so that $U(\omega_1) \geq U(\omega_2) \geq \cdots \geq U(\omega_N)$.

Motivated reasoning may shift beliefs toward states that give the agent higher utility. Following Kovach (2020), I use a strictly increasing function $f : \mathbb{R} \to \mathbb{R}_{++}$ to model preference-driven distortion. The term $f(U(\omega))$ determines the relative distortion associated with each state. Because f is strictly increasing, states with higher utility receive greater relative weight. For a given model m and observed signal s, the distorted model-specific posterior probability of state ω, $\hat{\mu}_m(\omega \mid s)$, is given by:

$$\hat{\mu}_m(\omega \mid s) = \frac{\mu_0(\omega)\pi_m(s \mid \omega)f(U(\omega))}{\sum_{\omega' \in \Omega} \mu_0(\omega')\pi_m(s \mid \omega')f(U(\omega'))} = \frac{\mu_m(\omega \mid s)f(U(\omega))}{\sum_{\omega' \in \Omega} \mu_m(\omega' \mid s)f(U(\omega'))}. \tag{5}$$

3.2.1 Interpretation of the Distortion

I interpret the distortion as modifying the agent's mental representation of the information structure. For each model m, the distorted mental representation is defined by:

$$\hat{\pi}_m(s \mid \omega) = \pi_m(s \mid \omega)f(U(\omega)). \tag{6}$$

Substituting $\hat{\pi}_m$ for π_m in Bayes' rule yields the distorted model-specific posterior in Equation (5). This posterior is well-defined, although $\hat{\pi}_m$ need not be a conditional probability

distribution.[16] $\hat{\pi}_m$ is a conditional probability distribution only when $f(U(\omega)) = 1$ for every state.

Preference-driven distortion also changes the agent's perception of how well each model fits the observed signal. For model m and signal s, subjective fit, $\hat{P}_m(s)$, aggregates the distorted mental representation across states using the state prior:

$$\begin{aligned}\hat{P}_m(s) &= \sum_{\omega\in\Omega} \mu_0(\omega)\hat{\pi}_m(s \mid \omega) \\ &= \sum_{\omega\in\Omega} \mu_0(\omega)\pi_m(s \mid \omega)f(U(\omega)) \\ &= P_m(s) \sum_{\omega\in\Omega} \mu_m(\omega \mid s)f(U(\omega)).\end{aligned} \tag{7}$$

Equation (7) shows that subjective fit is the product of objective fit and the posterior expectation of $f(U)$ under model m. Therefore, preference-driven distortion can favor models under which the observed signal indicates a higher probability on states the agent prefers. Subjective fits determine which model best-fit agents select. Together with model priors, they determine model weights Bayesian agents assign across models.[17]

3.2.2 Belief Updating under Preference-Driven Distortion

Preference-driven distortion can affect beliefs in two ways. It changes posterior probabilities within each model and changes how the agent chooses or weights competing models.

Within any model m, Equation (5) implies the following relationship between objective and distorted posterior odds. For any two states $\omega, \omega' \in \Omega$,

$$\frac{\hat{\mu}_m(\omega \mid s)}{\hat{\mu}_m(\omega' \mid s)} = \frac{\mu_m(\omega \mid s)}{\mu_m(\omega' \mid s)} \frac{f(U(\omega))}{f(U(\omega'))}. \tag{8}$$

If $U(\omega) > U(\omega')$, then $f(U(\omega))/f(U(\omega')) > 1$. The distorted posterior odds of ω relative to ω' therefore exceed the posterior odds without distortion. If $f(U(\omega))$ is constant across states, preference-driven distortion leaves every model-specific posterior unchanged.

Subjective fit, defined in Equation (7), enters the two behavioral rules as follows. A best-fit agent selects the model with the highest subjective fit and updates her posterior belief under

[16] Ba, Bohren and Imas (2024) use the term "pseudo-information structure" for a similar distorted mental representation. Their model overweights signal likelihoods in states that are more representative of the observed signal.

[17] Alternatively, the preference-driven distortion can be incorporated into the prior over states while leaving the information structure unchanged. Replacing the prior term $\mu_0(\omega)$ in Bayes' rule with $\mu_0(\omega)f(U(\omega))$ yields the same distorted posterior belief as Equation (5). Liu (2024) discusses the equivalence between placing a state-dependent multiplicative distortion in the prior and in the signal likelihoods. In my setting, the two interpretations generate the same model selections for Best-fit agents and the same distorted model weights for Bayesian agents.

that model. $\hat{m}_s^*$ denotes the model with the highest subjective fit to signal s, and $\hat{\mu}_F(\omega \mid s)$ denotes the distorted posterior belief of a best-fit agent:

$$\hat{m}_s^* \in \arg\max_{m \in M} \hat{P}_m(s), \qquad \hat{\mu}_F(\omega \mid s) = \hat{\mu}_{\hat{m}_s^*}(\omega \mid s). \tag{9}$$

A Bayesian agent uses subjective fits and model priors to update model weights. The agent then averages the distorted model-specific posteriors using these weights. $\hat{\rho}(m \mid s)$ denotes the distorted weight assigned to model m after observing signal s, and $\hat{\mu}_B(\omega \mid s)$ denotes the distorted posterior belief of a Bayesian agent:

$$\hat{\rho}(m \mid s) = \frac{\rho_0(m)\hat{P}_m(s)}{\sum_{m' \in M} \rho_0(m')\hat{P}_{m'}(s)}, \qquad \hat{\mu}_B(\omega \mid s) = \sum_{m \in M} \hat{\rho}(m \mid s)\hat{\mu}_m(\omega \mid s). \tag{10}$$

3.3 Competing Models

This section first examines how preference-driven distortion affects model selection and then examines its effects on posterior beliefs. I begin with a general result that allows any finite number of states, signals, and models. After that, the full-switch result requires two signals and two models, while the sensitivity result focuses on the $2 \times 2 \times 2$ environment that matches my experiment. Finally, I derive implications for posterior beliefs.

3.3.1 Definitions

I begin by defining a condition that compares how well two models fit the same signal.

Definition 1 (Relative Representativeness of Two Models)**.** Fix two distinct models $m, m' \in M$ and a signal $s \in S$. Model m is more representative than m' for s if

1. $\pi_m(s \mid \omega) \geq \pi_{m'}(s \mid \omega)$ for every $\omega \in \Omega$; and
2. $\pi_m(s \mid \omega) > \pi_{m'}(s \mid \omega)$ for some $\omega \in \Omega$.

This condition means that, in every state, signal s is at least as likely under model m as under model m'. As a result, m always has higher objective fit and subjective fit than m'. Proposition 1 formalizes this implication.

The following definition classifies how preference-driven distortion changes the model selected by a best-fit agent.

Definition 2 (Model Switch)**.** For best-fit agents, preference-driven distortion produces no switch if $\hat{m}_s^* = m_s^*$ for every $s \in S$. It produces a full switch if $\hat{m}_s^* \neq m_s^*$ for every $s \in S$. It produces a partial switch if the selected model changes for some signals and remains unchanged for others.

3.3.2 Predictions

I first link relative representativeness to subjective fit in a setting with any finite number of states, signals, and models.

Proposition 1 (Model Fit)**.** *Suppose m is more representative than m' for signal s. Then the following statements hold:*

1. $\hat{P}_m(s) > \hat{P}_{m'}(s)$;
2. $\frac{\hat{\rho}(m|s)}{\hat{\rho}(m'|s)} > \frac{\rho_0(m)}{\rho_0(m')}$;
3. $m' \notin \arg\max_{m'' \in M} \hat{P}_{m''}(s)$. *If* $M = \{m, m'\}$, $\hat{m}^*_s = m$.

Proof. See Appendix A.2.1. □

For Bayesian agents, part 2 implies that the posterior model odds of m relative to m' under preference-driven distortion are greater than the prior model odds. For Best-fit agents, part 3 implies that m' cannot be selected while m is available. When these are the only two models, the agent selects m.

I next consider the case with two signals, allowing any finite number of states and models. In this case, differences in objective and subjective fit between any two models $m, m' \in M$ reverse across signals:

$$P_m(s_2) - P_{m'}(s_2) = -\big[P_m(s_1) - P_{m'}(s_1)\big], \qquad \hat{P}_m(s_2) - \hat{P}_{m'}(s_2) = -\big[\hat{P}_m(s_1) - \hat{P}_{m'}(s_1)\big]. \quad (11)$$

Appendix A.1 formally derives these results, which lead to the following prediction for two models and any finite state space.

Prediction 1 (Full Switch)**.** *Suppose $S = \{s_1, s_2\}$, $M = \{m, m'\}$, and the models have distinct model fits. For a best-fit agent, preference-driven distortion produces either a full switch or no switch.*

Proof. See Appendix A.2.2. □

If the two models have equal model fit, the selected model is determined by the tie-breaking rule, and a partial switch may occur.

I next characterize the distortion required for a switch in the experimental $2 \times 2 \times 2$ environment. Let $\Omega = \{\omega_1, \omega_2\}$, $S = \{s_1, s_2\}$, $M = \{m, m'\}$. Suppose $U(\omega_1) > U(\omega_2)$, so that ω_1 is the preferred state.

For signal s_1, suppose neither model is more representative than the other. Suppose further that

$$P_m(s_1) > P_{m'}(s_1) \qquad \text{and} \qquad \mu_m(\omega_1 \mid s_1) < \mu_{m'}(\omega_1 \mid s_1).$$

Thus, model m has higher objective fit for s_1, while model m' yields a higher model-specific posterior probability on the preferred state.

Define

$$\begin{aligned} A &= \mu_0(\omega_1)\big[\pi_{m'}(s_1 \mid \omega_1) - \pi_m(s_1 \mid \omega_1)\big], \\ B &= \mu_0(\omega_2)\big[\pi_m(s_1 \mid \omega_2) - \pi_{m'}(s_1 \mid \omega_2)\big]. \end{aligned} \tag{12}$$

The term A captures how much more likely signal s_1 is under m' than under m when ω_1 is true, accounting for the prior probability $\mu_0(\omega_1)$. Similarly, B captures how much more likely s_1 is under m than under m' when ω_2 is true, accounting for $\mu_0(\omega_2)$. Because $P_m(s_1) - P_{m'}(s_1) = B - A$, we have $B > A > 0$.

Define

$$R = \frac{A}{B} \in (0,1), \qquad t = \frac{f(U(\omega_1))}{f(U(\omega_2))} > 1.$$

The ratio R compares the two differences in the objective fit calculation. Because $R < 1$, a value closer to one means that A and B more nearly offset each other. The ratio t compares the preference-driven distortion across the two states. A larger t places greater relative weight on the preferred state.

Proposition 2 (Sensitivity of Model Switch)**.** *Given $B > A > 0$, preference-driven distortion produces a full switch if and only if*

$$t > \frac{1}{R}.$$

Proof. See Appendix A.2.3. □

Preference-driven distortion produces a full switch when $tA > B$. I call $1/R = B/A$ the switching threshold. When R is close to one, A nearly offsets B, so the models have similar objective fits. A relatively small distortion toward ω_1 can then make m' have higher subjective fit than m. When R is smaller, A offsets less of B, which implies that a larger distortion is required to reverse the order of subjective fit.

I next examine how preference-driven distortion affects posterior beliefs. I return to the finite state, signal, and model spaces defined above and fix a signal $s \in S$. For Bayesian agents, preference-driven distortion changes both model-specific posterior probabilities and model weights. Relative model weights shift toward models under which the observed signal supports states the agent prefers. Appendix A.1 provides the pairwise relationship between posterior model odds before and after preference-driven distortion.

For every state $\omega \in \Omega$, the distorted Bayesian posterior belief satisfies

$$\hat{\mu}_B(\omega \mid s) = \frac{\mu_B(\omega \mid s) f(U(\omega))}{\sum_{\omega' \in \Omega} \mu_B(\omega' \mid s) f(U(\omega'))}. \tag{13}$$

Equation (13) combines the changes in model-specific posteriors and model weights into a direct distortion of the Bayesian agent's posterior.

Consider the experimental $2 \times 2 \times 2$ environment. Let $\Omega = \{\omega_1, \omega_2\}$, $S = \{s_1, s_2\}$, $M = \{m, m'\}$. Suppose $U(\omega_1) > U(\omega_2)$, so that ω_1 is the preferred state.

Proposition 3. *For every signal $s \in S$,*

$$\hat{\mu}_B(\omega_1 \mid s) > \mu_B(\omega_1 \mid s), \qquad \hat{\mu}_F(\omega_1 \mid s) > \mu_F(\omega_1 \mid s).$$

Moreover, the model selected under preference-driven distortion satisfies

$$\mu_{\hat{m}_s^*}(\omega_1 \mid s) \geq \mu_{m_s^*}(\omega_1 \mid s).$$

Proof. See Appendix A.2.4. □

For Bayesian agents, Equation (13) shows that preference-driven distortion shifts their posterior belief toward the preferred state. For best-fit agents, the distortion raises the posterior probability on the preferred state within every model, regardless of model switch. If a switch occurs, the new model yields at least as high a model-specific posterior probability on the preferred state as the original model. The within-model change and the change in model selection therefore move the distorted best-fit posterior belief in the same direction. A switch produces an additional discrete increase in the posterior belief.

I next examine how posterior beliefs change as preference-driven distortion toward the preferred state becomes stronger. Let $\Omega = \{\omega_1, \omega_2\}$ and $U(\omega_1) > U(\omega_2)$. Write

$$t = \frac{f(U(\omega_1))}{f(U(\omega_2))} > 1.$$

A larger t places greater relative weight on the preferred state ω_1.

Let $\hat{m}_s^*(t)$ denote the model selected by a best-fit agent after signal s when the relative distortion is t. The term $\mu_{\hat{m}_s^*(t)}(\omega_1 \mid s)$ denotes the model-specific posterior probability on ω_1 under that selected model when there is no distortion. The term $\hat{\mu}_F(\omega_1 \mid s)$ denotes the distorted posterior belief for best-fit agents.

Proposition 4. *For every signal $s \in S$:*

1. *The distorted Bayesian posterior belief $\hat{\mu}_B(\omega_1 \mid s)$ is continuous and strictly increasing in t on $(1, \infty)$.*

2. *The model-specific posterior probability on ω_1 under the selected model, $\mu_{\hat{m}_s^*(t)}(\omega_1 \mid s)$, is non-decreasing in t on $(1,\infty)$. It remains constant between switching thresholds.*

3. *The distorted best-fit posterior belief $\hat{\mu}_F(\omega_1 \mid s)$ is strictly increasing in t on $(1,\infty)$. It is continuous except at finitely many switching thresholds. A switch to a model with a strictly higher model-specific posterior probability on the preferred state produces an upward jump.*

Proof. See Appendix A.2.5. □

For Bayesian agents, Equation (13) shows that a higher t continuously raises the distorted Bayesian posterior belief on the preferred state. For Best-fit agents, a higher t makes the likelihood of s in ω_1 more influential in subjective fit. The selected model therefore remains unchanged or switches to a model with an equal or higher model-specific posterior probability on ω_1. This explains why $\mu_{\hat{m}_s^*(t)}(\omega_1 \mid s)$ is non-decreasing and may remain constant over an interval.

Even when the selected model remains unchanged, a higher t raises the distorted model-specific posterior probability on ω_1 within that model. The distorted best-fit posterior belief is therefore strictly increasing. A switch produces an upward jump when the new model has a strictly higher model-specific posterior probability on the preferred state.

Thus, both $\hat{\mu}_B(\omega_1 \mid s)$ and $\hat{\mu}_F(\omega_1 \mid s)$ strictly increase with t. The distorted Bayesian posterior belief changes continuously. The distorted best-fit posterior belief is continuous between switching thresholds and may jump upward when the selected model changes.

4 Experimental Design

To answer the research questions, this paper implements a lab experiment based on the classical ball-and-urn setup to simulate a multi-model learning environment. As formally defined in Section 3.1, a model is a data-generating process that maps unobserved states of the world into distributions over observable signals, which is often called an information structure in the belief-updating literature. Participants are explicitly informed about the priors over states and over models. This feature provides a clear Bayesian benchmark for belief updating and helps distinguish heterogeneous behavioral rules for model selection and belief updating in this environment. The experiment employs a within-subject design with respect to participants' preferences over states: Each participant completes one block of belief-updating tasks with equal state-dependent payoffs and another block with unequal state-dependent payoffs. This design allows the experiment to identify whether, and to what extent, preferences over unobserved states affect model selection and belief updating.

The experiment consists of five parts. In Part 1, participants complete two blocks of belief-updating tasks in random order: the *Symmetric Payoff Belief Updating Task (SP)* and the *Asymmetric Payoff Belief Updating Task (AP)*.[18] Each block contains 27 belief-updating scenarios. Earnings are summed across all scenarios. In Part 2, the experiment elicits participants' risk attitudes using the bomb risk task (Crosetto and Filippin 2013). In Part 3, participants complete eight cognitive tasks related to information processing and statistical reasoning, presented in random order. Seven of these tasks are adapted from Enke, Graeber and Oprea (2023), while the eighth task presents two identical carts and measures participants' Bayesian updating under the experiment's environment.[19] Each correct response adds \$0.40 to participants' final earnings. After completing these tasks, participants are asked to predict their task score from 0 to 8 and whether it exceeds the session average score. Each correct prediction earns an additional \$0.40. In Part 4, participants are presented with six pairs of conflicting economic narratives and report their views on each pair using an unincentivized 11-point Likert scale. Finally, Part 5 collects demographic information. Appendix D shows the instructions and the main screens participants saw.

4.1 Belief Updating Task

Blocks *SP* and *AP* each consist of 27 scenarios. I use the direct method to elicit posterior beliefs. In each scenario, participants observe one randomly drawn signal. They then report their posterior belief using the displayed priors, the two competing information structures, and the observed signal. Signals are drawn independently across scenarios. A decision problem represents a unique combination of the prior over states and the information structures provided by the pair of competing models. The same nine decision problems appear in both blocks, and participants face each decision problem 3 times within a block.[20] The order of the 27 scenarios within each block is randomized separately for each participant.

In each scenario, participants see **2** carts (Cart A and Cart B) and **4** urns (two urns per cart), as Figure 1 shows.[21] Each cart contains two urns: one labeled X and the other labeled Y. Each urn contains **20** balls, consisting of a mixture of purple and orange balls. Participants are informed

[18] The order of the *SP* and *AP* blocks was randomized at the individual level. The instructions inform participants that Part 1 consists of two blocks of 27 scenarios each. The details for Block 2 are not disclosed before participants complete Block 1.

[19] The seven tasks from Enke, Graeber and Oprea (2023) include base-rate neglect, correlation neglect, balls-and-urns updating, gambler's fallacy, sample-size neglect, regression to the mean, and the acquiring-company problem. Cason, Hudja and Rosokha (2025) also use a separate urn task to measure participants' ability to update beliefs in an environment similar to the main experiment.

[20] For each decision problem, three repetitions allow me to observe most participants' beliefs under both signal realizations within a block.

[21] The two information structures are randomly assigned to the cart labels and corresponding display positions, with equal probability. For example, the cart shown as Cart A in Figure 1 can instead be labeled Cart B and displayed at the bottom.

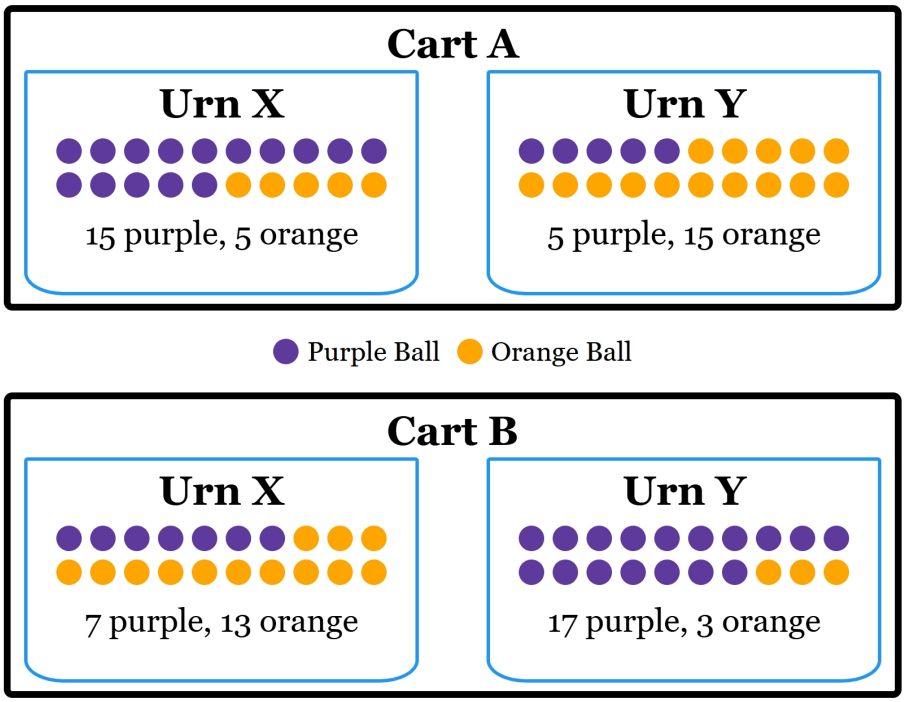


Figure 1: Ball Composition from One Scenario

that the computer randomly selects one urn using a two-step process. (1) The computer rolls a fair 10-sided die numbered 1 to 10 to determine whether Urn X or Urn Y is selected. The die ranges selecting Urns X and Y vary across decision problems and are displayed on the screen. For a prior probability of 0.50 on Urn X, rolls 1–5 select Urn X and rolls 6–10 select Urn Y. (2) Independently of the die roll, the computer flips a fair coin to determine which cart is selected. Heads selects Cart A, and tails selects Cart B, each with probability 0.50. Participants do not observe which urn is selected. One ball is randomly drawn from the selected urn, and participants observe only its color (purple or orange). After seeing the ball's color, participants report a probability from 0 to 100 using a slider. This number represents their guess of the probability, expressed as a percentage, that the computer selected Urn X from either cart.

In Part 1 of the experiment, participants are paid based on their decisions from all 54 scenarios across both blocks. Participants' earnings in each scenario come from two sources: the urn payment and the guess prize. The urn payment depends on the urn selected by the computer. In the *SP* block, participants receive 5 tokens regardless of which urn is selected. In the *AP* block, participants receive 25 tokens if Urn X is selected and 5 tokens if Urn Y is selected. Tokens are converted to dollars at a rate of 150 tokens per dollar. For the guess prize, I use the binarized scoring rule (Savage 1971, Hossain and Okui 2013) to elicit participants' beliefs. Participants' guesses determine the probability of winning the guess prize of 80 tokens in the corresponding scenario.[22] After each scenario, participants receive no feedback about the true state, their urn

[22] Following the discussion of belief elicitation methods by Danz, Vesterlund and Wilson (2022, 2024), Healy and Leo (2025), I do not directly provide participants with the details of the binarized scoring rule on the interface. However, they can access the rule and a calculator by clicking the corresponding buttons on the interface. The

payment, or whether they won the guess prize. After completing the experiment, they can view a summary table covering all 54 scenarios.

In this environment, the selected urn's label (X or Y) corresponds to the true state of the world, while the selected cart corresponds to the true model (signal-generating process). The observed ball's color corresponds to the observed data. The urn payment induces participants' preferences over the two states. It depends only on the selected urn, regardless of the selected cart or observed ball color. The urn payment changes payoffs across states without changing the objective information structure. Under Bayesian updating, it does not affect posterior beliefs. Under risk neutrality, it also does not affect the optimal report under the binarized scoring rule.[23] This experimental design allows me to examine whether these induced preferences affect participants' model selection and belief updating processes.

4.2 Treatment Overview

The primary objective in designing the experiment is to create a paradigm that allows observation of participants' posterior beliefs in a multi-model environment and to manipulate participants' preferences over states within that environment. In order to identify the behavioral rules participants follow in the experiment, certain parameters are varied across scenarios. The treatment structure is as follows.

Within-subject design. Participants complete 27 scenarios in each of two blocks; block order and scenario order within each block are randomized. The tasks vary across scenarios and blocks along the following dimensions:

- *Participants' preference over payoff-relevant states:* In the *SP* block, participants are indifferent between Urns X and Y; in the *AP* block, participants prefer Urn X over Urn Y.
- *Prior belief over states:* This parameter may vary across scenarios, allowing observation of different updating behaviors.
- *Pair of models provided:* Model pairs vary across decision problems. In 8 out of 9 decision problems, participants are presented with competing models that offer conflicting interpretations of the same observed signal.

rule description follows Wilson and Vespa (2018), while the interface of the belief-updating task is adapted from Bland and Rosokha (2024).

[23] With a concave utility function, the state-dependent payment may create hedging incentives. In the *AP* block, a risk-averse participant may report a lower belief in Urn X than they hold. A lower report raises the chance of winning the guess prize when Urn Y is selected, which offsets the lower urn payment in that case. This would move reported beliefs away from the preferred state, which is the opposite direction of motivated reasoning, so it would tend to make the estimated effects of motivated reasoning smaller. I elicit participants' risk attitudes and control for them in regression analysis to further address this concern.

Table 1: Decision Problems

No.	State prior $\mu_0(X)$	Model A		Model B		Objective Fit for s_p		Best-fit Agents	
		$\pi_A(s_p \mid X)$	$\pi_A(s_p \mid Y)$	$\pi_B(s_p \mid X)$	$\pi_B(s_p \mid Y)$	$P_A(s_p)$	$P_B(s_p)$	$\mu_F(X \mid s_p)$	$\mu_F(X \mid s_o)$
1	0.50	0.70	0.25	0.35	0.85	0.475	0.600	0.292	0.286
2	0.50	0.70	0.25	0.40	0.95	0.475	0.675	0.296	0.286
3	0.50	0.65	0.15	0.40	0.95	0.400	0.675	0.296	0.292
4	0.50	0.85	0.35	0.25	0.70	0.600	0.475	0.708	0.714
5	0.50	0.75	0.20	0.75	0.20	0.475	0.475	0.789	0.238
6	0.30	0.80	0.35	0.30	0.75	0.485	0.615	0.146	0.117
7	0.30	0.75	0.25	0.35	0.85	0.400	0.700	0.150	0.125
8	0.70	0.60	0.15	0.40	0.85	0.465	0.535	0.523	0.523
9	0.70	0.55	0.05	0.45	0.95	0.400	0.600	0.525	0.525

Notes: Urns X and Y are the two states, and s_p and s_o denote a purple and an orange ball. $\pi_m(s_p \mid \omega)$ is the probability of drawing a purple ball from Urn ω under Model m, so $\pi_m(s_o \mid \omega) = 1 - \pi_m(s_p \mid \omega)$. $P_m(s_p)$ is the objective fit of Model m after a purple ball, defined in Equation (1b). After an orange ball, $P_m(s_o) = 1 - P_m(s_p)$. $\mu_F(X \mid s)$ is the Best-fit benchmark, the posterior probability of Urn X under the model with the higher objective fit, defined in Equation (2). In Decision Problem 5, the two models are identical.

The same decision problem can generate different signal realizations in different scenarios. Table 1 provides a summary of the 9 decision problems. Let s_p and s_o denote the purple and orange signals, respectively.

In Decision Problem 4, after either signal, the objectively best-fit model also produces the larger posterior probability on Urn X. In Decision Problem 5, Cart A and Cart B have the exact same ball composition, reducing the problem to a classical single-model belief-updating task. This decision problem provides a single-model benchmark within the experiment.

4.3 Hypotheses

The experiment allows me to examine the behavioral rules participants follow and how asymmetric payoffs affect belief updating.

The Bayesian and best-fit rules in Section 3.1 generate benchmark posterior beliefs for each decision problem and signal realization. I use participants' 27 reported beliefs from the *SP* block to classify them into types based on the behavioral rule they are likely to follow. I then compare reported beliefs between the *SP* and *AP* blocks, controlling for the decision problem and signal realization. I test the following hypotheses using these within-subject comparisons, with participant types determined from *SP* block behavior.

Proposition 3 predicts that preference-driven distortion raises the posterior probability of the preferred state under both behavioral rules. In the *AP* block, the urn payment makes Urn X the preferred state. This yields the first hypothesis:

Hypothesis 1 (General Motivated Reasoning Effect)**.** *For identical decision problems, participants will report higher posterior beliefs that Urn X was selected in the AP block compared to the SP block, conditional on the same signal realization.*

As described in Section 3.1, Bayesian agents average model-specific posterior probabilities using updated model weights. Best-fit agents select a single model and update their beliefs using that model. Preference-driven distortion changes model-specific posterior beliefs under both rules. Proposition 4 shows that the beliefs of Bayesian agents change continuously with the strength of the distortion. The beliefs of best-fit agents change continuously when they do not switch models and can increase discretely when they switch models.

Prediction 1 and Proposition 2 show that sufficiently strong preference-driven distortion can produce a full switch in Decision Problems 1–3 and 6–9. Without distortion, best-fit agents select the model with the lower posterior probability of Urn X after either signal, as Table 1 shows. Following a full switch, they select the model with the higher posterior probability after each signal. The beliefs of Bayesian agents lie between the two model-specific posterior probabilities, both before and after distortion. For the same degree of distortion, a full switch therefore produces a larger increase in the beliefs of best-fit agents than in those of Bayesian agents.[24] This comparison motivates the following hypothesis about average motivated reasoning effects across agent types:

Hypothesis 2 (Heterogeneity in Motivated Reasoning Effect)**.** *Asymmetric payoffs have a larger average effect on reported beliefs among participants classified as best-fit agents than among those classified as Bayesian agents.*

Proposition 2 characterizes the distortion required for a full switch. Define the objective fit gap as $|P_A(s) - P_B(s)|$, the difference between the objective fits of the two models after signal s (Equation (1b)). A full switch is possible in Decision Problems 1–3 and 6–9. Among these problems, those with the same state prior (Decision Problems 1–3, 6–7, and 8–9) have lower switching thresholds when their objective fit gaps are smaller.[25] For a given distribution of preference-driven distortion, lower thresholds can increase the fraction of best-fit agents who switch models. These switches move agents toward the model that implies a higher posterior probability of selecting Urn X. More frequent switches can therefore amplify the average upward shift in reported beliefs. This switching channel motivates the following empirical hypothesis about average belief effects across decision problems:

Hypothesis 3 (Sensitivity in Motivated Reasoning)**.** *For participants classified as best-fit agents,*

[24] When there is no model switch, best-fit agents are affected by preference-driven distortion in a moderate and continuous way, as the distortion affects the model-specific posterior probabilities. However, in that case, the sizes of the belief changes for best-fit agents and Bayesian agents are not obvious.

[25] The switching threshold is $B/A = 1 + (B-A)/A$, with A and B defined in Equation (12). Its value depends on both the objective fit gap and A.

the average effect of motivated reasoning on reported beliefs increases as the objective fit gap narrows.

4.4 Experimental Procedure

I collected experimental data from 193 student participants at the Vernon Smith Experimental Economics Laboratory (VSEEL) at Purdue University in March 2025. The experiment was pre-registered at the AEA RCT Registry in January 2025 (https://doi.org/10.1257/rct.15136).[26] I drew participants from the VSEEL subject pool (administered using ORSEE; Greiner 2015). The sample is balanced by gender, with 49.7% female participants. The experiment was programmed in oTree (Chen, Schonger and Wickens 2016). Upon arrival, participants were assigned random participant IDs, and they electronically provided consent to participate. I provided the instructions to each participant on the computer screen, together with printed instructions. The experimenter read the instructions aloud at the front of the room. After that, participants completed ten incentivized quiz questions and earned $0.50 for each question answered correctly. Participants received immediate feedback to clarify any misunderstandings they might have had, regardless of whether they answered the questions correctly.

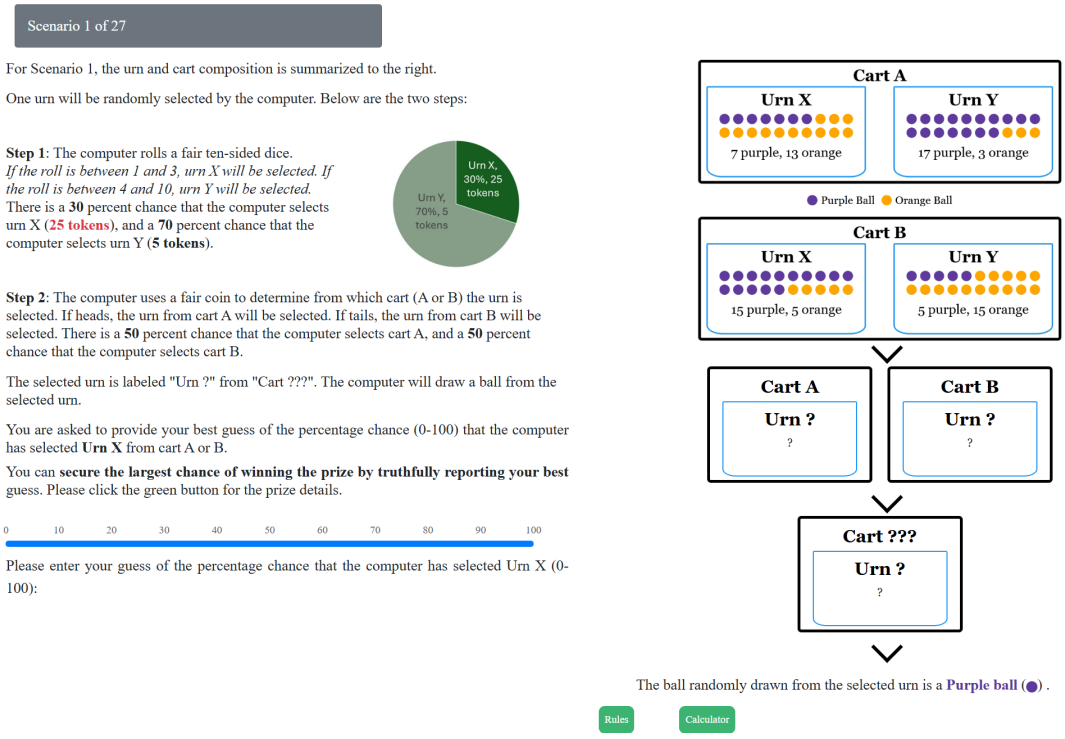


Figure 2: Screenshot of an Example Scenario

After the comprehension quiz, the main part of the experiment began. Each participant com-

[26] I pre-registered a target sample of 200 participants. The data were collected across 7 in-person sessions on March 5, March 6, and March 13, 2025. Due to fluctuations in the show-up rate, I eventually collected data from 193 participants.

pleted two blocks of belief-updating tasks and made one decision for each scenario. Figure 2 presents a screenshot of an example scenario.[27] The experiment lasted approximately 70 minutes, with an average total payment of $22.15. Table B-1 in Appendix B.1 provides a summary of the demographic statistics.

The experiment also measures participants' risk attitude, probabilistic reasoning, overconfidence, and opinions on economic narratives in Parts 2–4. In Part 2, I use the bomb risk task from Crosetto and Filippin (2013). Participants see 100 boxes, and one box is collected every second after they press Start. Each collected box is worth 15 tokens, but one randomly located box contains a bomb that destroys all earnings from the task. Participants decide when to stop collecting, and I use the number of boxes collected to measure risk attitude. A risk-neutral participant collects 50 boxes, and collecting more boxes indicates greater risk tolerance. In Part 3, I measure participants' probabilistic reasoning with eight cognitive tasks (see footnote 19 for the list of tasks). Participants have 10 minutes to complete this part. The cognitive score is the number of correct responses across the eight tasks. After these tasks, participants predict how many of the eight tasks they answered correctly. Overconfidence is an indicator equal to one if this predicted score exceeds the actual score. In Part 4, participants rate each of six pairs of conflicting economic narratives on a scale from zero to ten. Narrative extremeness is the mean absolute distance of the six ratings from the midpoint of five. It ranges from zero to five, and higher values indicate more extreme stated opinions.

5 Results

This section first describes participants' reported beliefs and classifies them using their decisions from the *Symmetric Payoff (SP)* block. It then examines the effect of motivated reasoning on reported beliefs.

5.1 Description of Behavioral Patterns

Each participant reports beliefs in 27 scenarios in the *SP* block. For each scenario, I derive benchmark beliefs under the Bayesian and Best-fit behavioral rules from the decision problem and signal realization. Table 2 shows the shares of reported beliefs within two or five percentage points of these benchmarks. Nearly 29% of symmetric payoff reports lie within two percentage points of the Bayesian benchmark, compared with 5.5% for the Best-fit benchmark. With a five-percentage-point window, these shares rise to 47.7% and 12.2%, respectively.[28]

[27] Figures D-8 – D-12 in Appendix D show the full sequence of screens in a scenario.

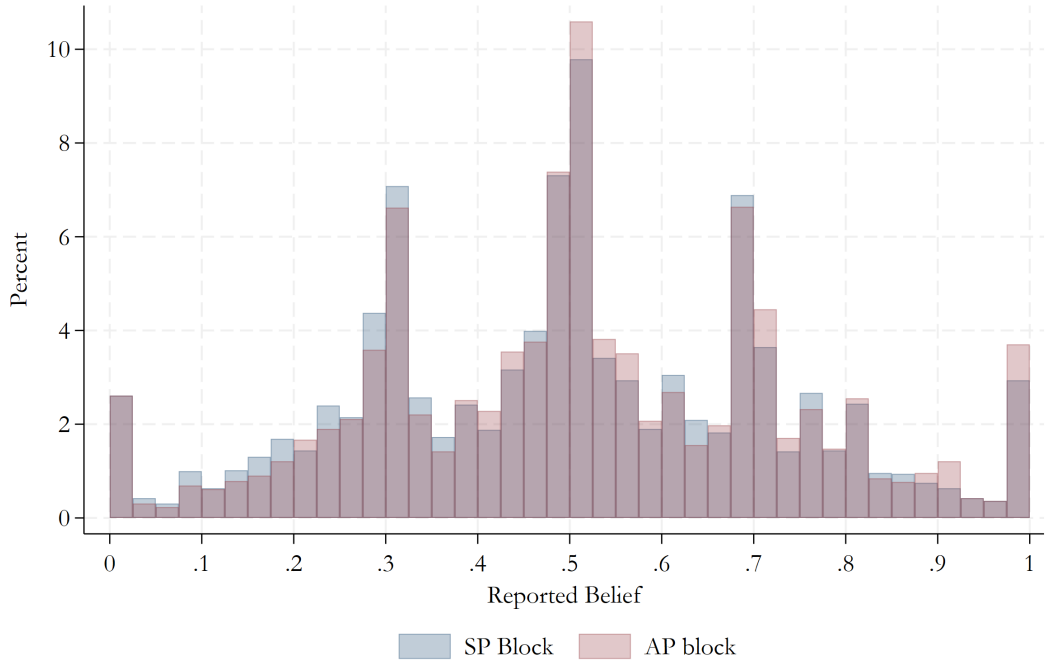
[28] A reported belief can lie within five percentage points of both benchmarks when they are sufficiently close. In Decision Problem 5, for example, both behavioral rules yield the same benchmark.

See Appendix C for details on the follow-up study supplying model-specific posteriors and its comparison with the main study.

Table 2: Classification of Round-level Beliefs in the *SP* Block

Type of Guess	Within 2 p.p.		Within 5 p.p.	
	%	95%-CI	%	95%-CI
Bayesian	28.98	[25.29, 32.67]	47.75	[43.21, 52.28]
Best-fit	5.51	[4.47, 6.54]	12.15	[10.35, 13.94]

Notes: The table reports the percentage of beliefs within the indicated distance of each benchmark. The sample contains 5,211 reports from 193 participants in the *SP* block. The 95% confidence intervals are based on standard errors calculated from participant-level shares.



Notes: The histograms pool all 10,422 scenarios from 193 participants. The horizontal axis shows the reported beliefs of Urn X. Blue bars represent the *SP* block and red bars represent the *AP* block.

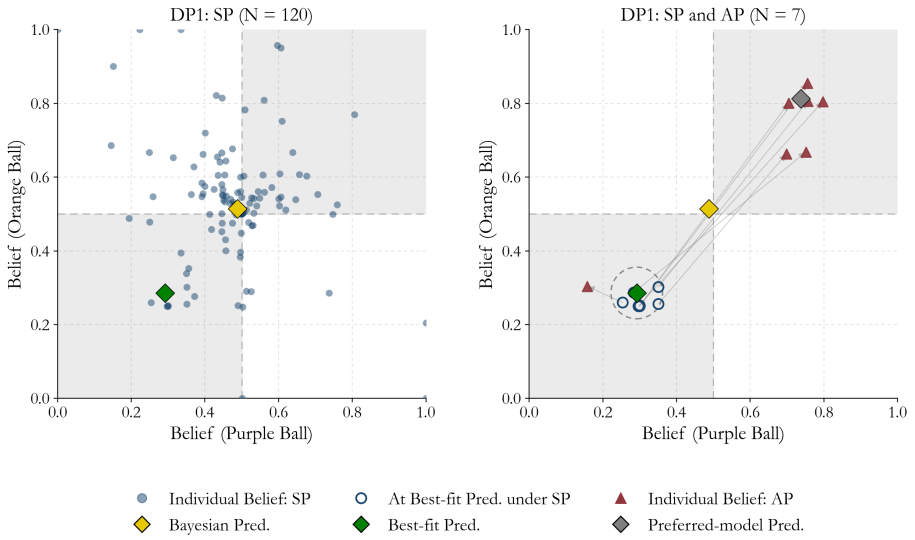
Figure 3: Distribution of Reported Beliefs Across All Scenarios

Figure 3 displays the distribution of reported beliefs in the *SP* and *AP* blocks. Reported beliefs concentrate around the state priors of 0.3, 0.5, and 0.7 listed in Table 1. A nontrivial share of beliefs also occurs at 0 and 1. The distributions overlap a lot, with the *AP* distribution shifted slightly to the right.

Figure 4 plots participants' reported beliefs in Decision Problem 1. Each point represents a participant's vector of beliefs following the two signal realizations.[29] The yellow diamond represents the vector of Bayesian benchmark beliefs, and the green diamond represents the vector

[29] Participants face this decision problem three times in each block. The figure includes only participants who

of Best-fit benchmark beliefs. Following Aina (2025), a vector of posterior beliefs is Bayes-consistent if the state prior is a strict convex combination of the posteriors across signals.[30] The shaded areas mark the region of vectors of beliefs that are not Bayes-consistent. The left panel plots each participant's vector of beliefs in the *SP* block. Many points cluster near the Bayesian benchmark, while a smaller group clusters near the Best-fit benchmark in the shaded area. Similar to Aina and Schneider (2025), about half of the vectors are Bayes-inconsistent.



Notes: Both panels use Decision Problem 1 and include only participants who observe both signal realizations in both blocks. Each belief is the participant's first reported belief after the given signal realization within a block. The horizontal axis shows beliefs after observing a purple ball, and the vertical axis shows beliefs after observing an orange ball. The left panel plots each participant's pair of beliefs in the *SP* block. The right panel includes the participants whose *SP* pair lies within a Euclidean distance of $0.05\sqrt{2}$ of the Best-fit benchmark, shown by the dashed circle. For each of these participants, a hollow circle marks the *SP* pair, a triangle marks the *AP* pair, and an arrow connects the two. The yellow and green diamonds represent the Bayesian and Best-fit benchmarks. The grey diamond represents the benchmark of using the model that implies a higher probability of Urn X. Shaded areas mark the Bayes-inconsistent region.

Figure 4: Vector of Reported Beliefs in Decision Problem 1

The right panel of Figure 4 provides a simple illustration of how participants near the Best-fit benchmark respond to asymmetric payoffs. It plots the seven participants whose *SP* beliefs in Decision Problem 1 lie inside the dashed circle around the Best-fit benchmark. In the *AP* block, six of them exhibit a large, discrete jump in their belief vectors, and their vectors are close to the grey diamond, which represents the beliefs implied by using the model with the higher posterior probability of Urn X after each signal realization. None of the seven moves to the Bayesian benchmark. This descriptive pattern is consistent with the full switch in Prediction 1, in which a best-fit agent changes the selected model after both signal realizations. Section 5.3

observe both signal realizations from Decision Problem 1 in both blocks (N = 120). For the signal realization observed twice within a block, the figure uses the first of the participant's two reported beliefs.

[30] With a state prior of 0.5, this requires one posterior probability of Urn X to lie above 0.5 and the other to lie below 0.5, or both to equal 0.5.

estimates the effects of asymmetric payoffs by behavioral type.

5.2 Structural Classification of Agent Types

To better understand these descriptive patterns and formally test the theoretical hypotheses, I use a finite mixture model to classify participants by their behavioral rules. This classification allows me to examine how asymmetric payoffs affect beliefs across behavioral types.

Classifying participants into distinct behavioral types presents several challenges. First, the behavioral rules participants follow are not directly observed. Second, different rules can yield similar or identical predictions in some decision problems. Third, computational and reporting errors can cause observed beliefs to deviate from a rule's predictions, even when a participant intends to follow that rule (Alós-Ferrer and Garagnani 2023).

The structural classification therefore models errors explicitly. Without errors, each rule predicts a single posterior belief in each scenario. Reported beliefs rarely match these predictions exactly. Only 47.7% of reports in the *SP* block lie within five percentage points of the Bayesian benchmark, and 12.1% lie within five percentage points of the Best-fit benchmark (Table 2). A classification without errors would therefore leave many participants inconsistent with every behavioral rule. As in El-Gamal and Grether (1995), allowing for errors gives each behavioral rule a positive likelihood for every observed behavior. For example, a Best-fit updater may misidentify the best-fitting model because of a mistake in calculating model fit. This mistake matters most when the two models fit the observed signal similarly, because a small error in fit calculation can switch the selected model and move the reported belief to the other model's posterior. Participants may also report beliefs that differ from their intended responses.[31] Allowing for these errors means that the same reported belief can be produced by multiple behavioral rules, with different likelihoods. The model evaluates each participant's beliefs across all 27 scenarios in the *SP* block to help distinguish behavioral rules from pure errors. As a result, a single report that deviates from a type's prediction carries limited weight in the classification.[32]

In the finite mixture model, each participant belongs to one of three latent types specified *ex ante*: Bayesian, Best-fit, and Non-mover.[33] Bayesian agents average model-specific posteriors using their model weights; Best-fit agents select the model with the highest fit; and Non-movers retain the state prior. Following El-Gamal and Grether (1995), the model assumes that each

[31] When participants see the same decision problem and signal more than once in the *SP* block, their reports differ by 9.3 percentage points on average, and only 14% of repeated reports match exactly (Appendix B.6.2).

[32] The classification excludes all reports from the *AP* block, because deriving benchmark beliefs for this block would require additional assumptions about participants' preference distortion term $f(U(\omega))$.

[33] The Bayesian and Best-fit types are directly informed by the theory. The main specification includes a separate Non-mover type because Bayesian benchmarks lie close to the state prior in some decision problems. This type helps distinguish Bayesian updating from non-moving behaviors. See Appendix B.5 for details on the estimation of the alternative two-, four-, and five-type classifications.

participant follows the same underlying rule across scenarios.

The model allows two layers of mean-zero normal errors, with standard deviations shared across types. The first layer is an error in calculating model fit. In the *SP* block, a participant perceives the objective fit of each model, defined in Equation (1b), with a normal error. Bayesian agents weight the model-specific posteriors using these perceived fits, and Best-fit agents select the model with the higher perceived fit. The second layer is a reporting error. It is added to the belief implied by the participant's rule, so the reported belief can deviate from that belief.[34] Non-movers have only reporting errors, as they only focus on the state prior.

For example, consider a purple signal in Decision Problem 1 (see Table 1). The objective fits of Models A and B are 0.475 and 0.600, and their model-specific posterior probabilities of Urn X are 0.737 and 0.292. Without errors, a Best-fit agent selects Model B and reports 0.292, and a Bayesian agent reports 0.488. A fit-calculation error can reverse the order of the two perceived fits. A Best-fit agent then selects Model A, and the belief implied by the rule becomes 0.737. Such a reversal is more likely when the two objective fits are close. A reporting error can move the reported belief away from the belief implied by the rule, for example from 0.292 to 0.302.

I estimate population shares and the two error standard deviations, one for each layer, by maximum likelihood, using all 5,211 reports from the *SP* block. Population shares are the estimated fractions of each type in the population. Table 3 presents the estimates, and Appendix B.2 describes the likelihood and classification procedure.

Table 3: Three-Type Parameter Estimates

Parameter	**Estimate**	**Bootstrap SE**
Population shares		
p_1 (Bayesian)	0.683	0.037
p_2 (Best-fit)	0.174	0.026
p_3 (Non-mover)	0.144	0.029
First-layer (fit-calculation) error SD		
σ_1 (Bayesian and Best-fit)	0.143	0.011
$\sigma_{1,\text{Non-mover}}$	—	
Second-layer (reporting) error SD		
σ_2 (all types)	0.132	0.006

Notes: The table reports maximum likelihood estimates of population shares and error standard deviations shared across types. The sample contains 5,211 beliefs from 193 participants in the *SP* block, with 27 beliefs per participant. Population shares are the estimated fractions of each type in the population. Standard errors are calculated from 500 bootstrap samples. See Appendix B.2 for details.

[34] With the reporting error, the belief implied by the rule and the error can fall below zero or above one, but participants can only report beliefs between zero and one. Therefore, the likelihood uses normal densities for reports strictly between zero and one. For reports at zero and one, it uses the normal probabilities below zero and above one, respectively.

Using Decision Problem 1 again as an example, the objective fits of the two models differ by 0.125 after either signal. At the estimated σ_l of 0.143, the model implies that a Best-fit agent selects the model with the lower objective fit after about 27% of signals, and that a Bayesian agent places a higher weight on the model with the lower objective fit after about 27% of signals in this problem.

For each participant, I use the estimated population shares as prior probabilities of belonging to each behavioral type. I then update these priors using the likelihood of the participant's 27 reports in the *SP* block under each type. This likelihood measures how well the type's predictions and estimated error standard deviations account for the participant's beliefs in the *SP* block. A higher likelihood increases a type's posterior probability relative to other types, holding the prior probabilities fixed. Appendix B.2 explains how these likelihoods are calculated. I assign each participant to the type with the highest posterior probability and report the classification in Table 4. Of 193 participants, 132 are classified as Bayesian, 33 are classified as Best-fit, and 28 are classified as Non-movers.[35]

Table 4: Participant Classification

Type	Assigned	Above 50%	Above 70%	Above 90%
Bayesian	132	132	129	118
Best-fit	33	32	32	31
Non-mover	28	28	25	18
Unclassified	-	1	7	26

Notes: The table reports the number of participants classified as each behavioral type. The sample contains 193 participants. The Assigned column classifies each participant as the type with the highest posterior probability. The Above 50%, Above 70%, and Above 90% columns count only participants whose highest posterior probability exceeds 50%, 70%, and 90%, respectively.

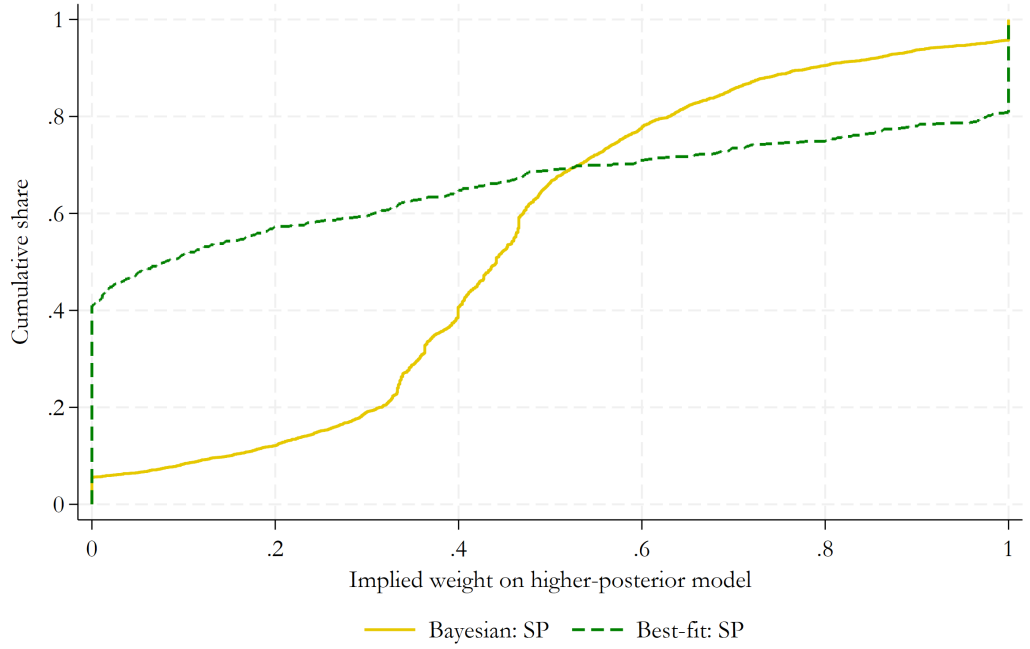
Result 1. *Approximately two-thirds of participants are classified as Bayesian, while approximately one-sixth are classified as Best-fit.*

The structural classification suggests that most participants weight information from competing models, while about one-sixth of participants rely on a single model when updating beliefs. This proportion relies on the assumption that all types share the same error standard deviations. This assumption is strong, because participants who follow different rules may also differ in how precisely they calculate model fits and report beliefs. I therefore also estimate a version that allows type-specific error standard deviations. This version classifies 63 participants (33%) as Best-fit (see Appendix B.4 for details). All 33 participants classified as Best-fit in the main

[35] For 167 of the 193 participants, the assigned type has a posterior probability above 90%. For 192 of the 193 participants, the assigned type has a posterior probability above 50%. For one participant, the assigned type has a posterior probability of 0.42. Among the 167 participants whose assigned type has a posterior probability above 90%, 31 (19%) are classified as Best-fit and 118 (71%) as Bayesian.

specification remain Best-fit, and 26 of the 30 additional Best-fit participants are classified as Bayesian in the main specification.

More participants are classified as Best-fit because, when error standard deviations differ across types, the Best-fit type can fit noisy reports with larger errors. The estimated Best-fit type has a much larger fit-calculation error than the Bayesian type (Table B-7 in Appendix B.4), so it often selects the model with the lower objective fit, whose posterior is closer to the Bayesian benchmark than to the Best-fit benchmark. It can therefore fit participants whose reports are more dispersed around the Bayesian benchmark. With shared error standard deviations, the classification relies less on how dispersed a participant's reports are, so I use shared error standard deviations in the main specification and report the type-specific version in Appendix B.4.



Notes: The figure shows the cumulative distributions of implied weights on the higher model-specific posterior, separately for participants classified as Bayesian and Best-fit. The sample includes reports from Decision Problems 1–4 and 6–9 in the *SP* block.

Figure 5: Report-level Implied Weights in the *SP* Block

Figure 5 provides descriptive evidence on the classification results by showing how participants' beliefs relate to the two model-specific posteriors. For each report, I derive the implied weight on the higher model-specific posterior, which expresses the report as a weighted average of the two model-specific posteriors. In the example of Decision Problem 1 above, the two model-specific posteriors after a purple signal are 0.737 and 0.292. A reported belief of 0.5 is therefore a weighted average with a weight of $(0.5-0.292)/(0.737-0.292) \approx 0.47$ on the higher posterior.[36] Decision Problem 5 is excluded because the two model-specific posteriors

[36] For beliefs below the lower model-specific posterior, the implied weight is set to zero. For beliefs above the higher model-specific posterior, it is set to one. In the *SP* block, 17.3% of reports in Decision Problems 1–4 and

coincide. The figure shows that implied weights for participants classified as Bayesian concentrate at intermediate values, while implied weights for Best-fit participants have larger masses at zero and one. These patterns are broadly consistent with the distinction between the two behavioral rules. See Appendix B.6.3 for report-level implied weights in the *AP* block.

5.3 Effects of Motivated Reasoning

This section compares beliefs between the *SP* and *AP* blocks to test the hypotheses about motivated reasoning. In the *AP* block, participants receive a higher urn payment when Urn X is selected. Hypothesis 1 predicts higher beliefs of Urn X in the *AP* block than in the *SP* block. Hypothesis 2 predicts a larger average effect of motivated reasoning on reported beliefs among Best-fit participants than among Bayesian participants.

I first estimate the following specification:

$$\text{Belief}_{it} = \alpha + \beta_1 \mathbf{1}_{\text{Asym},it} + \beta_2 \mathbf{X}_i + \lambda_{d_{it}} + \eta_{s_{it}} + \tau_t + \varepsilon_{it}. \tag{14}$$

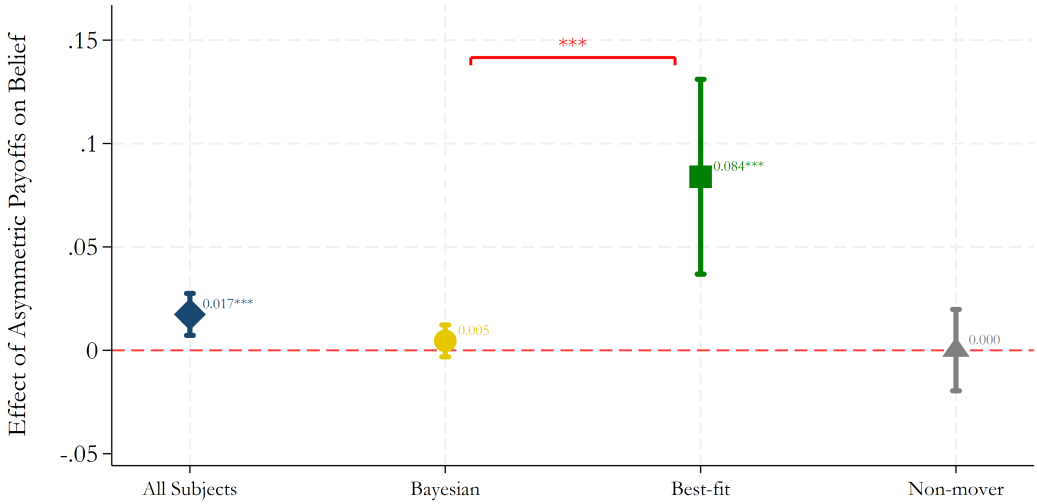
Here, $\text{Belief}_{it} \in [0,1]$ is participant i's reported belief of Urn X in round t. The indicator $\mathbb{1}_{\text{Asym},it}$ equals one in the *AP* block. The participant covariates $\mathbf{X}_i$ include gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness, defined in Section 4.4. I also include fixed effects for decision problem, signal color, and round. Standard errors are clustered by participant.[37]

Figure 6 reports the estimated coefficient on the *AP* block indicator for the pooled sample and separately for each behavioral type. Conditional on the specified participant covariates and fixed effects, this coefficient measures how asymmetric payoffs affect posterior beliefs of Urn X, and I refer to this coefficient as the effect of motivated reasoning. Consistent with Hypothesis 1, asymmetric payoffs increase the reported belief of Urn X by 1.7 percentage points overall ($p < 0.001$). The estimated effect is 0.5 percentage points among Bayesian participants ($p = 0.245$) and 8.4 percentage points among Best-fit participants ($p = 0.001$). In addition, the estimated effect is significantly larger among Best-fit participants than among Bayesian participants ($p < 0.001$), consistent with Hypothesis 2.[38]

6–9 lie outside the range of the two model-specific posteriors. In a related experiment, Aina and Schneider (2025) find that 12.1% of guesses lie outside the range of the two model predictions.

[37] The results are robust to participant fixed effects. See Appendix B.3 for details on these estimates and the first-block comparison.

[38] One may be concerned that participants who complete the *AP* block first carry its effects into the *SP* block, which would affect both their classification and the estimated effect of motivated reasoning. The order of the two blocks was randomized, and 91 participants completed the *SP* block first. For these participants, types are classified only from beliefs reported before they faced asymmetric payoffs. Among them, the effect of motivated reasoning is 6.5 percentage points larger among Best-fit participants than among Bayesian participants ($p = 0.014$).



Notes: The figure reports estimated coefficients on $\mathbb{1}_{\text{Asym}}$ in Equation (14) for the pooled sample and separately for each behavioral type. The pooled, Bayesian, Best-fit, and Non-mover samples contain 193, 132, 33, and 28 participants, and each participant contributes 54 beliefs. Regressions control for gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness, and include fixed effects for signal color, round, and decision problem. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level. The red bracket reports the significance of the difference between the Best-fit and Bayesian coefficients. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure 6: Motivated Reasoning Effect by Type

Result 2. *Asymmetric payoffs have a significantly larger effect on Best-fit participants' reported beliefs than on Bayesian participants' reported beliefs.*

I next examine how the effect of motivated reasoning varies with posterior type probabilities. I interact $\mathbb{1}_{\text{Asym}}$ with each participant's posterior probabilities of being the Best-fit and Non-mover types, using the Bayesian type as the reference group. These probabilities are estimated from each participant's 27 reports in the *SP* block. Table 5 reports a positive coefficient of 0.082 on the interaction between $\mathbb{1}_{\text{Asym}}$ and the Best-fit probability ($p < 0.001$). Thus, the estimated effect of motivated reasoning increases with the Best-fit probability, which is consistent with the separate estimates in Figure 6.[39]

[39] See Appendix B.3 for details on effects of motivated reasoning and closeness to Best-fit versus Bayesian predictions measured using reports from other decision problems.

Table 5: Motivated Reasoning Effect by Likelihood of Type

DV: Reported Belief	(1)
$\mathbb{1}_{\text{Asym}}$	0.004
	(0.004)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Best-fit Type	0.082***
	(0.023)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Non-mover Type	-0.007
	(0.011)
Likelihood of Best-fit Type	-0.069***
	(0.018)
Likelihood of Non-mover Type	0.004
	(0.008)
Observations	10422
Participants	193
Mean of DV	0.506
R^2	0.343

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with posterior type probabilities. The dependent variable is reported belief. Unreported controls include gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. The specification includes fixed effects for signal color, round, and decision problem. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

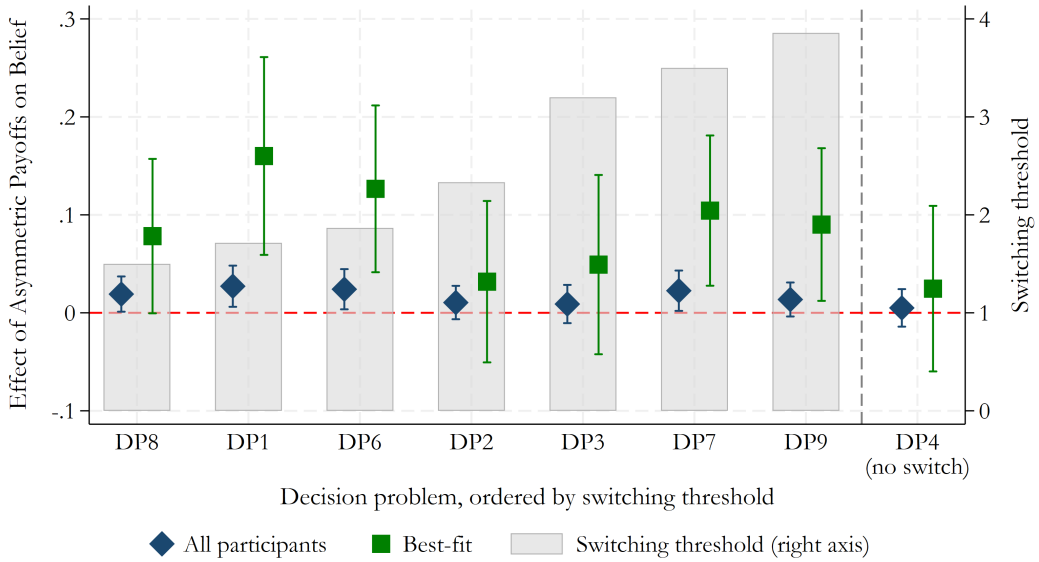
The main results on motivated reasoning are robust to alternative two-, four-, and five-type classifications. Asymmetric payoffs have a significantly larger effect among Best-fit participants than among Bayesian participants under all four specifications. See Appendix B.5 for details on the corresponding estimated effects of motivated reasoning and regressions using posterior type probabilities.

As described in Section 5.2, the main specification assumes that the error standard deviations are shared across types. When I relax this assumption and allow type-specific error standard deviations, asymmetric payoffs still have a significantly larger effect among Best-fit participants than among Bayesian participants. The estimated effects are 4.9 and 0.1 percentage points, respectively, compared with 8.4 and 0.5 percentage points in Figure 6, and their difference is statistically significant ($p < 0.001$). See Appendix B.4 for details on this specification.

5.4 Tests of the Model-Switch Predictions

I next examine how the effect of motivated reasoning varies across decision problems. In Decision Problems 1–3 and 6–9, the objectively best-fitting model assigns a lower posterior probability to Urn X than the other model after either signal. A sufficiently strong preference-driven distortion may therefore lead Best-fit agents to select the model with the higher posterior

probability. Prediction 1 and Proposition 2 establish the conditions for this switch. In Decision Problem 4, the objectively best-fitting model already assigns the higher posterior probability to Urn X. In Decision Problem 5, participants see two identical models, so the two model-specific posteriors coincide. Neither problem therefore allows a switch from a model with a lower posterior probability to one with a higher posterior probability. Preference-driven distortion may still affect Best-fit agents' beliefs through changes in model-specific posterior probabilities.[40]



Notes: The figure reports estimated effects of motivated reasoning on the reported beliefs of Urn X for each decision problem (left axis). Navy diamonds represent estimates for the pooled sample, and green squares represent estimates for participants classified as Best-fit. Decision problems are ordered by the switching threshold, shown by the grey bars (right axis). Decision Problem 4 appears last because no switch is possible. Decision Problem 5 is excluded because participants see identical models. The regressions include participant, decision-problem-by-signal, and round fixed effects. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level.

Figure 7: Motivated Reasoning Effects Across Decision Problems

Figure 7 reports the estimated effects of motivated reasoning in each decision problem separately. The decision problems are ordered by their switching threshold, $1/R$, discussed in Proposition 2. The switching threshold is the smallest preference-driven distortion toward Urn X that leads a Best-fit agent to select the model with the higher posterior probability of Urn X. A lower threshold means that a weaker preference is enough to switch models. In Decision Problem 4, the best-fitting model already implies the higher posterior probability of Urn X, so no switch is possible. Several Best-fit estimates are significantly positive, but they are imprecise and do not decline steadily as the threshold rises.[41]

[40] Decision Problems 1 and 4 have the same state prior, objective fit gap, and gap between model-specific posteriors. The best-fitting model implies the lower posterior probability of Urn X in Decision Problem 1 and the higher posterior probability in Decision Problem 4. Appendix B.3 compares their estimated effects to examine the role of model switching.

[41] State priors may affect how easily participants assess model fit. Decision Problems 1–3 share a state prior of 0.5, under which participants can determine each model's fit by counting the balls of the observed color in the

I next test Hypothesis 3 with regressions. Hypothesis 3 implies that the effect among Best-fit participants should be larger in problems with lower thresholds. I use two measures of how much preference-driven distortion is needed to switch models: the objective fit gap, $|P_A(s) - P_B(s)|$, and the switching threshold, $1/R$, discussed in Proposition 2 and used in Hypothesis 3. Using Decision Problems 1–3 and 6–9, I interact $\mathbb{1}_{\text{Asym}}$ with each measure in a separate regression. The regressions include participant, decision-problem-by-signal, and round fixed effects.[42] They also interact $\mathbb{1}_{\text{Asym}}$ with state-prior indicators, which allows the effect of motivated reasoning to vary across state priors.

Table 6: Motivated Reasoning, Model Fit, and Switching Thresholds

DV: Reported Belief	(1)	(2)	(3)	(4)
$\mathbb{1}_{\text{Asym}}$	0.035**	0.177***	0.034**	0.157**
	(0.015)	(0.063)	(0.015)	(0.064)
$\mathbb{1}_{\text{Asym}} \times$ objective fit gap	-0.054	-0.269		
	(0.044)	(0.168)		
$\mathbb{1}_{\text{Asym}} \times$ switching threshold			-0.004	-0.014
			(0.004)	(0.014)
Sample	All	Best-fit	All	Best-fit
Observations	8106	1386	8106	1386
Participants	193	33	193	33
Mean of DV	0.505	0.465	0.505	0.465
R^2	0.513	0.380	0.513	0.379

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with the objective fit gap and switching threshold. The dependent variable is reported belief. The sample includes Decision Problems 1–3 and 6–9. Columns (1) and (3) use all participants, while columns (2) and (4) use participants classified as Best-fit. The objective fit gap is $|P_A(s) - P_B(s)|$, and the switching threshold is $1/R$, discussed in Proposition 2. All specifications include interactions between $\mathbb{1}_{\text{Asym}}$ and state-prior indicators, together with participant, decision-problem-by-signal, and round fixed effects. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 6 reports negative interaction coefficients for both measures. The estimates are consistent with Hypothesis 3 directionally, although none of the interaction coefficients is statistically significant. With 33 Best-fit participants and seven decision problems, the design has limited power to detect how the effect of motivated reasoning varies across decision problems. Another possible reason is that preference-driven distortion differs across participants. A Best-fit participant with a strong distortion switches models in every decision problem, and one with a weak distortion never switches. Only participants with an intermediate distortion switch in decision problems with low thresholds but not in those with high thresholds. Differences in the average effect across decision problems therefore come only from this intermediate group,

cart. Among these three problems, the Best-fit estimate is largest in Decision Problem 1, which has the lowest threshold.

[42] Decision Problems 4–5 are excluded as no switch is possible.

which may be quite small.

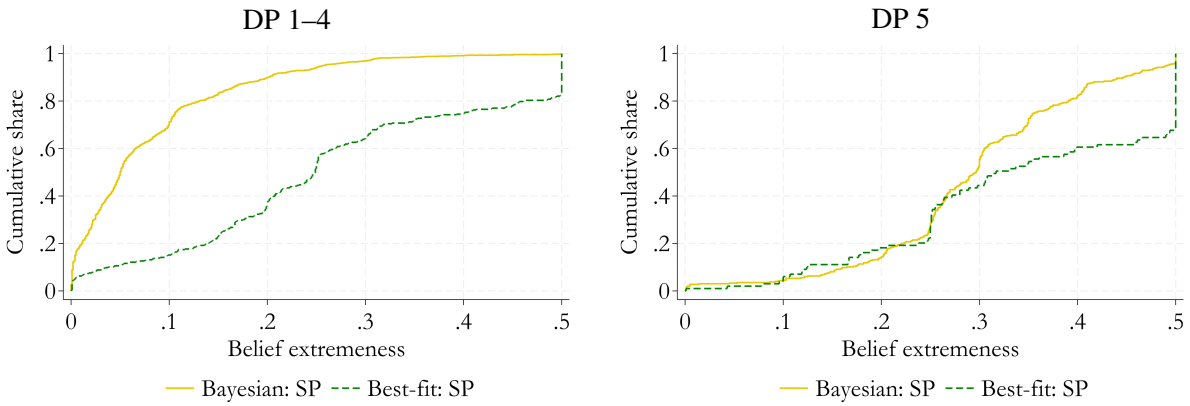
5.5 Belief Extremeness and Individual Characteristics

This section examines the extremeness of reported beliefs, measured by their absolute distance from 0.5 (Fan and Fries 2026). A reported belief of 0.1 or 0.9 has an extremeness of 0.4, while a reported belief of 0.5 has an extremeness of zero. In Decision Problems 1–4 and 6–7, the state priors are 0.5 and 0.3, and the Best-fit benchmark is further from 0.5 than the Bayesian benchmark. In Decision Problems 8–9, where the state prior is 0.7, the Best-fit benchmark instead predicts less extreme beliefs than the Bayesian benchmark. The Bayesian and Best-fit benchmarks coincide in Decision Problem 5. Across all nine decision problems in the *SP* block, mean belief extremeness is 0.292 among participants classified as Best-fit and 0.162 among those classified as Bayesian ($p < 0.001$). In Decision Problems 1–4, the corresponding means are 0.263 and 0.079 ($p < 0.001$).

Figure 8 shows the cumulative distributions of report-level extremeness from Decision Problems 1–5 in the *SP* block. In Decision Problems 1–4, the Best-fit rule generates more extreme benchmark beliefs than the Bayesian rule. In Decision Problem 5, both rules produce the same benchmark. Best-fit participants report higher extremeness levels in both Decision Problems 1–4 ($p < 0.001$) and Decision Problem 5 ($p = 0.038$), while the difference in mean belief extremeness between Best-fit and Bayesian participants is significantly larger in Decision Problems 1–4 than in Decision Problem 5 ($p < 0.001$). One may be concerned that the difference in belief extremeness is driven entirely by the parameters of the decision problems. The larger gap in Decision Problems 1–4 suggests that differences between the rules' predictions help explain why Best-fit participants report more extreme beliefs. However, Best-fit participants also report more extreme beliefs on average in Decision Problem 5, where both rules predict the same beliefs. This finding suggests that the difference in belief extremeness is not solely driven by benchmark predictions.

I also explore which individual characteristics are associated with behavioral type classification. Table 7 reports average marginal effects from separate logit regressions for Bayesian and Best-fit types. The regressions include gender, race indicators, cognitive score, risk attitude, overconfidence, narrative extremeness, quiz score, and previous experimental participation. Narrative extremeness, defined in Section 4.4, comes from the unincentivized narrative survey at the end of the experiment.

The regression results show that narrative extremeness is positively associated with being classified as the Best-fit type in the belief updating tasks in this experiment. The corresponding association with Bayesian classification is negative, and both associations are statistically significant at the 5% level. Risk attitude is positively associated with Bayesian classification



Notes: The curves show the cumulative distributions of belief extremeness, $|\text{Belief}_{it} - 0.5|$, in the *SP* block for participants classified as Bayesian and Best-fit. The left panel uses Decision Problems 1–4, and the right panel uses Decision Problem 5.

Figure 8: Belief Extremeness by Type in the Symmetric Payoff Block

Table 7: Individual Characteristics and Type Classification

DV: Type Classification	Bayesian	Best-fit
	(1)	(2)
Female	0.032	-0.004
	(0.066)	(0.054)
Quiz Score	0.056	0.025
	(0.040)	(0.037)
Cognitive Score	0.010	-0.008
	(0.028)	(0.023)
Risk Attitude	0.005***	-0.001
	(0.002)	(0.001)
Overconfidence	-0.075	-0.034
	(0.080)	(0.064)
Narrative Extremeness	-0.094**	0.067**
	(0.037)	(0.029)
Observations	193	193
Mean of DV	0.684	0.171
Pseudo R^2	0.109	0.121

Notes: The table reports average marginal effects from binary logit regressions. The dependent variables are binary indicators for Bayesian and Best-fit types in columns (1) and (2). Each regression uses all 193 participants. All displayed characteristics enter both regressions. The regressions include gender, race indicators, cognitive score, risk attitude, overconfidence, narrative extremeness, quiz score, and previous experimental participation. Delta-method standard errors appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

($p < 0.01$).[43] The estimated coefficients of quiz score, cognitive score, and overconfidence are not statistically significant for either type.

This finding suggests that the behavioral rules participants use in the laboratory may sometimes reflect tendencies that extend beyond the abstract belief-updating tasks. Participants classified as Best-fit also tend to express stronger views on economic narratives. However, narrative extremeness is measured through an unincentivized survey. This association therefore provides weak evidence of a broader connection. See Appendix B.6.1 for details on type-assignment regressions including Non-movers and associations between narrative extremeness and belief updating.

5.6 Further Discussion

The results connect differences in belief updating and model selection to differences in responses to asymmetric payoffs. Asymmetric payoffs increase reported beliefs in the preferred state, which supports Hypothesis 1. This effect is significantly larger among participants classified as Best-fit than among those classified as Bayesian, which supports Hypothesis 2.

The results do not provide clear support for Hypothesis 3. Among Best-fit participants, the point estimates suggest a larger effect of motivated reasoning when the objective fit gap is smaller or the switching threshold is lower, but these estimates are imprecise and never statistically significant. The objective fit gap may not be the only determinant of Best-fit participants' model-switching behavior. Salience, complexity, or model-specific posteriors might also affect this process. In addition, Hypothesis 3 tests a further link between model switching and the effect of motivated reasoning. Preference-driven distortion affects Best-fit agents through two channels: model selection and model-specific posterior beliefs. The second channel still operates when model switching is less frequent. Together, these issues limit how clearly I can test this hypothesis.

I also explore the extremeness of reported beliefs across behavioral rules and find that Best-fit participants report more extreme beliefs on average. I further explore the potential connection between behavior in this belief updating task and behavior in broader settings. The results suggest that participants classified as Best-fit tend to hold stronger opinions on economic policy narratives.

[43] This association is economically modest. A one-standard-deviation increase in the number of boxes collected in the bomb risk task (about 21 boxes) raises the probability of Bayesian classification by about 10 percentage points. Hedging against the urn payment is also unlikely to explain this association. The classification uses only beliefs from the *SP* block, where the urn payment is the same for both urns, so participants have little reason to hedge. In the *AP* block, hedging could lower reported beliefs of Urn X, especially among more risk-averse participants. Participants classified as Best-fit are more risk-averse on average, so hedging would, if anything, reduce their estimated effect of motivated reasoning relative to Bayesian participants.

6 Conclusion

This paper examines how preferences over payoff-relevant states affect model selection and belief updating when multiple competing models can explain the same information. I develop a theoretical framework in which preference-driven distortion affects how individuals perceive the information structure. The same distortion applies to Bayesian and Best-fit behavioral rules, and it distorts Bayesian agents' model weights and may switch Best-fit agents' model selection. In the experimental setting with two states, two signals, and two models, stronger bias shifts beliefs toward the preferred state under both rules. Bayesian beliefs change continuously as the distortion increases. Best-fit beliefs can also jump when individuals switch to a model that assigns a higher probability to the preferred state. This theoretical framework characterizes when these model switches occur and their dependence on the information structure.

The experiment shows that asymmetric payoffs have a larger effect on participants classified as best-fit updaters than on those classified as Bayesian updaters. Behavior in the symmetric payoff block therefore helps predict whose beliefs are more affected by preferences. This finding suggests that differences in model selection and belief updating matter for understanding motivated reasoning. Even when individuals face the same information and competing explanations, their responses to preferences can differ substantially.

The framework offers one explanation for this difference. Preferences can change how well a model appears to explain the data, making a model that favors the preferred state more likely to be selected. Individuals may therefore arrive at a preferred belief while following what they perceive as the best explanation of the evidence. The experimental results are consistent with this explanation, although the variation across decision problems predicted by Hypothesis 3 is not statistically significant. However, the experiment elicits only beliefs, so I cannot directly measure how much of the effect of motivated reasoning comes from model switches. Eliciting model choices alongside beliefs would help separate model switches from changes in model-specific posteriors within the selected model.

These results have implications for efforts to reduce biased beliefs by presenting competing explanations. Both models are available in the experiment, and individuals differ in whether they weight them or select one model to update beliefs. The effectiveness of providing alternative explanations may therefore depend on how individuals use them when updating beliefs. A useful next step is to examine whether encouraging individuals to weight multiple models could reduce the effect of motivated reasoning.

A further question is when individuals adopt different updating rules. The experiment uses simple models that participants can compare directly. In many settings, individuals face several complex models whose assumptions and predictions are difficult to assess. Comparing and evaluating these models may require considerable time and effort. As a result, individuals

may then rely on a single model, and their preferences may influence which model they select. Future experiments could vary the number of models or the complexity of the model. This would help us understand whether individuals change their updating rules across settings and whether these changes affect the magnitude of motivated reasoning.

The supply of competing models also deserves further attention. In this experiment, the models are given by the experimenter, and their availability does not depend on participants' preferences. In political and financial settings, interested parties often choose which explanations to present. A sender who knows the receiver's preferences may offer a model that supports the receiver's preferred state while fitting the observed data. The optimal sending strategy may also depend on whether the receiver weights competing models or selects one. Future research could allow senders to choose models and explore how senders respond to differences in receivers' preferences and updating behavior, and whether their choices strengthen the effect of motivated reasoning.

References

Aina, Chiara. 2025. "Tailored stories." Working Paper.

Aina, Chiara, and Florian Schneider. 2025. "Weighting competing models." CESifo Working Paper 11757.

Alós-Ferrer, Carlos, and Michele Garagnani. 2023. "Part-time Bayesians: incentives and behavioral heterogeneity in belief updating." Management Science, 69(9): 5523–5542.

Ambuehl, Sandro, and Heidi C Thysen. 2025. "Choice with Competing Models: An Experimental Study." Working Paper, University of Zurich.

Andre, Peter, Ingar Haaland, Christopher Roth, Mirko Wiederholt, and Johannes Wohlfart. 2026. "Narratives about the Macroeconomy." Review of Economic Studies, rdag014.

Angrisani, Marco, Anya Samek, and Ricardo Serrano-Padial. 2024. "Competing Narratives in Action: An Empirical Analysis of Model Adoption Dynamics." Working Paper.

Avtar, Ruchi, and Alex Ballyk. 2026. "Attribution Errors: Context-Dependence and Choice Distortions." Unpublished Manuscript.

Babcock, Linda, George Loewenstein, Samuel Issacharoff, and Colin Camerer. 1995. "Biased judgments of fairness in bargaining." The American Economic Review, 85(5): 1337–1343.

Ba, Cuimin. 2026. "Robust Misspecified Models." American Economic Review, 116(4): 1340–79.

Ba, Cuimin, J Aislinn Bohren, and Alex Imas. 2024. "Over-and underreaction to information." Available at SSRN 4274617.

Barron, Kai. 2021. "Belief updating: does the 'good-news, bad-news' asymmetry extend to purely financial domains?" Experimental Economics, 24(1): 31–58.

Barron, Kai, and Tilman Fries. 2024a. "Narrative persuasion." Berlin School of Economics Discussion Papers 0039.

Barron, Kai, and Tilman Fries. 2024b. "Narrative Persuasion: A Brief Introduction." In Encyclopedia of Experimental Social Science.

Barron, Kai, Anna Becker, and Steffen Huck. 2025. "Motivated political reasoning: On the emergence of belief-value constellations." European Economic Review, 172: 104929.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2009. "Dynamic psychological games." Journal of Economic Theory, 144(1): 1–35.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2022. "Belief-dependent motivations and psychological game theory." Journal of Economic Literature, 60(3): 833–882.

Battigalli, Pierpaolo, Roberto Corrao, and Martin Dufwenberg. 2019. "Incorporating belief-dependent motivation in games." Journal of Economic Behavior & Organization, 167: 185–218.

Bénabou, Roland, and Jean Tirole. 2016. "Mindful economics: The production, consumption, and value of beliefs." Journal of Economic Perspectives, 30(3): 141–164.

Benjamin, Daniel J. 2019. "Errors in probabilistic reasoning and judgment biases." Handbook of Behavioral Economics: Applications and Foundations 1, 2: 69–186.

Bland, James, and Yaroslav Rosokha. 2024. "Rounding the (Non) Bayesian Curve: Unraveling the Effects of Rounding Errors in Belief Updating." Purdue University Economics Working Papers 1353.

Bland, James R, and Yaroslav Rosokha. 2021. "Learning under uncertainty with multiple priors: experimental investigation." Journal of Risk and Uncertainty, 62(2): 157–176.

Bohren, J Aislinn, and Daniel N Hauser. 2021. "Learning with heterogeneous misspecified models: Characterization and robustness." Econometrica, 89(6): 3025–3077.

Bolte, Lukas, and Tony Q Fan. 2024. "Motivated mislearning: The case of correlation neglect." Journal of Economic Behavior & Organization, 217: 647–663.

Brunnermeier, Markus K, and Jonathan A Parker. 2005. "Optimal expectations." American Economic Review, 95(4): 1092–1118.

Cason, Timothy N, Stanton Hudja, and Yaroslav Rosokha. 2025. "Innovation and Information Free-Riding: An Experimental Investigation of Collective Experimentation." Available at SSRN 5253468.

Chambers, Christopher P, Yusufcan Masatlioglu, and Collin Raymond. 2023. "Coherent distorted beliefs." arXiv preprint arXiv:2310.09879.

Charles, Constantin, and Chad Kendall. 2025. "Causal narratives." Available at SSRN 4669371.

Charness, Gary, and Chetan Dave. 2017. "Confirmation bias with motivated beliefs." Games and Economic Behavior, 104: 1–23.

Chen, Daniel L, Martin Schonger, and Chris Wickens. 2016. "oTree—An open-source platform for laboratory, online, and field experiments." Journal of Behavioral and Experimental Finance, 9: 88–97.

Conlon, John J. 2026. "Memory rehearsal and belief biases." Available at SSRN 7050598.

Coutts, Alexander. 2019. "Good news and bad news are still news: Experimental evidence on belief updating." Experimental Economics, 22(2): 369–395.

Coutts, Alexander, Leonie Gerhards, and Zahra Murad. 2024. "What to blame? Self-serving attribution bias with multi-dimensional uncertainty." The Economic Journal, 134(661): 1835–1874.

Crosetto, Paolo, and Antonio Filippin. 2013. "The "bomb" risk elicitation task." Journal of Risk and Uncertainty, 47: 31–65.

Danz, David, Lise Vesterlund, and Alistair J Wilson. 2022. "Belief elicitation and behavioral incentive compatibility." American Economic Review, 112(9): 2851–2883.

Danz, David, Lise Vesterlund, and Alistair J Wilson. 2024. "Evaluating Behavioral Incentive Compatibility: Insights from Experiments." Journal of Economic Perspectives, 38(4): 131–154.

Dominiak, Adam, Matthew Kovach, and Gerelt Tserenjigmid. 2023. "Inertial updating." arXiv preprint arXiv:2303.06336.

Dominiak, Adam, Matthew Kovach, and Gerelt Tserenjigmid. 2025. "Inertial updating with general information." arXiv preprint arXiv:2502.00958.

Drobner, Christoph. 2022. "Motivated beliefs and anticipation of uncertainty resolution." American Economic Review: Insights, 4(1): 89–105.

Drobner, Christoph, and Sebastian J Goerg. 2024. "Motivated belief updating and rationalization of information." Management Science, 70(7): 4583–4592.

Eil, David, and Justin M Rao. 2011. "The good news-bad news effect: asymmetric processing of objective information about yourself." American Economic Journal: Microeconomics, 3(2): 114–138.

El-Gamal, Mahmoud A, and David M Grether. 1995. "Are people Bayesian? Uncovering behavioral strategies." Journal of the American Statistical Association, 90(432): 1137–1145.

Eliaz, Kfir, and Ran Spiegler. 2020. "A model of competing narratives." American Economic Review, 110(12): 3786–3816.

Eliaz, Kfir, and Ran Spiegler. 2026. "News media as suppliers of narratives (and information)." Theoretical Economics, 21(3): 928–972.

Engelmann, Jan B, Maël Lebreton, Nahuel A Salem-Garcia, Peter Schwardmann, and Joël J van der Weele. 2024. "Anticipatory anxiety and wishful thinking." American Economic Review, 114(4): 926–960.

Enke, Benjamin, Thomas Graeber, and Ryan Oprea. 2023. "Confidence, self-selection, and bias in the aggregate." American Economic Review, 113(7): 1933–1966.

Epstein, Larry G, and Martin Schneider. 2007. "Learning under ambiguity." The Review of Economic Studies, 74(4): 1275–1303.

Epstein, Larry G, and Yoram Halevy. 2024. "Hard-to-interpret signals." Journal of the European Economic Association, 22(1): 393–427.

Ertac, Seda. 2011. "Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback." Journal of Economic Behavior & Organization, 80(3): 532–545.

Esponda, Ignacio, Emanuel Vespa, and Sevgi Yuksel. 2024. "Mental models and learning: The case of base-rate neglect." American Economic Review, 114(3): 752–782.

Fan, Tony Q. 2024. "Forming Misspecified Mental Models: The Power of Choice." Available at SSRN 4497643.

Fan, Tony Q, and Tilman Fries. 2026. "Narratives, Belief Movements, and Economic Fluctuations." Available at SSRN 6603858.

Farina, Agata, and Clément Herman. 2026. "Communicating with Data-Generating Processes: An Experimental Analysis." Available at SSRN 6482578.

Filiz-Ozbay, Emel, and Zhenxun Liu. 2026. "Information manipulation of wishful thinkers." Games and Economic Behavior, 158: 763–786.

Fréchette, Guillaume R, Emanuel Vespa, and Sevgi Yuksel. 2024. "Extracting Models From Data Sets: An Experiment." Available at SSRN 5026751.

Fryer Jr, Roland G, Philipp Harms, and Matthew O Jackson. 2019. "Updating beliefs when evidence is open to interpretation: Implications for bias and polarization." Journal of the European Economic Association, 17(5): 1470–1501.

Fudenberg, Drew, and Giacomo Lanzani. 2023. "Which misspecifications persist?" Theoretical Economics, 18(3): 1271–1315.

Gaines, Brian J, James H Kuklinski, Paul J Quirk, Buddy Peyton, and Jay Verkuilen. 2007. "Same facts, different interpretations: Partisan motivation and opinion on Iraq." The Journal of Politics, 69(4): 957–974.

Gilboa, Itzhak, and David Schmeidler. 1993. "Updating ambiguous beliefs." Journal of Economic Theory, 59(1): 33–49.

Greiner, Ben. 2015. "Subject pool recruitment procedures: organizing experiments with ORSEE." Journal of the Economic Science Association, 1(1): 114–125.

Grether, David M. 1980. "Bayes rule as a descriptive model: The representativeness heuristic." The Quarterly Journal of Economics, 95(3): 537–557.

Hastorf, Albert H, and Hadley Cantril. 1954. "They saw a game; a case study." The Journal of Abnormal and Social Psychology, 49(1): 129–134.

Healy, Paul J, and Greg Leo. 2025. "Belief elicitation: a user's guide." In Handbook of Experimental Methodology. Vol. 1, 81–162. North-Holland.

He, Simin, Sherry Xin Li, Junze Sun, and Xinlu Zou. 2026. "Consensus and Truth-Finding in Narrative Communication." Working Paper.

Hossain, Tanjim, and Ryo Okui. 2013. "The binarized scoring rule." Review of Economic Studies, 80(3): 984–1001.

Huffman, David, Collin Raymond, and Julia Shvets. 2022. "Persistent overconfidence and biased memory: Evidence from managers." American Economic Review, 112(10): 3141–3175.

Hu, Yanshuo, and Stella Papadokonstantaki. 2025. "Motivated Beliefs: The Role of Uncertainty about Information Accuracy." Available at SSRN 5268027.

Jain, Atulya. 2023. "Informing agents amidst biased narratives." Working Paper.

Kendall, Chad, and Ryan Oprea. 2024. "On the complexity of forming mental models." Quantitative Economics, 15(1): 175–211.

Kimball, Miles S, Collin B Raymond, Jiannan Zhou, Junya Zhou, Fumio Ohtake, and Yoshiro Tsutsui. 2024. "Happiness dynamics, reference dependence, and motivated beliefs in US presidential elections." NBER Working Paper 32078.

Kovach, Matthew. 2020. "Twisting the truth: Foundations of wishful thinking." Theoretical Economics, 15(3): 989–1022.

Kovach, Matthew. 2024. "Ambiguity and partial Bayesian updating." Economic Theory, 78(1): 155–180.

Kovach, Matthew, Daniel Martin, and Gerelt Tserenjigmid. 2026. "Learning from an Unknown DGP: Experimental Evidence on Belief Updating with AI Recommendations." arXiv preprint arXiv:2607.10460.

Kunda, Ziva. 1990. "The case for motivated reasoning." Psychological Bulletin, 108(3): 480–498.

Le Yaouanq, Yves, Peter Schwardmann, and Joël J. van der Weele. 2025. "Rationalizations and Political Polarization." CESifo Working Paper 11897.

Liang, Yucheng. 2025. "Learning from unknown information sources." Management Science, 71(5): 3873–3890.

Liu, Manwei, and Sili Zhang. 2026. "Counteracting Narratives: Evidence from An Online Experiment." The Economic Journal, 136(673): 125–162.

Liu, Zhenxun. 2024. "Essays on Information and Non-Bayesian Beliefs." PhD diss. University of Maryland, College Park.

Liu, Zhenxun. 2026. "When Confirmation Bias Meets Motivated Reasoning: A Unified Perspective." Working Paper.

Lord, Charles G, Lee Ross, and Mark R Lepper. 1979. "Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence." Journal of Personality and Social Psychology, 37(11): 2098–2109.

Mayraz, Guy. 2011. "Wishful thinking." Available at SSRN 1955644.

Möbius, Markus M, Muriel Niederle, Paul Niehaus, and Tanya S Rosenblat. 2022. "Managing self-confidence: Theory and experimental evidence." Management Science, 68(11): 7793–7817.

Montiel Olea, José Luis, Pietro Ortoleva, Mallesh M Pai, and Andrea Prat. 2022. "Competing models." The Quarterly Journal of Economics, 137(4): 2419–2457.

Moreno, Othon M, and Yaroslav Rosokha. 2016. "Learning under compound risk vs. learning under ambiguity–an experiment." Journal of Risk and Uncertainty, 53: 137–162.

Musolff, Robin, and Florian Zimmermann. 2025. "Model uncertainty." CESifo Working Paper 12041.

Ngangoué, M Kathleen. 2021. "Learning under ambiguity: An experiment in gradual information processing." Journal of Economic Theory, 195: 105282.

Oprea, Ryan, and Sevgi Yuksel. 2022. "Social exchange of motivated beliefs." Journal of the European Economic Association, 20(2): 667–699.

Qiao, Zhongheng. 2025. "Sunk-cost, Responsibility, and Belief." Available at SSRN 5987694.

Rabin, Matthew, and Joel L Schrag. 1999. "First impressions matter: A model of confirmatory bias." The Quarterly Journal of Economics, 114(1): 37–82.

Roos, Michael, and Matthias Reccius. 2024. "Narratives in economics." Journal of Economic Surveys, 38(2): 303–341.

Savage, Leonard J. 1971. "Elicitation of personal probabilities and expectations." Journal of the American Statistical Association, 66(336): 783–801.

Schwardmann, Peter, and Joël J van der Weele. 2019. "Deception and self-deception." Nature Human Behaviour, 3(10): 1055–1061.

Schwardmann, Peter, Egon Tripodi, and Joël J van der Weele. 2022. "Self-persuasion: Evidence from field experiments at international debating competitions." American Economic Review, 112(4): 1118–1146.

Schwartzstein, Joshua, and Adi Sunderam. 2021. "Using models to persuade." American Economic Review, 111(1): 276–323.

Schwartzstein, Joshua, and Adi Sunderam. 2022. "Shared models in networks, organizations, and groups." NBER Working Paper 30642.

Shiller, Robert J. 2017. "Narrative economics." American Economic Review, 107(4): 967–1004.

Shishkin, Denis, and Pietro Ortoleva. 2023. "Ambiguous information and dilation: An experiment." Journal of Economic Theory, 208: 105610.

Suleymanov, Elchin. 2025. "Robust maximum likelihood updating." arXiv preprint arXiv:2504.17151.

Thaler, Michael. 2021. "The supply of motivated beliefs." arXiv Preprint arXiv:2111.06062.

Thaler, Michael. 2024. "The fake news effect: Experimentally identifying motivated reasoning using trust in news." American Economic Journal: Microeconomics, 16(2): 1–38.

Thaler, Michael. 2026. "Good news is not a sufficient condition for motivated reasoning." Review of Economics and Statistics, 1–24.

Wang, Zhaoqi, and Vivian Juehui Zheng. 2026. "Motivated inference of ambiguous information." Journal of Behavioral and Experimental Finance, 101213.

Wilson, Alistair, and Emanuel Vespa. 2018. "Paired-uniform scoring: Implementing a binarized scoring rule with non-mathematical language." Working Paper.

Yang, Jeffrey. 2023. "A Criterion of Model Decisiveness." Available at SSRN 4425088.

Zimmermann, Florian. 2020. "The dynamics of motivated beliefs." American Economic Review, 110(2): 337–363.

A Further Theoretical Analysis

This appendix derives several identities used in Section 3 and proves each proposition and prediction. The assumptions specified in Sections 3.1 and 3.2 apply throughout.

A.1 Model Fit

For every model, objective and subjective fits satisfy

$$\sum_{s\in S} P_m(s) = 1, \qquad \sum_{s\in S} \hat{P}_m(s) = \sum_{\omega\in\Omega} \mu_0(\omega) f(U(\omega)). \tag{A-1}$$

Both equalities follow by exchanging the state and signal sums and using $\sum_{s\in S} \pi_m(s \mid \omega) = 1$. Their right-hand sides are independent of the model.

Suppose $S = \{s_1, s_2\}$. Subtracting Equation (A-1) for models m and m' gives Equation (11). This derivation allows any finite number of states.

Taking the ratio of the weights assigned to m and m', before and after preference-driven distortion, and using Equation (7) gives

$$\frac{\hat{\rho}(m \mid s)}{\hat{\rho}(m' \mid s)} = \frac{\rho(m \mid s)}{\rho(m' \mid s)} \frac{\sum_{\omega\in\Omega} \mu_m(\omega \mid s) f(U(\omega))}{\sum_{\omega\in\Omega} \mu_{m'}(\omega \mid s) f(U(\omega))}.$$

Preference-driven distortion increases the weight on m relative to m' exactly when the posterior expectation of $f(U)$ is larger under m.

A related observation holds with any finite signal space. Suppose $M = \{m, m'\}$ and $\pi_m(s \mid \omega) = \pi_{m'}(s \mid \omega)$ for every $s \in S$ and $\omega \neq \bar{\omega}$. Define $\delta(s) = \pi_m(s \mid \bar{\omega}) - \pi_{m'}(s \mid \bar{\omega})$. Then

$$P_m(s) - P_{m'}(s) = \mu_0(\bar{\omega})\delta(s), \qquad \hat{P}_m(s) - \hat{P}_{m'}(s) = f(U(\bar{\omega}))\big[P_m(s) - P_{m'}(s)\big].$$

The objective and subjective fit differences therefore have the same sign after every signal. Preference-driven distortion cannot change the model selected by a best-fit agent.

A.2 Proofs

A.2.1 Proof of Proposition 1

Proof. Suppose m is more representative than m' for s. The subjective fit difference is

$$\hat{P}_m(s) - \hat{P}_{m'}(s) = \sum_{\omega \in \Omega} \mu_0(\omega) f(U(\omega)) \big[\pi_m(s \mid \omega) - \pi_{m'}(s \mid \omega)\big].$$

Every term is nonnegative. At least one is positive because $\mu_0(\omega) > 0$, $f(U(\omega)) > 0$, and one likelihood inequality is strict. Hence, $\hat{P}_m(s) > \hat{P}_{m'}(s)$, which proves part 1.

The common denominator in Equation (10) cancels from the ratio of model weights. Therefore,

$$\frac{\hat{\rho}(m \mid s)}{\hat{\rho}(m' \mid s)} = \frac{\rho_0(m)}{\rho_0(m')} \frac{\hat{P}_m(s)}{\hat{P}_{m'}(s)} > \frac{\rho_0(m)}{\rho_0(m')}.$$

This proves part 2. Part 1 also implies that m' cannot maximize subjective fit, which proves part 3. □

A.2.2 Proof of Prediction 1

Proof. By assumption, the objective and subjective fit differences between m and m' are both nonzero. Equation (11) implies that each difference changes sign across the two signals.

Suppose the selections based on objective and subjective fit differ after s_1. Both fit rankings reverse for s_2, so the selections also differ after s_2. The switch is therefore full.

If the selections agree after s_1, both ranking reversals make them agree after s_2. The selected model is then unchanged after both signals. □

A.2.3 Proof of Proposition 2

Proof. Because S has two signals and the models are distinct, their likelihoods of s_1 differ in at least one state. If they were equal in the other state, one model would be more representative. Thus, the likelihood differences are nonzero and have opposite signs.

To determine their signs, suppose that

$$\pi_m(s_1 \mid \omega_1) > \pi_{m'}(s_1 \mid \omega_1), \qquad \pi_m(s_1 \mid \omega_2) < \pi_{m'}(s_1 \mid \omega_2).$$

Bayes' rule would imply

$$\frac{\mu_m(\omega_1 \mid s_1)}{\mu_m(\omega_2 \mid s_1)} > \frac{\mu_{m'}(\omega_1 \mid s_1)}{\mu_{m'}(\omega_2 \mid s_1)}.$$

With two states, the posterior odds of ω_1 relative to ω_2 are strictly increasing in the posterior probability of ω_1. The displayed inequality contradicts $\mu_m(\omega_1 \mid s_1) < \mu_{m'}(\omega_1 \mid s_1)$. Hence,

$$\pi_m(s_1 \mid \omega_1) < \pi_{m'}(s_1 \mid \omega_1), \qquad \pi_m(s_1 \mid \omega_2) > \pi_{m'}(s_1 \mid \omega_2).$$

Equation (12) therefore gives $A > 0$ and $B > 0$.

The objective fit difference is

$$P_m(s_1) - P_{m'}(s_1) = B - A > 0.$$

Thus, $B > A > 0$ and $R = A/B \in (0,1)$.

The subjective fit difference equals

$$\hat{P}_m(s_1) - \hat{P}_{m'}(s_1) = f(U(\omega_2))B - f(U(\omega_1))A = f(U(\omega_2))(B - tA).$$

It is negative exactly when

$$t > \frac{B}{A} = \frac{1}{R}.$$

This condition reverses the selected model after s_1. Prediction 1 then implies a full switch. Below the threshold, the selections agree after s_1, so no switch occurs.

At equality, subjective fits tie after both signals. The fixed tie-breaking rule selects the same model after both signals. Objective selections differ across signals, so the switch is partial. □

A.2.4 Proof of Proposition 3

Proof. Fix $s \in S$. I first derive Equation (13). Let

$$D(s) = \sum_{m \in M} \rho_0(m) P_m(s), \qquad \hat{D}(s) = \sum_{m \in M} \rho_0(m) \hat{P}_m(s).$$

Equations (3), (4), and (1a) give

$$\mu_B(\omega \mid s) = \frac{\mu_0(\omega) \sum_{m \in M} \rho_0(m) \pi_m(s \mid \omega)}{D(s)}.$$

Likewise, Equations (5) and (10) give

$$\hat{\mu}_B(\omega \mid s) = \frac{\mu_0(\omega) f(U(\omega)) \sum_{m \in M} \rho_0(m) \pi_m(s \mid \omega)}{\hat{D}(s)}.$$

Using the definition of $\hat{D}(s)$ and rearranging the sums gives

$$\begin{aligned}\hat{D}(s) &= \sum_{\omega' \in \Omega} \mu_0(\omega') f(U(\omega')) \sum_{m \in M} \rho_0(m) \pi_m(s \mid \omega') \\ &= D(s) \sum_{\omega' \in \Omega} \mu_B(\omega' \mid s) f(U(\omega')).\end{aligned}$$

Substituting this equality into the expression for $\hat{\mu}_B$ gives Equation (13).

Now write $x_m = \mu_m(\omega_1 \mid s)$ and $t = f(U(\omega_1))/f(U(\omega_2)) > 1$. Full support implies $x_m \in (0,1)$ and $\mu_B(\omega_1 \mid s) \in (0,1)$. Equation (13) gives

$$\hat{\mu}_B(\omega_1 \mid s) = h(t, \mu_B(\omega_1 \mid s)), \qquad h(t,x) = \frac{tx}{tx + (1-x)}.$$

For $t > 1$ and $x \in (0,1)$,

$$h(t,x) - x = \frac{(t-1)x(1-x)}{1+(t-1)x} > 0.$$

Therefore, $\hat{\mu}_B(\omega_1 \mid s) > \mu_B(\omega_1 \mid s)$.

Let $j = m_s^*$ and $k = \hat{m}_s^*$. Equation (7) becomes

$$\hat{P}_m(s) = P_m(s) f(U(\omega_2)) \big[1 + (t-1)x_m\big].$$

Optimality gives $P_j(s) \geq P_k(s)$ and $\hat{P}_k(s) \geq \hat{P}_j(s)$. If $x_k < x_j$, then

$$\begin{aligned}\hat{P}_k(s) &= P_k(s) f(U(\omega_2)) \big[1 + (t-1)x_k\big] \\ &\leq P_j(s) f(U(\omega_2)) \big[1 + (t-1)x_k\big] \\ &< P_j(s) f(U(\omega_2)) \big[1 + (t-1)x_j\big] \\ &= \hat{P}_j(s).\end{aligned}$$

This contradicts the selection of k, so $x_k \geq x_j$. Hence,

$$\mu_{\hat{m}_s^*}(\omega_1 \mid s) \geq \mu_{m_s^*}(\omega_1 \mid s).$$

Finally,

$$\hat{\mu}_F(\omega_1 \mid s) = h(t, x_k) > x_k \geq x_j = \mu_F(\omega_1 \mid s).$$

This proves both claims for best-fit agents. □

A.2.5 Proof of Proposition 4

Proof. Fix $s \in S$. Multiplying all values of $f(U(\omega))$ by the same positive constant leaves posterior beliefs, model weights, and model selection unchanged. Set $f(U(\omega_2)) = 1$ and

$f(U(\omega_1)) = t$.

The function

$$h(t,x) = \frac{tx}{tx+(1-x)}$$

is continuous. For $t > 0$ and $x \in (0,1)$,

$$\frac{\partial h}{\partial t} = \frac{x(1-x)}{[tx+(1-x)]^2} > 0, \qquad \frac{\partial h}{\partial x} = \frac{t}{[tx+(1-x)]^2} > 0.$$

Equation (13) gives

$$\hat{\mu}_B(\omega_1 \mid s) = h(t, \mu_B(\omega_1 \mid s)).$$

Full support places $\mu_B(\omega_1 \mid s)$ in $(0,1)$. This proves part 1.

For the best-fit rule, define $x_m = \mu_m(\omega_1 \mid s)$ and

$$q_m(t) = P_m(s)\big[1 + (t-1)x_m\big].$$

Equation (7) gives $\hat{P}_m(s) = q_m(t)$, so the selected model maximizes $q_m(t)$. Fix $1 < t_1 < t_2$, and let models a and b be selected at these values.

Optimality gives $q_a(t_1) \geq q_b(t_1)$ and $q_b(t_2) \geq q_a(t_2)$. Combining these inequalities gives

$$\frac{1+(t_2-1)x_b}{1+(t_1-1)x_b} \geq \frac{1+(t_2-1)x_a}{1+(t_1-1)x_a}.$$

For fixed t_1 and t_2, the ratio on each side is strictly increasing in x, with derivative

$$\frac{t_2 - t_1}{[1+(t_1-1)x]^2} > 0.$$

Therefore, $x_b \geq x_a$.

The model-specific posterior probability on ω_1 under the selected model is therefore non-decreasing in t. It remains constant whenever the selected model does not change. This proves part 2.

The distorted best-fit posterior belief at t_1 is $h(t_1, x_a)$, while the belief at t_2 is $h(t_2, x_b)$. Because $x_b \geq x_a$,

$$h(t_2, x_b) \geq h(t_2, x_a) > h(t_1, x_a).$$

The strict inequality follows because $x_a \in (0,1)$. Thus, the distorted best-fit posterior belief is strictly increasing in t.

Each $q_m(t)$ is affine in t, and two distinct affine functions intersect at most once. Since M is finite, only finitely many switching thresholds exist. Identical functions do not create switches because the fixed tie-breaking rule selects the same model for every t.

Between switching thresholds, the selected model is fixed. The distorted best-fit posterior belief is therefore continuous on each such interval. A switch to a model with a strictly higher model-specific posterior probability on ω_1 produces an upward jump. This proves part 3. □

B Further Empirical Analysis

B.1 Sample Characteristics

Table B-1 summarizes the characteristics of the 193 participants.

	count	mean	sd	min	max
Female	193	0.497	0.501	0	1
Asian	193	0.415	0.494	0	1
White	193	0.472	0.500	0	1
Cognitive Score	193	2.751	1.283	0	6
Prior Experiments	193	3.710	3.072	0	17
Quiz Score	193	9.508	0.758	7	10

Notes: The table reports summary statistics for the 193 participants. Means of indicator variables represent participant proportions. Quiz scores measure comprehension of the experimental instructions.

Table B-1: Sample Characteristics

B.2 Structural Classification

This subsection presents the error structure and estimation procedure for the classification in Section 5.2. The sample contains 5,211 reported beliefs from 193 participants, with 27 beliefs per participant in the *SP* block. Each participant follows one behavioral rule across these scenarios. Let $k \in \mathscr{T} = \{\text{Bayesian}, \text{Best-fit}, \text{Non-mover}\}$ index behavioral types, and let t index rounds.

First layer: errors in model-fit calculations. For each decision problem, the objective fit of model m after signal s is

$$P_m(s) = \sum_{\omega \in \Omega} \mu_0(\omega)\pi_m(s \mid \omega). \tag{B-2}$$

Participant i of type k assesses this fit with a mean-zero normal error:

$$\hat{P}_{mit}(s) = P_m(s) + \varepsilon^1_{mit}, \qquad \varepsilon^1_{mit} \sim N(0, \sigma^2_{1,k}). \tag{B-3}$$

The errors are independent across models, participants, and rounds. Their standard deviation is common to the Bayesian and Best-fit types, $\sigma_{1,k} = \sigma_1$. The model prior assigns equal probabilities to carts A and B. Bayesian agents therefore weight the model-specific posteriors by their relative subjective fits. For positive perceived fits, the weight on model A is[44]

$$\rho_{it}(A \mid s) = \frac{\rho_0(A)\hat{P}_{Ait}(s)}{\rho_0(A)\hat{P}_{Ait}(s) + \rho_0(B)\hat{P}_{Bit}(s)}, \qquad \rho_{it}(B \mid s) = 1 - \rho_{it}(A \mid s). \tag{B-4}$$

In the experiment, $\rho_0(A) = \rho_0(B) = 1/2$, so the model-prior terms cancel. Bounding the weight between zero and one keeps the weighted belief between the two model-specific posteriors when fit errors are large. Best-fit agents select model A when $\hat{P}_{Ait}(s) > \hat{P}_{Bit}(s)$ and model B otherwise. Let $\hat{m}^*_{it}$ denote this selected model. Non-movers retain the state prior and do not use model-fit calculations.

Second layer: reporting errors. Let ω_1 denote selection of Urn X and ω_2 selection of Urn Y. The internal belief about Urn X for a participant of type k is $\mu_{k,it}(\omega_1 \mid s)$. The three rules imply

$$\mu_{\text{Bayesian},it}(\omega_1 \mid s) = \sum_{m \in \{A,B\}} \rho_{it}(m \mid s)\mu_m(\omega_1 \mid s), \tag{B-5}$$

$$\mu_{\text{Best-fit},it}(\omega_1 \mid s) = \mu_{\hat{m}^*_{it}}(\omega_1 \mid s), \tag{B-6}$$

$$\mu_{\text{Non-mover},it}(\omega_1 \mid s) = \mu_0(\omega_1). \tag{B-7}$$

[44] In structural estimation, subjective fits below 10^{-10} are set to 10^{-10}, and model weights are therefore restricted to $[0, 1]$. The main results are robust when I modify the lower bound.

A second layer of mean-zero normal error is added to the internal belief before censoring at zero and one:

$$\text{Belief}_{it} = \mu_{k,it}(\omega_1 \mid s) + \varepsilon_{it}^2, \qquad \varepsilon_{it}^2 \sim N(0, \sigma_{2,k}^2). \tag{B-8}$$

Reporting errors are independent across participants and rounds and independent of the first-layer errors. The reporting-error standard deviation is common to all three types, $\sigma_{2,k} = \sigma_2$. The Non-mover type has no fit-calculation error.

Likelihood and estimation. Let ϕ denote the standard normal density. Conditional on the first-layer errors, the density of a reported belief $b \in (0,1)$ is

$$\frac{1}{\sigma_{2,k}} \phi\left(\frac{b - \mu_{k,it}(\omega_1 \mid s)}{\sigma_{2,k}}\right). \tag{B-9}$$

Reported beliefs at zero and one contribute the corresponding lower and upper normal tail probabilities to the likelihood. For Bayesian and Best-fit types, I integrate the conditional likelihood contribution over the two first-layer errors. For Non-movers, I evaluate the same likelihood contribution with $\mu_{k,it}(\omega_1 \mid s) = \mu_0(\omega_1)$ and reporting-error standard deviation σ_2. Let $\ell_{it}(k;\boldsymbol{\sigma}_k)$ denote the resulting report likelihood.[45] Conditional on a participant's type, I assume that reported beliefs are independent across rounds. The participant likelihood is therefore the product of the report likelihoods:

$$L_i(k;\boldsymbol{\sigma}_k) = \prod_{t \in \mathscr{T}_i^{SP}} \ell_{it}(k;\boldsymbol{\sigma}_k), \tag{B-10}$$

where $\mathscr{T}_i^{SP}$ contains participant i's 27 rounds in the *SP* block. With population shares $p_k \geq 0$ and $\sum_k p_k = 1$, the sample likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i \in \mathscr{I}} \sum_{k \in \mathscr{T}} p_k L_i(k;\boldsymbol{\sigma}_k). \tag{B-11}$$

The parameter vector contains two independent population shares and two error standard deviations. Standard errors are calculated from 500 bootstrap samples drawn by resampling participants. Table 3 in Section 5.2 reports the parameter estimates and standard errors.

Participant classification. Using estimated population shares as prior type probabilities, participant i's posterior probability of type k is

$$\Pr(k \mid \{\text{Belief}_{it}\}_{t \in \mathscr{T}_i^{SP}}) = \frac{\hat{p}_k L_i(k;\hat{\boldsymbol{\sigma}}_k)}{\sum_{j \in \mathscr{T}} \hat{p}_j L_i(j;\hat{\boldsymbol{\sigma}}_j)}. \tag{B-12}$$

[45] In structural estimation and participant classification, for numerical stability, report likelihoods below 10^{-10} are set to 10^{-10} before taking logarithms. The main results are robust to modifying this lower bound.

I assign each participant to the type with the highest posterior probability. This procedure classifies 132 participants as Bayesian, 33 as Best-fit, and 28 as Non-movers. Table 4 in Section 5.2 reports how many participants remain classified when the highest posterior probability must exceed 50%, 70%, or 90%.

B.3 Effects of Motivated Reasoning: Robustness Checks

This subsection reports additional estimates of the effect of motivated reasoning. The regressions follow Equation (14) unless stated otherwise, and standard errors are clustered by participant. Regressions with participant fixed effects also include decision-problem-by-signal and round fixed effects.

Table B-2 reports the pooled effect of asymmetric payoffs under three specifications. Column (1) repeats the pooled estimate in Figure 6. Column (2) replaces the participant covariates and decision-problem fixed effects with participant and decision-problem-by-signal fixed effects. Column (3) uses only beliefs from each participant's first block, so each participant contributes beliefs under only one payoff condition, and participant fixed effects cannot be included. The estimated effects are 1.7, 1.6, and 1.3 percentage points, respectively. The first-block estimate is less precise and not statistically significant, but it is similar in size.

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.017^{***}	0.016^{***}	0.013
	(0.005)	(0.005)	(0.009)
Observations	10422	10422	5211
Participants	193	193	193
Mean of DV	0.506	0.506	0.503
SP comparison mean	0.498	0.498	0.496
R^2	0.336	0.511	0.424
Participant FE	No	Yes	No
DP $\times$ signal FE	No	Yes	Yes
DP FE	Yes	No	No
Sample	All reports	All reports	First block
Participant covariates	Yes	No	No

Notes: The table reports OLS estimates. The dependent variable is reported belief. Columns (1) and (2) use all beliefs; column (3) uses beliefs from the first block. Unreported controls in column (1) include signal color, gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness. Column (1) also includes unreported round and decision-problem fixed effects. Column (2) includes unreported participant, decision-problem-by-signal, and round fixed effects. Column (3) includes unreported decision-problem-by-signal and round fixed effects. The *SP* comparison mean uses beliefs in the *SP* block in the same estimation sample. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-2: Overall Effects and First-Block Comparison

Table B-3 repeats the estimates by type in Figure 6, replacing the participant covariates with participant fixed effects. As in Figure 6, types are assigned as in Section 5.2. The estimated effect is 7.9 percentage points among Best-fit participants ($p < 0.01$), compared with 0.5 percentage points among Bayesian participants and -0.2 percentage points among Non-movers. These estimates are close to those in Figure 6.

DV: Reported Belief	(1)	(2)	(3)	(4)
$\mathbb{1}_{\text{Asym}}$	0.016***	0.005	0.079***	-0.002
	(0.005)	(0.003)	(0.023)	(0.009)
Observations	10422	7128	1782	1512
Participants	193	132	33	28
Mean of DV	0.506	0.513	0.476	0.511
R^2	0.511	0.627	0.372	0.579
Sample	All participants	Bayesian	Best-fit	Non-mover

Notes: The table reports OLS estimates. The dependent variable is reported belief. Columns (1)–(4) use all participants, Bayesian participants, Best-fit participants, and Non-movers, respectively. The specification includes participant, decision-problem-by-signal, and round fixed effects. The Non-mover regression contains 28 participants. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-3: Motivated Reasoning Effects with Participant Fixed Effects

Table B-4 extends Table 5, which interacts $\mathbb{1}_{\text{Asym}}$ with each participant's posterior type probabilities. Column (2) repeats the specification in Table 5. Column (1) omits the participant covariates, and column (3) adds participant fixed effects, which absorb the posterior type probabilities. The coefficient on the interaction between $\mathbb{1}_{\text{Asym}}$ and the Best-fit probability ranges from 0.077 to 0.082 across the three columns ($p < 0.01$ in each).

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.004	0.004	0.004
	(0.004)	(0.004)	(0.003)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Best-fit Type	0.082***	0.082***	0.077***
	(0.022)	(0.023)	(0.023)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Non-mover Type	-0.007	-0.007	-0.014
	(0.011)	(0.011)	(0.011)
Likelihood of Best-fit Type	-0.076***	-0.069***	
	(0.019)	(0.018)	
Likelihood of Non-mover Type	-0.001	0.004	
	(0.007)	(0.008)	
Observations	10422	10422	10422
Participants	193	193	193
Mean of DV	0.506	0.506	0.506
R^2	0.340	0.343	0.515
Participant FE	No	No	Yes
DP $\times$ signal FE	No	No	Yes
DP FE	Yes	Yes	No
Participant covariates	No	Yes	Absorbed
Narrative extremeness	No	Yes	Absorbed

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with posterior type probabilities. The dependent variable is reported belief. Each column uses 10,422 beliefs from 193 participants. Posterior type probabilities are estimated from the 27 beliefs per participant in the *SP* block. The Bayesian type is the reference category. Each regression includes $\mathbb{1}_{\text{Asym}}$, the other types' posterior probabilities, and their interactions with $\mathbb{1}_{\text{Asym}}$. Columns (1) and (2) include fixed effects for signal color, round, and decision problem. Column (2) also controls for gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. Column (3) includes participant, decision-problem-by-signal, and round fixed effects. Participant fixed effects absorb the posterior type probabilities in column (3). Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-4: Posterior-Probability Interactions: Three Types

The estimates above rely on the classification from the finite mixture model. As a robustness check that does not use this model, I construct a simple measure of how close each participant's reports are to Best-fit predictions. For each report in the *SP* block, I calculate its absolute distance from the Bayesian benchmark minus its absolute distance from the Best-fit benchmark. For each participant and decision problem, the closeness to Best-fit predictions is the average of this difference over the participant's *SP* reports in the other eight decision problems. Excluding the current decision problem ensures that the measure does not use reports from the decision problem in which the effect is estimated. Higher values mean that a participant's reports are closer to Best-fit predictions relative to Bayesian predictions. Most participants have negative values, and the median participant average is -0.11.

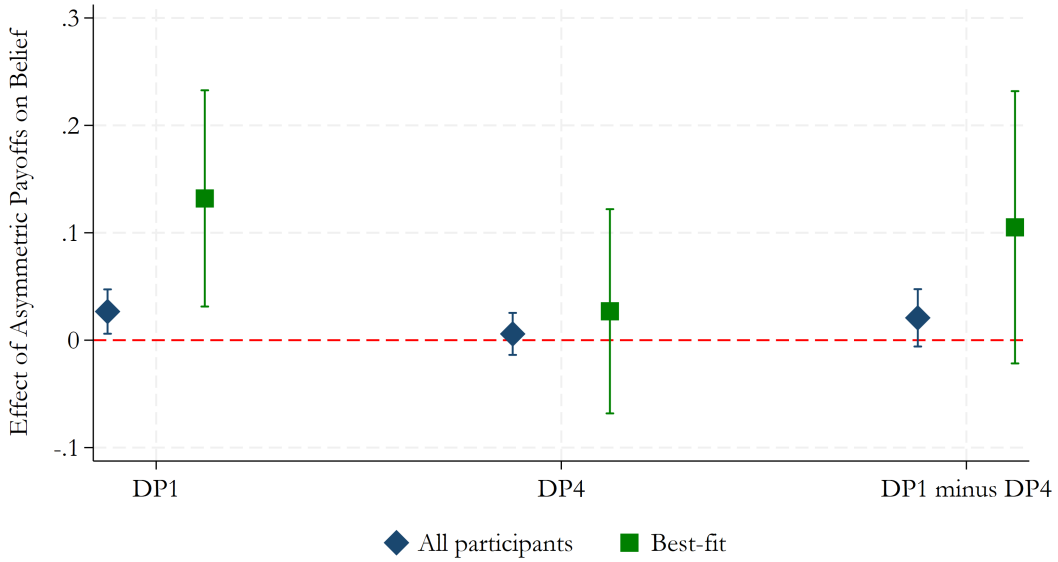
Table B-5 adds this measure and its interaction with $\mathbb{1}_{\text{Asym}}$ to Equation (14). The interaction coefficient is 0.544 ($p < 0.01$). At the lower and upper quartiles of participant averages, -0.15 and -0.06, the implied effect of asymmetric payoffs is -1.3 and 3.5 percentage points, respectively. Participants whose reports are closer to Best-fit predictions therefore respond more strongly to asymmetric payoffs, consistent with the estimates by type.

DV: Reported Belief	(1)
$\mathbb{1}_{\text{Asym}}$	0.068***
	(0.011)
Closeness to Best-fit predictions	-0.406***
	(0.067)
$\mathbb{1}_{\text{Asym}} \times$ Closeness	0.544***
	(0.090)
Observations	10422
Participants	193
Mean of DV	0.506
R^2	0.345

Notes: The table reports OLS estimates. The dependent variable is reported belief. The sample contains 10,422 beliefs from 193 participants. Closeness to Best-fit predictions averages absolute Bayesian error minus absolute Best-fit error across beliefs in the *SP* block outside the current decision problem. The regression includes $\mathbb{1}_{\text{Asym}}$, closeness to Best-fit predictions, and their interaction. Unreported controls include gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. The specification includes fixed effects for signal color, round, and decision problem. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-5: Motivated Reasoning Effects and Closeness to Best-fit Predictions

Decision Problems 1 and 4. Decision Problem 4 reproduces Decision Problem 1 after exchanging the state labels and model labels. The two problems therefore share the state prior, objective fit gap, and separation between model-specific posteriors. Their objectively best-fitting models favor Urn Y and Urn X, respectively. Figure B-1 displays the effects in each problem and their difference. For Best-fit participants, the estimated effect of motivated reasoning is approximately 13.2 percentage points in Decision Problem 1 and 2.7 percentage points in Decision Problem 4. Their difference is 10.5 percentage points, with a two-sided p-value of 0.101. The pooled difference is 2.1 percentage points and has $p = 0.126$.



Notes: The figure reports estimated effects of motivated reasoning using beliefs from Decision Problems 1 and 4. Navy diamonds show pooled estimates and green squares show Best-fit estimates. The samples contain 2,316 beliefs from 193 participants and 396 beliefs from 33 participants, respectively. Regressions include participant, decision-problem-by-signal, and round fixed effects and interact $\mathbb{1}_{\text{Asym}}$ with an indicator for Decision Problem 1. The first two positions show the implied effects of motivated reasoning; the third shows their directly estimated difference. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level.

Figure B-1: Motivated Reasoning Effects in Decision Problems 1 and 4

Table B-6 reports the estimates plotted in Figure B-1. For Best-fit participants, the difference between Decision Problems 1 and 4 is 10.5 percentage points with a standard error of 6.2 percentage points. The estimates have the predicted sign, but the difference is imprecisely estimated.

DV: Reported Belief	(1)	(2)
$\mathbb{1}_{\text{Asym}}$ (DP4)	0.006	0.027
	(0.010)	(0.047)
DP1 minus DP4 effect	0.021	0.105
	(0.014)	(0.062)
Observations	2316	396
Participants	193	33
Mean of DV	0.518	0.525
R^2	0.228	0.378
Sample	All participants	Best-fit

Notes: The table reports OLS estimates. The dependent variable is reported belief. The sample contains only Decision Problems 1 and 4. Columns (1) and (2) use all participants and Best-fit participants, respectively. The first row reports the effect of motivated reasoning in Decision Problem 4. The second reports the difference between Decision Problems 1 and 4. The specification includes participant, decision-problem-by-signal, and round fixed effects. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-6: Motivated Reasoning Effects in Decision Problems 1 and 4: Regression Estimates

B.4 Type-Specific Error Standard Deviations

This subsection reports the three-type classification when the error standard deviations in Equations (B-3) and (B-8) differ across types. The fit-calculation standard deviation $\sigma_{1,k}$ is estimated separately for the Bayesian and Best-fit types, and the reporting standard deviation $\sigma_{2,k}$ is estimated separately for each of the three types. The parameter vector therefore contains two independent population shares and five error standard deviations. The behavioral rules, participant likelihood, and posterior assignment otherwise follow Appendix B.2.[46] Table B-7 reports the estimates.

This specification classifies 123 participants as Bayesian, 63 as Best-fit, and seven as Non-movers, compared with 132, 33, and 28 in the main specification (Table 4). Table B-8 compares these assignments with the main classification. All 33 participants classified as Best-fit in the main specification remain Best-fit. The other 30 Best-fit participants are classified as Bayesian (26) or Non-movers (4) in the main specification. Figure B-2 reports the effects of motivated reasoning by type. With participant covariates, the estimated effect is 4.9 percentage points among Best-fit participants ($p < 0.001$) and 0.1 percentage points among Bayesian participants ($p = 0.693$), compared with 8.4 and 0.5 percentage points in the main specification (Figure 6).

Parameter	**Estimate**	**Bootstrap SE**
Population Shares		
p_1 (Bayesian)	0.641	0.059
p_2 (Best-fit)	0.322	0.059
p_3 (Non-mover)	0.036	0.013
First-layer (fit-calculation) error SD		
$\sigma_{1,\text{Bayes}}$	0.004	0.001
$\sigma_{1,\text{Best-fit}}$	0.278	0.067
$\sigma_{1,\text{Non-mover}}$		—
Second-layer (reporting) error SD		
$\sigma_{2,\text{Bayes}}$	0.103	0.010
$\sigma_{2,\text{Best-fit}}$	0.185	0.008
$\sigma_{2,\text{Non-mover}}$	0.001	0.000

Notes: The table reports maximum likelihood estimates of population shares and type-specific error standard deviations. The sample contains 5,211 beliefs from 193 participants in the *SP* block, with 27 beliefs per participant. Population shares are the estimated fractions of each type in the population from which participants are drawn. Standard errors are calculated from 500 bootstrap samples drawn by resampling participants.

Table B-7: Type-Specific Errors: Three-Type Parameter Estimates

[46] With the reporting error, the belief implied by the rule and the error can fall below zero or above one, but participants can only report beliefs between zero and one. Therefore, the likelihood uses normal densities for reports strictly between zero and one. For reports at zero and one, it uses the normal probabilities below zero and above one, respectively.

	Bayesian	Best-fit	Non-mover	Total
Bayesian	106	26	0	132
Best-fit	0	33	0	33
Non-mover	17	4	7	28
Total	123	63	7	193

Notes: The table compares participant classifications under the main three-type specification and the three-type specification with type-specific error standard deviations. Rows report the main classification, and columns report the alternative classification. Each entry gives the number of participants. Both specifications assign each of the 193 participants to the type with the highest posterior probability, using beliefs from the *SP* block.

Table B-8: Type-Specific Errors and Main Three-Type Assignments

DV: Reported Belief	(1)	(2)	(3)	(4)
$\mathbb{1}_{\text{Asym}}$	0.017^{***}	0.001	0.049^{***}	0.002
	(0.005)	(0.004)	(0.014)	(0.001)
Observations	10422	6642	3402	378
Participants	193	123	63	7
Mean of DV	0.506	0.506	0.507	0.502
R^2	0.336	0.458	0.261	0.978
Sample	All participants	Bayesian	Best-fit	Non-mover

Notes: The table reports OLS estimates. The dependent variable is reported belief. Columns (1)–(4) use the pooled sample, Bayesian participants, Best-fit participants, and Non-movers, respectively. Each participant contributes 54 beliefs. Unreported controls include gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. The specification includes fixed effects for signal color, round, and decision problem. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

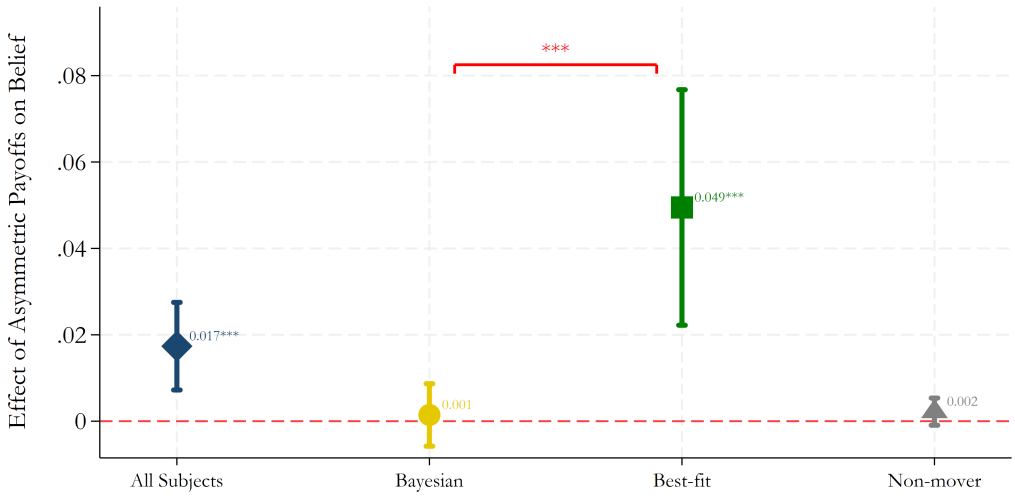
Table B-9: Type-Specific Errors: Motivated Reasoning Regressions, Participant Covariates

With participant covariates, the effect of motivated reasoning is 4.8 percentage points larger among Best-fit participants than among Bayesian participants ($p < 0.001$). A test that the effect of motivated reasoning is equal across all types gives $p = 0.003$. Both tests come from a single regression that interacts each control and fixed effect with type.

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.001	0.001	0.001
	(0.004)	(0.004)	(0.003)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Best-fit Type	0.050***	0.050***	0.045***
	(0.015)	(0.015)	(0.014)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Non-mover Type	0.005	0.005	0.001
	(0.005)	(0.005)	(0.010)
Likelihood of Best-fit Type	-0.026**	-0.024**	
	(0.013)	(0.012)	
Likelihood of Non-mover Type	-0.008**	-0.006	
	(0.004)	(0.008)	
Observations	10422	10422	10422
Participants	193	193	193
Mean of DV	0.506	0.506	0.506
R^2	0.335	0.339	0.513
Participant FE	No	No	Yes
DP $\times$ signal FE	No	No	Yes
DP FE	Yes	Yes	No
Participant covariates	No	Yes	Absorbed
Narrative extremeness	No	Yes	Absorbed

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with posterior type probabilities. The dependent variable is reported belief. Each column uses 10,422 beliefs from 193 participants. Posterior type probabilities are estimated from the 27 beliefs per participant in the *SP* block. The Bayesian type is the reference category. Each regression includes $\mathbb{1}_{\text{Asym}}$, the other types' posterior probabilities, and their interactions with $\mathbb{1}_{\text{Asym}}$. Columns (1) and (2) include fixed effects for signal color, round, and decision problem. Column (2) also controls for gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. Column (3) includes participant, decision-problem-by-signal, and round fixed effects. Participant fixed effects absorb the posterior type probabilities in column (3). Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-10: Type-Specific Errors: Posterior-Probability Interactions



Notes: The figure reports estimated coefficients on $\mathbb{1}_{\text{Asym}}$ in Equation (14) for the pooled sample and separately for each behavioral type, using the classification with type-specific error standard deviations. The pooled, Bayesian, Best-fit, and Non-mover samples contain 193, 123, 63, and 7 participants, respectively. Each participant contributes 54 beliefs. Regressions control for gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness. All specifications include fixed effects for signal color, round, and decision problem. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level. The red bracket reports the significance of the difference between the Best-fit and Bayesian coefficients. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure B-2: Type-Specific Errors: Motivated Reasoning Effects, All Types

B.5 Alternative Mixture Specifications

This subsection examines how the results change under alternative sets of behavioral types. Each specification estimates population shares, the two error standard deviations shared across types, and the parameters of any Grether components using the 27 beliefs per participant in the *SP* block. The Gaussian error structure, participant likelihood, and posterior assignment follow Appendix B.2. Each additional Grether component has two parameters, $\beta_{1,k}$ and $\beta_{2,k}$, common to participants in that component, and shares the two error standard deviations with the other types.

Grether components. Following Grether (1980), I allow prior odds and likelihood ratios to receive different weights when updating beliefs within each model. The following equations define the Grether components. For each model, the transformation applies to its Bayesian posterior $\mu_m(\omega_1 \mid s)$:

$$\mu^G_{m,k}(\omega_1 \mid s) = \frac{\mu_0(\omega_1)^{\beta_{1,k}} \mu_m(\omega_1 \mid s)^{\beta_{2,k}}}{\mu_0(\omega_1)^{\beta_{1,k}} \mu_m(\omega_1 \mid s)^{\beta_{2,k}} + \mu_0(\omega_2)^{\beta_{1,k}} \mu_m(\omega_2 \mid s)^{\beta_{2,k}}}. \tag{B-13}$$

Here, $\beta_{1,k}$ weights the state prior, and $\beta_{2,k}$ weights the model-specific posterior, which already incorporates the signal and prior. In log odds, this transformation is equivalent to

$$\log \frac{\mu^G_{m,k}(\omega_1 \mid s)}{\mu^G_{m,k}(\omega_2 \mid s)} = (\beta_{1,k} + \beta_{2,k}) \log \frac{\mu_0(\omega_1)}{\mu_0(\omega_2)} + \beta_{2,k} \log \frac{\pi_m(s \mid \omega_1)}{\pi_m(s \mid \omega_2)}. \tag{B-14}$$

Thus, the effective coefficient on prior log odds is $\beta_{1,k} + \beta_{2,k}$.

The Grether-Bayesian component applies the same transformation to posterior model probabilities. Let $\hat{P}^+_{mit}(s)$ denote the positive perceived fit used to evaluate this transformation.[47] Posterior model probabilities are proportional to $\rho_0(m)\hat{P}^+_{mit}(s)$. The transformation multiplies their $\beta_{2,k}$ powers by the model prior raised to $\beta_{1,k}$. Its weight on model A is therefore

$$\rho^G_{kit}(A \mid s) = \frac{\rho_0(A)^{\beta_{1,k}+\beta_{2,k}} [\hat{P}^+_{Ait}(s)]^{\beta_{2,k}}}{\rho_0(A)^{\beta_{1,k}+\beta_{2,k}} [\hat{P}^+_{Ait}(s)]^{\beta_{2,k}} + \rho_0(B)^{\beta_{1,k}+\beta_{2,k}} [\hat{P}^+_{Bit}(s)]^{\beta_{2,k}}}. \tag{B-15}$$

The weight on model B is $1 - \rho^G_{kit}(A \mid s)$. Here, $\rho_0(m)$ is the prior probability of model m.

The Grether-Bayesian internal belief is the weighted average of the transformed model-specific posteriors:

$$\mu_{k,it}(\omega_1 \mid s) = \rho^G_{kit}(A \mid s)\mu^G_{A,k}(\omega_1 \mid s) + [1 - \rho^G_{kit}(A \mid s)]\mu^G_{B,k}(\omega_1 \mid s). \tag{B-16}$$

[47] For numerical stability, $\hat{P}^+_{mit}(s) = \max\{\hat{P}_{mit}(s), 10^{-10}\}$.

The Grether-Best-fit component selects the model with the larger transformed model probability. It selects model A when

$$\rho_0(A)^{\beta_{1,k}+\beta_{2,k}}[\hat{P}^+_{Ait}(s)]^{\beta_{2,k}} > \rho_0(B)^{\beta_{1,k}+\beta_{2,k}}[\hat{P}^+_{Bit}(s)]^{\beta_{2,k}}, \tag{B-17}$$

and model B otherwise.

In the experiment, $\rho_0(A) = \rho_0(B) = 1/2$. The model-prior terms therefore cancel from both the model weights and the selection comparison. Because $\beta_{2,k} > 0$, the selection comparison reduces to $\hat{P}^+_{Ait}(s) > \hat{P}^+_{Bit}(s)$. Thus, the Grether-Best-fit component selects the model with the larger positive perceived fit.

Its internal belief is the selected model's transformed posterior from Equation (B-13). Both components add the shared normal reporting error and integrate over the first-layer errors when evaluating report likelihoods. Under this parameterization, $\beta_{1,k} = 0$ and $\beta_{2,k} = 1$ recover the Bayesian rule in the absence of errors.

With equal model priors, these parameter values also recover the Best-fit rule in the absence of errors. Setting both parameters to one gives an effective prior coefficient of two and a signal-likelihood coefficient of one.

Alternative sets of types. The two-type specification contains Bayesian and Best-fit components, allowing participants classified as Non-movers in the main specification to be reassigned. The four-type specification adds one Grether-Bayesian component to the main three-type model. Five-type specification I adds both Grether-Bayesian and Grether-Best-fit components to the main three-type model. Five-type specification II adds two Grether-Bayesian components, A and B, with separately estimated parameters. Component A imposes $\beta_{1,A} > \beta_{2,A}$, while component B imposes $\beta_{2,B} > \beta_{1,B}$. These restrictions distinguish the two components within the implemented transformation.

The additional components change which participants receive the Bayesian and Best-fit classifications. The comparisons below therefore assess sensitivity to the set of behavioral types and the resulting participant assignments. Five-type specification II assigns 11 participants to Grether-Bayesian A. The four-type specification and five-type specifications I and II classify 29, 28, and 24 participants as Non-movers, respectively, compared with 28 in the main specification.

Each specification reports an assignment table, a figure of motivated reasoning effects, and two regression tables. The regression tables use participant covariates and interactions with posterior type probabilities, respectively. Participant covariates include gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness. These regressions also include fixed effects for signal color, round, and decision problem. Standard errors are clustered at the

participant level.

B.5.1 Two-Type Specification

	Bayesian	Best-fit	Total
Bayesian	132	0	132
Best-fit	1	32	33
Non-mover	28	0	28
Total	161	32	193

Notes: The table compares participant classifications under the main three-type specification and the alternative specification. Rows report the main classification, and columns report the alternative classification. Each entry gives the number of participants. Both specifications assign each of the 193 participants to the type with the highest posterior probability, using beliefs from the *SP* block.

Table B-11: Two-Type and Main Three-Type Assignments

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.017^{***}	0.004	0.089^{***}
	(0.005)	(0.004)	(0.024)
Observations	10422	8694	1728
Participants	193	161	32
Mean of DV	0.506	0.512	0.476
R^2	0.336	0.414	0.230
Sample	All participants	Bayesian	Best-fit

Notes: The table reports OLS estimates. The dependent variable is reported belief. Columns (1)–(3) use the pooled sample, Bayesian participants, and Best-fit participants, respectively. Each participant contributes 54 beliefs. Unreported controls include gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. The specification includes fixed effects for signal color, round, and decision problem. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

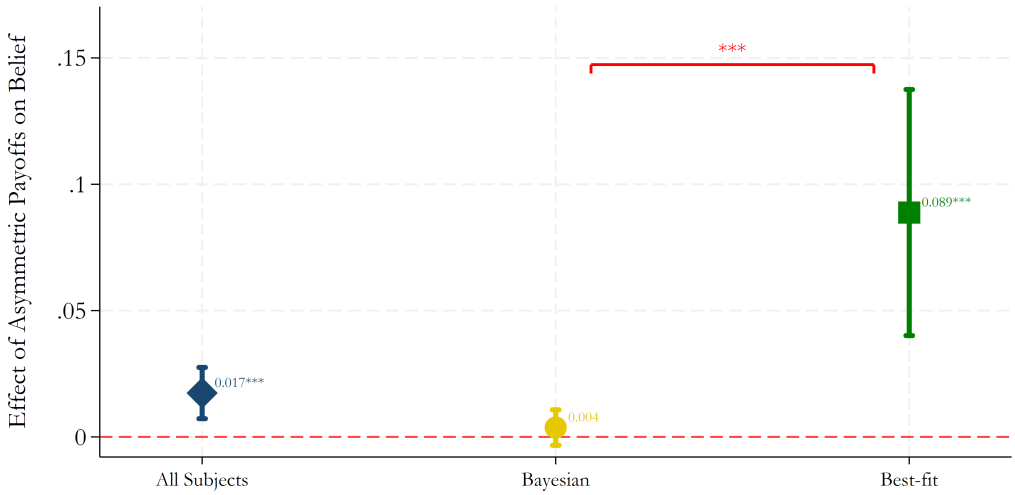
Table B-12: Two-Type Motivated Reasoning Regressions: Participant Covariates

With participant covariates, the effect of motivated reasoning is 8.5 percentage points larger among Best-fit participants than among Bayesian participants ($p < 0.001$). A test that the effect of motivated reasoning is equal across all types gives $p < 0.001$. Both tests come from a single regression that interacts each control and fixed effect with type.

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.003	0.003	0.002
	(0.004)	(0.004)	(0.003)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Best-fit Type	0.086***	0.086***	0.082***
	(0.023)	(0.023)	(0.023)
Likelihood of Best-fit Type	-0.078***	-0.072***	
	(0.019)	(0.019)	
Observations	10422	10422	10422
Participants	193	193	193
Mean of DV	0.506	0.506	0.506
R^2	0.341	0.343	0.515
Participant FE	No	No	Yes
DP $\times$ signal FE	No	No	Yes
DP FE	Yes	Yes	No
Participant covariates	No	Yes	Absorbed
Narrative extremeness	No	Yes	Absorbed

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with posterior type probabilities. The dependent variable is reported belief. Each column uses 10,422 beliefs from 193 participants. Posterior type probabilities are estimated from the 27 beliefs per participant in the *SP* block. The Bayesian type is the reference category. Each regression includes $\mathbb{1}_{\text{Asym}}$, the other types' posterior probabilities, and their interactions with $\mathbb{1}_{\text{Asym}}$. Columns (1) and (2) include fixed effects for signal color, round, and decision problem. Column (2) also controls for gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. Column (3) includes participant, decision-problem-by-signal, and round fixed effects. Participant fixed effects absorb the posterior type probabilities in column (3). Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-13: Two-Type Posterior-Probability Interactions



Notes: The figure reports estimated coefficients on $\mathbb{1}_{\text{Asym}}$ in Equation (14) for the pooled sample and separately for each behavioral type. The pooled, Bayesian, and Best-fit samples contain 193, 161, and 32 participants, respectively. Each participant contributes 54 beliefs. Regressions control for gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness. All specifications include fixed effects for signal color, round, and decision problem. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level. The red bracket reports the significance of the difference between the Best-fit and Bayesian coefficients. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure B-3: Two-Type Motivated Reasoning Effects: All Types

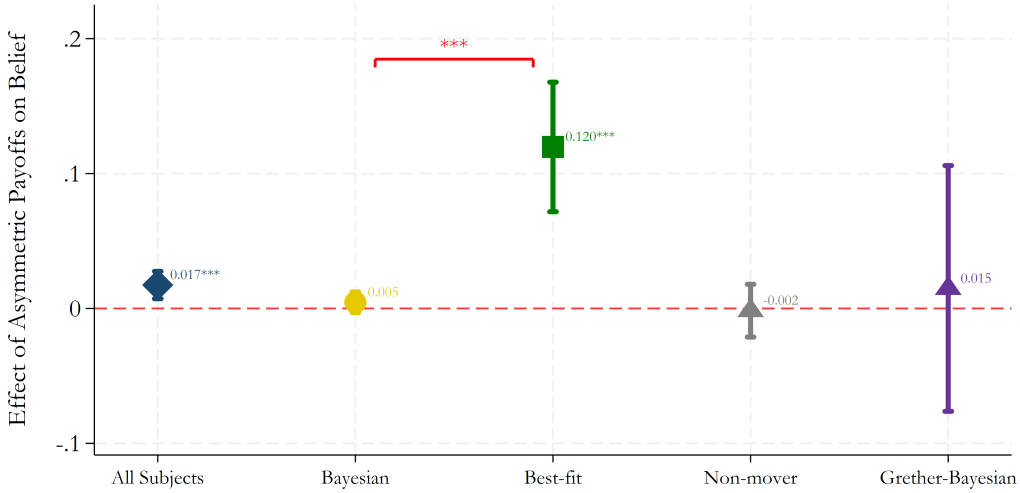
B.5.2 Four-Type Specification

	Bayesian	Best-fit	Non-mover	Grether-Bayesian	Total
Bayesian	128	3	1	0	132
Best-fit	1	19	0	13	33
Non-mover	0	0	28	0	28
Total	129	22	29	13	193

Notes: The table compares participant classifications under the main three-type specification and the alternative specification. Rows report the main classification, and columns report the alternative classification. Each entry gives the number of participants. Both specifications assign each of the 193 participants to the type with the highest posterior probability, using beliefs from the *SP* block.

Table B-14: Four-Type and Main Three-Type Assignments

With participant covariates, the effect of motivated reasoning is 11.5 percentage points larger among Best-fit participants than among Bayesian participants ($p < 0.001$). A test that the effect of motivated reasoning is equal across all types gives $p < 0.001$. Both tests come from a single regression that interacts each control and fixed effect with type.



Notes: The figure reports estimated coefficients on $\mathbb{1}_{\text{Asym}}$ in Equation (14) for the pooled sample and separately for each behavioral type. The pooled sample contains 193 participants. The Bayesian, Best-fit, Non-mover, and Grether-Bayesian samples contain 129, 22, 29, and 13 participants, respectively. Each participant contributes 54 beliefs. Regressions control for gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness. All specifications include fixed effects for signal color, round, and decision problem. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level. The red bracket reports the significance of the difference between the Best-fit and Bayesian coefficients. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure B-4: Four-Type Motivated Reasoning Effects: All Types

DV: Reported Belief	(1)	(2)	(3)	(4)	(5)
$\mathbb{1}_{\text{Asym}}$	0.017***	0.005	0.120***	-0.002	0.015
	(0.005)	(0.004)	(0.023)	(0.010)	(0.042)
Observations	10422	6966	1188	1566	702
Participants	193	129	22	29	13
Mean of DV	0.506	0.511	0.472	0.510	0.511
R^2	0.336	0.417	0.262	0.519	0.403
Sample	All participants	Bayesian	Best-fit	Non-mover	Grether-Bayesian

Notes: The table reports OLS estimates. The dependent variable is reported belief. Columns (1)–(4) use the pooled sample, Bayesian participants, Best-fit participants, and Non-movers, respectively. Column (5) uses Grether-Bayesian participants. Each participant contributes 54 beliefs. Unreported controls include gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. The specification includes fixed effects for signal color, round, and decision problem. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-15: Four-Type Motivated Reasoning Regressions: Participant Covariates

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.004	0.004	0.004
	(0.004)	(0.004)	(0.003)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Best-fit Type	0.117***	0.117***	0.114***
	(0.023)	(0.023)	(0.022)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Non-mover Type	-0.007	-0.007	-0.013
	(0.011)	(0.011)	(0.010)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Grether-Bayesian Type	0.009	0.009	0.003
	(0.036)	(0.036)	(0.036)
Likelihood of Best-fit Type	-0.102***	-0.097***	
	(0.022)	(0.021)	
Likelihood of Non-mover Type	0.001	0.006	
	(0.007)	(0.007)	
Likelihood of Grether-Bayesian Type	-0.003	0.002	
	(0.024)	(0.024)	
Observations	10422	10422	10422
Participants	193	193	193
Mean of DV	0.506	0.506	0.506
R^2	0.343	0.346	0.518
Participant FE	No	No	Yes
DP $\times$ signal FE	No	No	Yes
DP FE	Yes	Yes	No
Participant covariates	No	Yes	Absorbed
Narrative extremeness	No	Yes	Absorbed

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with posterior type probabilities. The dependent variable is reported belief. Each column uses 10,422 beliefs from 193 participants. Posterior type probabilities are estimated from the 27 beliefs per participant in the *SP* block. The Bayesian type is the reference category. Each regression includes $\mathbb{1}_{\text{Asym}}$, the other types' posterior probabilities, and their interactions with $\mathbb{1}_{\text{Asym}}$. Columns (1) and (2) include fixed effects for signal color, round, and decision problem. Column (2) also controls for gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. Column (3) includes participant, decision-problem-by-signal, and round fixed effects. Participant fixed effects absorb the posterior type probabilities in column (3). Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

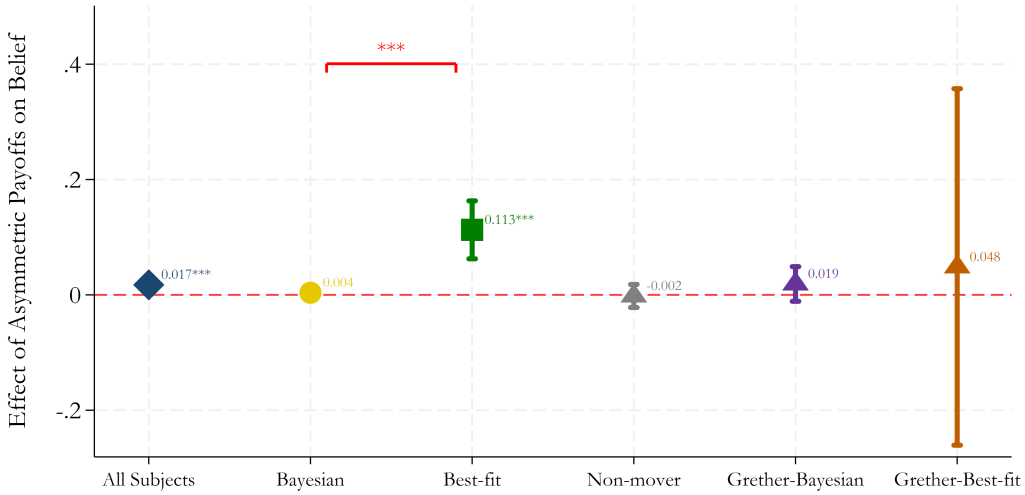
Table B-16: Four-Type Posterior-Probability Interactions

B.5.3 Five-Type I Specification

	Bayesian	Best-fit	Non-mover	Grether-Bayesian	Grether-Best-fit	Total
Bayesian	117	2	1	12	0	132
Best-fit	1	21	0	6	5	33
Non-mover	0	0	27	1	0	28
Total	118	23	28	19	5	193

Notes: The table compares participant classifications under the main three-type specification and the alternative specification. Rows report the main classification, and columns report the alternative classification. Each entry gives the number of participants. Both specifications assign each of the 193 participants to the type with the highest posterior probability, using beliefs from the *SP* block.

Table B-17: Five-Type I and Main Three-Type Assignments



Notes: The figure reports estimated coefficients on $\mathbb{1}_{\text{Asym}}$ in Equation (14) for the pooled sample and separately for each behavioral type. The pooled sample contains 193 participants. The Bayesian, Best-fit, and Non-mover samples contain 118, 23, and 28 participants, respectively. The Grether-Bayesian and Grether-Best-fit samples contain 19 and 5 participants, respectively. Each participant contributes 54 beliefs. Regressions control for gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness. All specifications include fixed effects for signal color, round, and decision problem. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level. The red bracket reports the significance of the difference between the Best-fit and Bayesian coefficients. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure B-5: Five-Type I Motivated Reasoning Effects: All Types

With participant covariates, the effect of motivated reasoning is 10.9 percentage points larger among Best-fit participants than among Bayesian participants ($p < 0.001$). A test that the effect of motivated reasoning is equal across all types gives $p < 0.001$. Both tests come from a single regression that interacts each control and fixed effect with type.

DV: Reported Belief	(1)	(2)	(3)	(4)	(5)	(6)
$\mathbb{1}_{\text{Asym}}$	0.017***	0.004	0.113***	-0.002	0.019	0.048
	(0.005)	(0.004)	(0.024)	(0.010)	(0.014)	(0.111)
Observations	10422	6372	1242	1512	1026	270
Participants	193	118	23	28	19	5
Mean of DV	0.506	0.508	0.477	0.509	0.533	0.493
R^2	0.336	0.422	0.243	0.530	0.498	0.560
Sample	All participants	Bayesian	Best-fit	Non-mover	Grether-Bayesian	Grether-Best-fit

Notes: The table reports OLS estimates. The dependent variable is reported belief. Columns (1)–(4) use the pooled sample, Bayesian participants, Best-fit participants, and Non-movers, respectively. Columns (5) and (6) use Grether-Bayesian and Grether-Best-fit participants, respectively. Each participant contributes 54 beliefs. Unreported controls include gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. The specification includes fixed effects for signal color, round, and decision problem. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-18: Five-Type I Motivated Reasoning Regressions: Participant Covariates

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.003	0.003	0.004
	(0.004)	(0.004)	(0.003)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Best-fit Type	0.108***	0.108***	0.105***
	(0.024)	(0.024)	(0.024)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Non-mover Type	-0.006	-0.006	-0.012
	(0.011)	(0.011)	(0.010)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Grether-Bayesian Type	0.007	0.007	0.004
	(0.015)	(0.015)	(0.014)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Grether-Best-fit Type	0.034	0.034	0.016
	(0.074)	(0.074)	(0.077)
Likelihood of Best-fit Type	-0.089***	-0.083***	
	(0.024)	(0.023)	
Likelihood of Non-mover Type	0.002	0.008	
	(0.007)	(0.007)	
Likelihood of Grether-Bayesian Type	0.019*	0.019*	
	(0.011)	(0.011)	
Likelihood of Grether-Best-fit Type	-0.031	-0.027	
	(0.042)	(0.044)	
Observations	10422	10422	10422
Participants	193	193	193
Mean of DV	0.506	0.506	0.506
R^2	0.342	0.345	0.517
Participant FE	No	No	Yes
DP $\times$ signal FE	No	No	Yes
DP FE	Yes	Yes	No
Participant covariates	No	Yes	Absorbed
Narrative extremeness	No	Yes	Absorbed

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with posterior type probabilities. The dependent variable is reported belief. Each column uses 10,422 beliefs from 193 participants. Posterior type probabilities are estimated from the 27 beliefs per participant in the *SP* block. The Bayesian type is the reference category. Each regression includes $\mathbb{1}_{\text{Asym}}$, the other types' posterior probabilities, and their interactions with $\mathbb{1}_{\text{Asym}}$. Columns (1) and (2) include fixed effects for signal color, round, and decision problem. Column (2) also controls for gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. Column (3) includes participant, decision-problem-by-signal, and round fixed effects. Participant fixed effects absorb the posterior type probabilities in column (3). Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

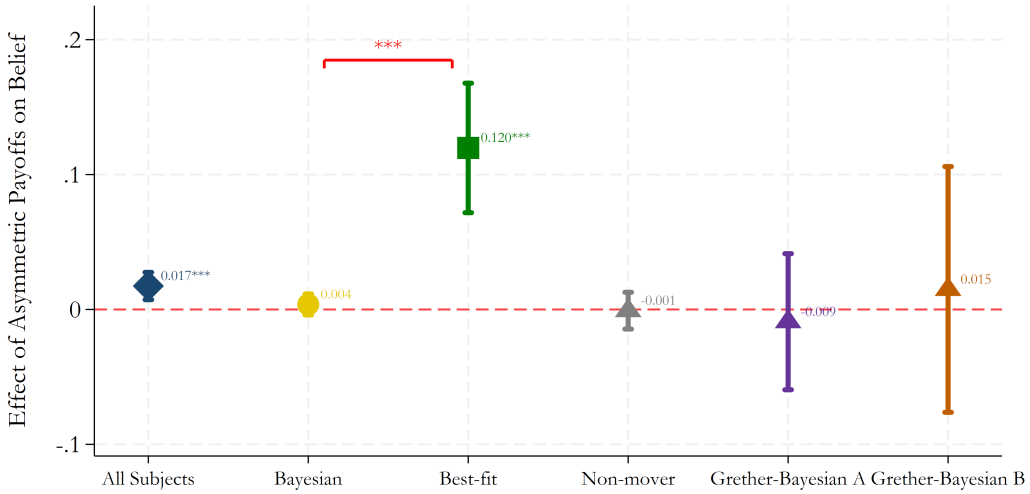
Table B-19: Five-Type I Posterior-Probability Interactions

B.5.4 Five-Type II Specification

	Bayesian	Best-fit	Non-mover	Grether-Bayesian A	Grether-Bayesian B	Total
Bayesian	121	3	1	7	0	132
Best-fit	1	19	0	0	13	33
Non-mover	1	0	23	4	0	28
Total	123	22	24	11	13	193

Notes: The table compares participant classifications under the main three-type specification and the alternative specification. Rows report the main classification, and columns report the alternative classification. Each entry gives the number of participants. Both specifications assign each of the 193 participants to the type with the highest posterior probability, using beliefs from the *SP* block.

Table B-20: Five-Type II and Main Three-Type Assignments



Notes: The figure reports estimated coefficients on $\mathbb{1}_{\text{Asym}}$ in Equation (14) for the pooled sample and separately for each behavioral type. The pooled sample contains 193 participants. The Bayesian, Best-fit, and Non-mover samples contain 123, 22, and 24 participants, respectively. The Grether-Bayesian A and B samples contain 11 and 13 participants, respectively. Each participant contributes 54 beliefs. Regressions control for gender, race, cognitive score, risk attitude, overconfidence, and narrative extremeness. All specifications include fixed effects for signal color, round, and decision problem. Bars indicate 95% confidence intervals based on standard errors clustered at the participant level. The red bracket reports the significance of the difference between the Best-fit and Bayesian coefficients. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Figure B-6: Five-Type II Motivated Reasoning Effects: All Types

With participant covariates, the effect of motivated reasoning is 11.6 percentage points larger among Best-fit participants than among Bayesian participants ($p < 0.001$). A test that the effect of motivated reasoning is equal across all types gives $p < 0.001$. Both tests come from a single regression that interacts each control and fixed effect with type.

DV: Reported Belief	(1)	(2)	(3)	(4)	(5)	(6)
$\mathbb{1}_{\text{Asym}}$	0.017***	0.004	0.120***	-0.001	-0.009	0.015
	(0.005)	(0.004)	(0.023)	(0.007)	(0.023)	(0.042)
Observations	10422	6642	1188	1296	594	702
Participants	193	123	22	24	11	13
Mean of DV	0.506	0.512	0.472	0.509	0.505	0.511
R^2	0.336	0.441	0.262	0.680	0.287	0.403
Sample	All participants	Bayesian	Best-fit	Non-mover	Grether-Bayesian A	Grether-Bayesian B

Notes: The table reports OLS estimates. The dependent variable is reported belief. Columns (1)–(4) use the pooled sample, Bayesian participants, Best-fit participants, and Non-movers, respectively. Columns (5) and (6) use Grether-Bayesian A and Grether-Bayesian B participants, respectively. Each participant contributes 54 beliefs. Unreported controls include gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. The specification includes fixed effects for signal color, round, and decision problem. Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-21: Five-Type II Motivated Reasoning Regressions: Participant Covariates

DV: Reported Belief	(1)	(2)	(3)
$\mathbb{1}_{\text{Asym}}$	0.003	0.003	0.003
	(0.004)	(0.004)	(0.003)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Best-fit Type	0.117***	0.117***	0.114***
	(0.023)	(0.023)	(0.022)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Non-mover Type	-0.001	-0.001	-0.008
	(0.008)	(0.008)	(0.008)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Grether-Bayesian A Type	-0.001	-0.001	-0.004
	(0.022)	(0.022)	(0.021)
$\mathbb{1}_{\text{Asym}} \times$ Likelihood of Grether-Bayesian B Type	0.010	0.010	0.004
	(0.035)	(0.035)	(0.036)
Likelihood of Best-fit Type	-0.101***	-0.096***	
	(0.022)	(0.021)	
Likelihood of Non-mover Type	-0.004	0.002	
	(0.006)	(0.007)	
Likelihood of Grether-Bayesian A Type	-0.006	-0.004	
	(0.013)	(0.012)	
Likelihood of Grether-Bayesian B Type	-0.004	0.001	
	(0.024)	(0.024)	
Observations	10422	10422	10422
Participants	193	193	193
Mean of DV	0.506	0.506	0.506
R^2	0.343	0.345	0.518
Participant FE	No	No	Yes
DP $\times$ signal FE	No	No	Yes
DP FE	Yes	Yes	No
Participant covariates	No	Yes	Absorbed
Narrative extremeness	No	Yes	Absorbed

Notes: The table reports OLS estimates of how the effect of motivated reasoning varies with posterior type probabilities. The dependent variable is reported belief. Each column uses 10,422 beliefs from 193 participants. Posterior type probabilities are estimated from the 27 beliefs per participant in the *SP* block. The Bayesian type is the reference category. Each regression includes $\mathbb{1}_{\text{Asym}}$, the other types' posterior probabilities, and their interactions with $\mathbb{1}_{\text{Asym}}$. Columns (1) and (2) include fixed effects for signal color, round, and decision problem. Column (2) also controls for gender, race indicators, cognitive score, risk attitude, overconfidence, and narrative extremeness. Column (3) includes participant, decision-problem-by-signal, and round fixed effects. Participant fixed effects absorb the posterior type probabilities in column (3). Standard errors clustered at the participant level appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-22: Five-Type II Posterior-Probability Interactions

B.6 Individual Characteristics, Narrative Opinions, and Other Behavioral Measures

B.6.1 Type Associations and Narrative Opinions

The type regressions describe associations between participant characteristics and modal type assignment. Narrative opinions are collected after the laboratory tasks and provide an observational comparison across domains. Their extremeness is the mean absolute distance from five across six responses on the eleven-point scale.

DV: Type Classification	(1)	(2)	(3)
Female	0.032	-0.004	-0.033
	(0.066)	(0.054)	(0.051)
Asian	0.248**	-0.165**	-0.026
	(0.101)	(0.069)	(0.083)
White	0.238**	-0.234***	0.058
	(0.101)	(0.070)	(0.082)
Quiz Score	0.056	0.025	-0.074***
	(0.040)	(0.037)	(0.028)
Cognitive Score	0.010	-0.008	-0.003
	(0.028)	(0.023)	(0.022)
Risk Attitude	0.005***	-0.001	-0.004***
	(0.002)	(0.001)	(0.001)
Overconfidence	-0.075	-0.034	0.112
	(0.080)	(0.064)	(0.069)
Narrative Extremeness	-0.094**	0.067**	0.024
	(0.037)	(0.029)	(0.029)
Prior Experiments	0.014	-0.016	0.002
	(0.011)	(0.010)	(0.009)
Observations	193	193	193
Pseudo R^2	0.109	0.121	0.099
Mean of DV	0.684	0.171	0.145
Type	Bayesian	Best-fit	Non-mover

Notes: The table reports average marginal effects from binary logit regressions. The dependent variables indicate Bayesian, Best-fit, and Non-mover classification in columns (1)–(3), respectively. Each regression uses all 193 participants, with types assigned from their 27 beliefs in the *SP* block. All displayed characteristics enter each regression. Narrative extremeness is the mean absolute distance from five across six zero-to-ten survey responses. Standard errors calculated using the delta method appear in parentheses. The dependent-variable mean gives the corresponding type's participant share. Pseudo R^2 refers to the underlying logit. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-23: Participant Characteristics and Type Assignment

	Spearman rho	p-value	Observations
Laboratory and Narrative Extremeness	0.136	0.060	193

Notes: The table reports the Spearman rank correlation between laboratory belief extremeness and narrative extremeness. Belief extremeness averages absolute distance from one half across all 27 beliefs in the *SP* block. Narrative extremeness averages absolute distance from five across all six zero-to-ten survey responses. The calculation uses 193 participants and includes no controls.

Table B-24: Rank Association between Laboratory and Narrative Extremeness

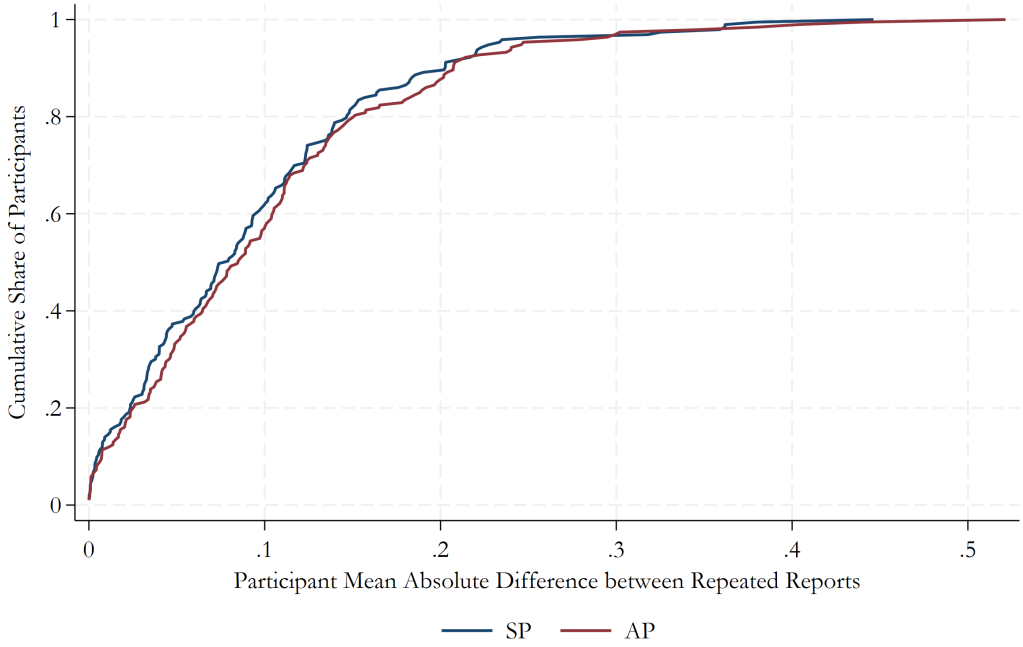
DV: Narrative Extremeness	(1)	(2)	(3)	(4)
Best-fit	0.409**			
	(0.184)			
Non-mover	0.170			
	(0.163)			
SP Relative Benchmark Distance		2.775**	2.780**	
		(1.074)	(1.084)	
Quiz Score			0.073	0.082
			(0.071)	(0.073)
Other Seven Cognitive Tasks			0.013	0.006
			(0.051)	(0.052)
Female			0.056	0.037
			(0.127)	(0.130)
SP Absolute Bayesian Error				1.399*
				(0.721)
Constant	2.217***	2.573***	1.813**	1.348*
	(0.069)	(0.126)	(0.704)	(0.740)
Observations	193	193	193	193
Mean of DV	2.312	2.312	2.312	2.312
R^2	0.034	0.058	0.063	0.025
Participant covariates	No	No	Yes	Yes

Notes: The table reports OLS estimates. The dependent variable is narrative extremeness, the mean absolute distance from five across six zero-to-ten responses. Each participant contributes one observation. Bayesian participants are the reference group in column (1). Relative benchmark distance equals mean absolute Bayesian error minus mean absolute Best-fit error over the 27 beliefs in the *SP* block. Columns (2) and (3) use relative benchmark distance; column (4) uses mean absolute Bayesian error. Columns (3) and (4) add quiz score, performance on the other seven cognitive tasks, and gender. All covariates and constants are displayed. Robust standard errors appear in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table B-25: Narrative Opinions, Type Assignment, and Updating Performance

B.6.2 Repeated Reports

Repeated-report differences measure consistency within identical observed information and preference conditions.

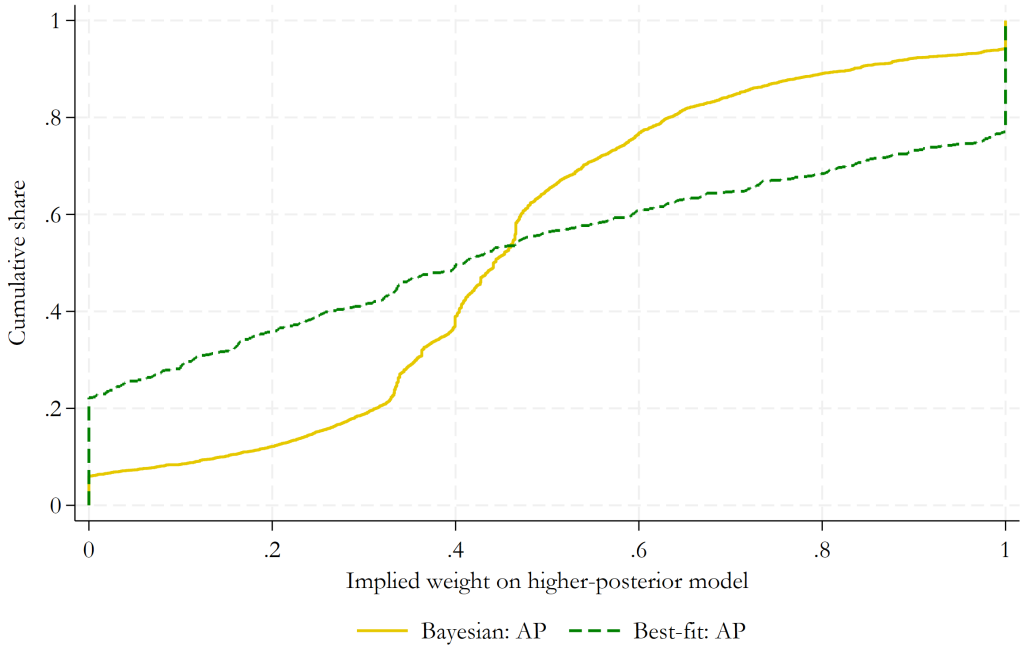


Notes: The curves show cumulative distributions of participant-average absolute differences between beliefs reported under identical information in the same block. Repeated information conditions receive equal weight within each participant. The *SP* and *AP* curves each contain 193 participants, with equal weight per participant.

Figure B-7: Distribution of Repeated-Report Differences

B.6.3 Implied Model Weights

An implied weight expresses a report as a linear combination of the lower and higher model-specific posterior probabilities. The weight on the higher posterior equals the report's distance above the lower posterior divided by the gap between the two posteriors. The single-model problem is excluded because its two posterior predictions coincide. The report-level distributions retain every eligible report and cap weights below zero at zero and above one at one.



Notes: The curves show cumulative distributions of weights on the higher model-specific posterior using beliefs from Decision Problems 1–4 and 6–9 in the *AP* block. The Bayesian and Best-fit samples contain 3,168 beliefs from 132 participants and 792 beliefs from 33 participants, respectively. Weights below zero are set to zero, and weights above one are set to one. Each belief receives equal weight. The masses at zero and one include beliefs outside the interval between the two model-specific posteriors.

Figure B-8: Report-level Implied Weights in the Asymmetric Payoff Block

C Follow-up Study

The follow-up study was conducted at Purdue University in October 2025 with 85 participants. It is separately registered in the AEA RCT Registry (RCT 16882). Participants completed the symmetric payoff block and supplementary tasks in a session lasting approximately 35 minutes. The belief-updating interface additionally displayed the correct posterior implied by each individual model in every scenario (Aina and Schneider 2025). This change reduces the computational demands of within-model updating while retaining the comparison between Bayesian averaging and Best-fit selection.

Table C-1 compares the shares of reports near each benchmark in the main and follow-up studies. Beliefs continue to cluster around both benchmarks when individual-model posteriors are supplied. The follow-up share near the Bayesian benchmark is lower, while the share near the Best-fit benchmark is higher. More reports fall within each window around the Bayesian benchmark than around the Best-fit benchmark. Because the studies involve different cohorts, these comparisons describe the persistence of the behavioral patterns across studies.

Study	Type of Guess	Within 2 p.p.		Within 5 p.p.	
		%	95%-CI	%	95%-CI
Main Study	Bayesian	28.98	[25.29, 32.67]	47.75	[43.21, 52.28]
	Best-fit	5.51	[4.47, 6.54]	12.15	[10.35, 13.94]
Follow-up Study	Bayesian	20.65	[16.75, 24.56]	36.17	[30.88, 41.45]
	Best-fit	10.54	[7.29, 13.80]	17.25	[13.00, 21.51]

Notes: The table reports the percentage of beliefs within the indicated distance of each benchmark in the main and follow-up studies. The samples contain 5,211 beliefs from 193 main-study participants and 2,295 beliefs from 85 follow-up participants, all from the *SP* block. The 95% confidence intervals are based on standard errors calculated from participant-level shares. The follow-up interface displays the posterior under each individual model.

Table C-1: Classification of Round-level Beliefs Across Studies

D Experimental Instructions and Interface

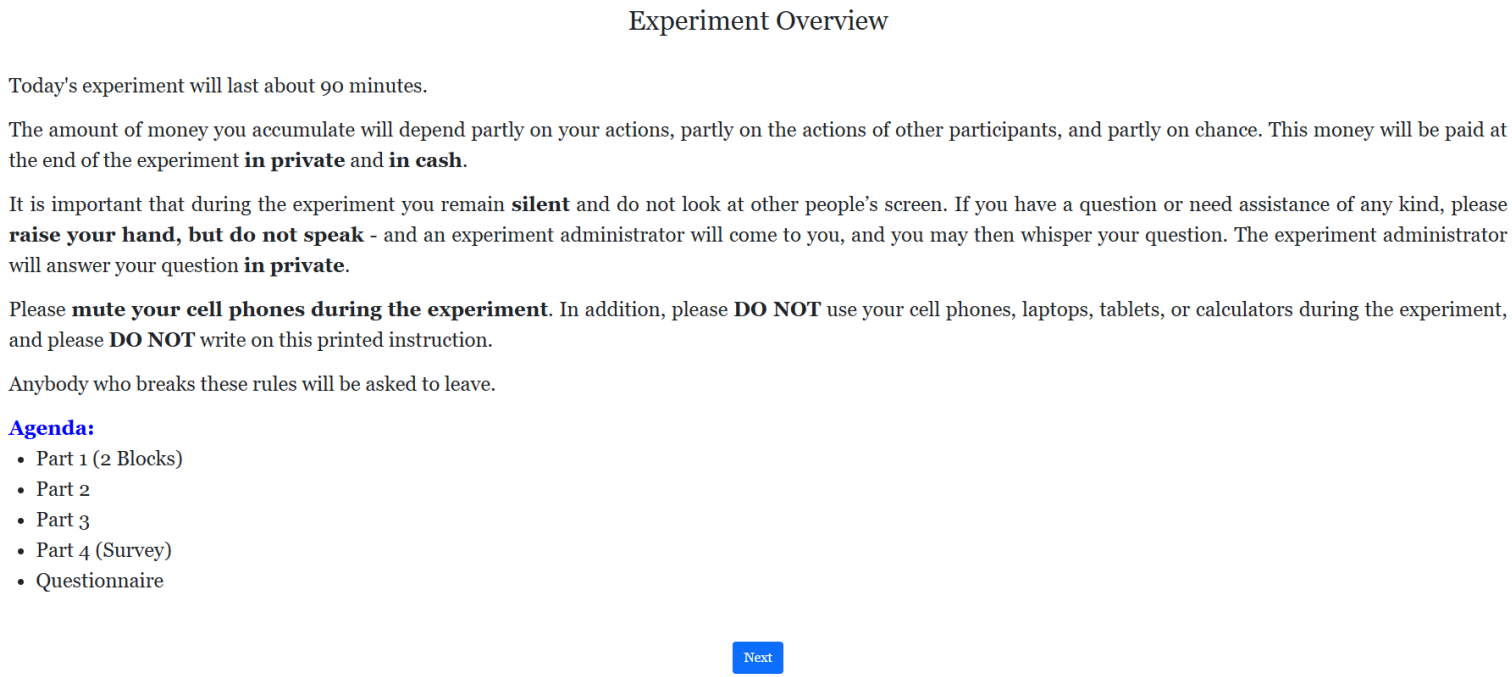


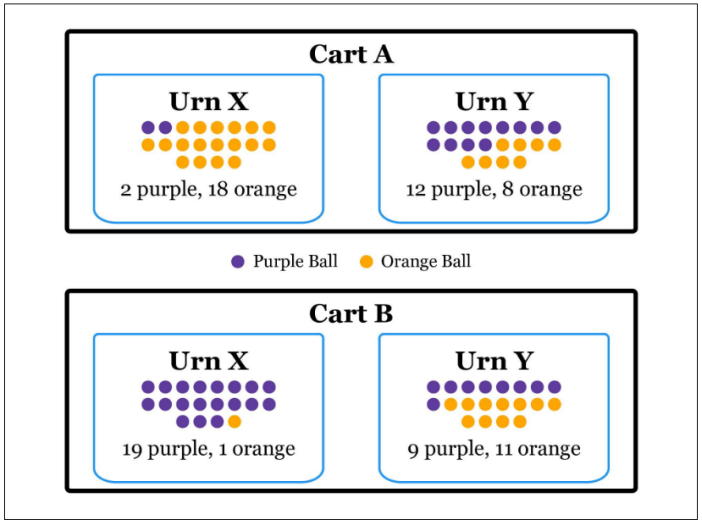
Figure D-1: Experiment Overview

Part 1 Instructions

General Information:

There is a 10-question quiz after Part 1 instructions. You will receive instructions for other parts once Part 1 is completed. Your decisions in Part 1 will NOT affect your earnings in other parts.

Your earnings in the experiment are denominated in experimental tokens, which will be exchanged at a rate of **150** tokens = **$1**.



Carts, Urns, and Balls

Part 1 contains **2** blocks, with **27 scenarios** in each block.

In each scenario, you will see **two carts** on your screen, cart A and cart B. Each cart contains **two urns**, urn X and urn Y.

There are purple and orange balls inside each urn. Each urn has **20** balls in total. You will see the exact number of purple and orange balls in each urn, as shown in the figure below.

The ball composition may vary across scenarios, so please check them carefully for each scenario.

- Example: *According to the Figure, urn X from Cart A has 2 purple balls and 18 orange balls; urn Y from Cart A has 12 purple balls and 8 orange balls. Urn X from Cart B has 19 purple balls and 1 orange balls; urn Y from Cart B has 9 purple balls and 11 orange balls.*

In each scenario, the computer randomly selects **one** urn using a two-step procedure, which is discussed on Page 2.

To Page 2

Figure D-2: Part 1 Instruction P1

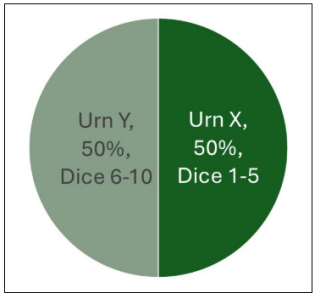
Part 1 Instructions

Urn Selection:

In each scenario, the computer randomly selects **one** urn from the two carts using a two-step procedure:

- **STEP 1 (Selecting an urn)**: The computer rolls a fair 10-sided die numbered 1 to 10, which determines which urn (X or Y) is selected.

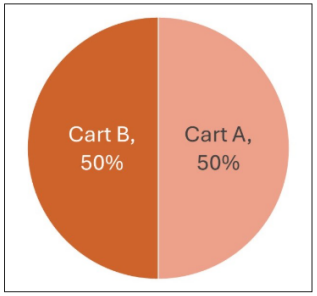
- Each die roll is independent.
- Please check the pie chart carefully, as it might change across scenarios.
- Example: *As the figure shows, in the example scenario, if the dice outcome is 1-5, urn X from cart A or urn X from cart B is selected (50% chance); if the dice outcome is 6-10, urn Y from cart A or urn Y from cart B is selected (50% chance).*



- **STEP 2 (Selecting a cart)**:

The computer uses a fair coin to determine from which cart (A or B) the urn is selected:
(1) If the coin toss is heads (50% chance), cart A is selected.
(2) If the coin toss is tails (50% chance), cart B is selected.

In each scenario, the coin toss and the die roll are independent.



To Page 1 To Page 3

Figure D-3: Part 1 Instruction P2

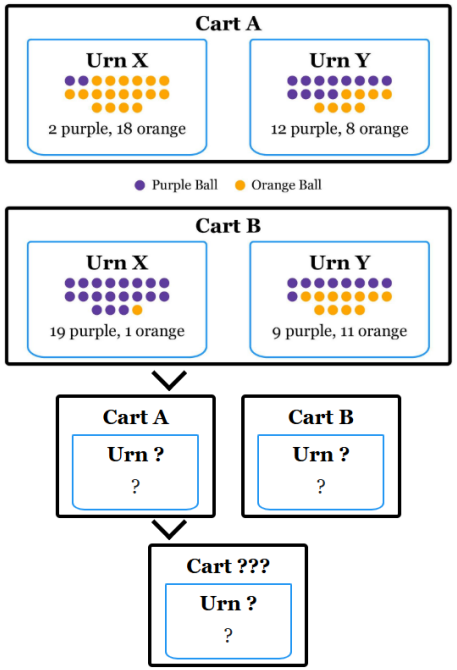
Part 1 Instructions

Example:

- **STEP 1 (Selecting an urn)**: in the example scenario, if the die roll is 1-5, urn X from cart A or cart B is selected (50% chance); if the die roll is 6-10, urn Y from cart A or cart B is selected (50% chance).
- **STEP 2 (Selecting a cart)**: The computer uses a fair coin determine from which cart the urn is selected. *If heads (50% chance), cart A is selected; if tails (50% chance), cart B is selected.*
- In this example scenario:

Suppose the die roll is 3 and the coin toss is heads, urn X from cart A is selected.
Suppose the die roll is 7 and the coin toss is heads, urn Y from cart A is selected.
Suppose the die roll is 1 and the coin toss is tails, urn X from cart B is selected.
Suppose the die roll is 9 and the coin toss is tails, urn Y from cart B is selected.

- Note 1: The die roll and coin toss are independently and randomly determined.
- Note 2: The selection rule for STEP 1 may vary across scenarios. *For instance, in one scenario, dice outcome 1-5 leads to urn X; in another scenario, dice outcome 1-7 leads to urn X.*
- Note 3: As for STEP 2, in each scenario, the computer tosses a coin to determine the cart.
- Note 4: You will NOT see which urn was selected, and you will NOT observe any die roll or coin toss. **One ball** is randomly drawn from the selected urn, and you will only see the **color** of the ball (purple or orange).



To Page 2 To Page 4

Figure D-4: Part 1 Instruction P3

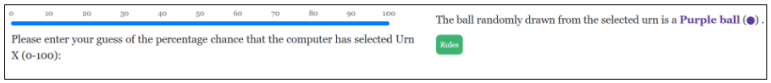
Part 1 Instructions

Your Guess:

After seeing the ball's color, you must use the slider to enter a number from **0** to **100**, as shown in the figure below.

This number represents your guess of the **probability (in percentage) that the computer selects Urn X**.

- Example: *Entering 0 means you are certain that the selected urn is Urn Y (from either cart), while entering 100 means you're certain that the selected urn is Urn X (from either cart).*
- Note: Your guess doesn't affect which urn was selected, as that was decided before your guess.



Earnings in Part 1:

Your total earnings in Part 1 will be the **sum** of your earnings from **all scenarios** in **both blocks**.

Your earnings in each scenario come from two sources, **urn payment** and **guess prize**.

(1) Urn payment: In each scenario, you will receive tokens based on which urn was selected by the computer, regardless of your guess.

- In Block 1, you will receive **5 tokens** if **urn Y** was selected. Urn payment for urn X is available **on your screen**.

(2) Guess prize: In each scenario, your guess determines the chance of winning the guess prize, **80 tokens**. The payment rule is designed so that you can secure the largest chance of winning the prize by truthfully reporting your best guess. You don't need to understand the payment rules, but if you're interested, you can click the button in the experiment interface to learn more.

- Example: *Suppose in a scenario, the selected urn is urn Y, you will receive 5 tokens as the urn payment. If you also win the guess prize, your total earning in this scenario is 5 + 80 = 85 tokens; otherwise, your total earning in this scenario is 5 + 0 = 5 tokens.*

To Page 3 To Page 5

Figure D-5: Part 1 Instruction P4

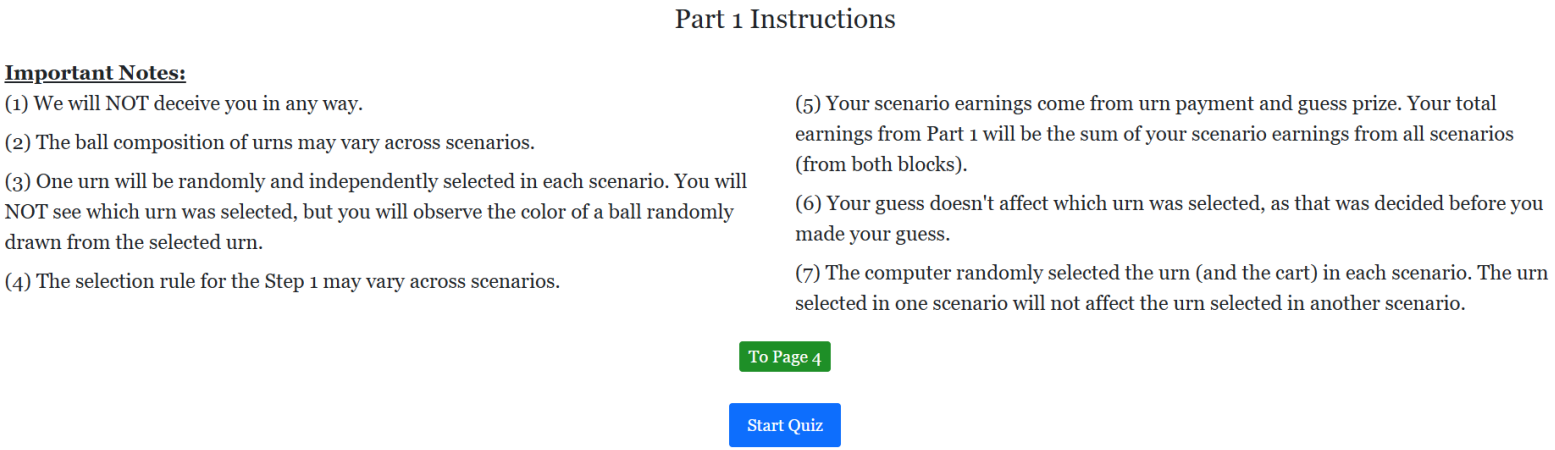


Figure D-6: Part 1 Instruction P5

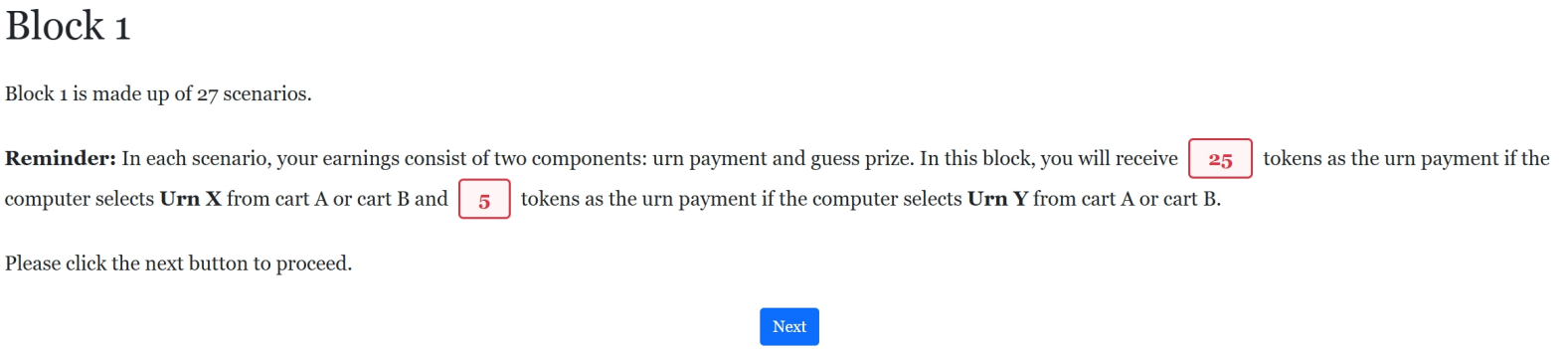


Figure D-7: Part 1 Block 1 Overview

Notes: Overview screen at the start of Block 1. The page initially has no Next button. Once all participants have arrived, the experimenter advances the session. The Next button then appears, and participants proceed to the first task.

Scenario 1 of 27

For Scenario 1, the urn and cart composition is summarized to the right.

One urn will be randomly selected by the computer. Below are the two steps:

Click for Step 1

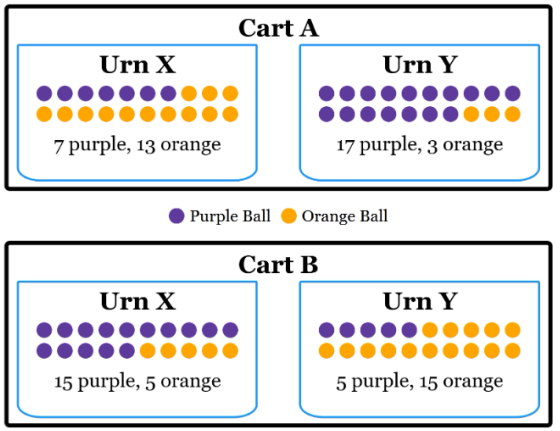


Figure D-8: Part 1 Example Task

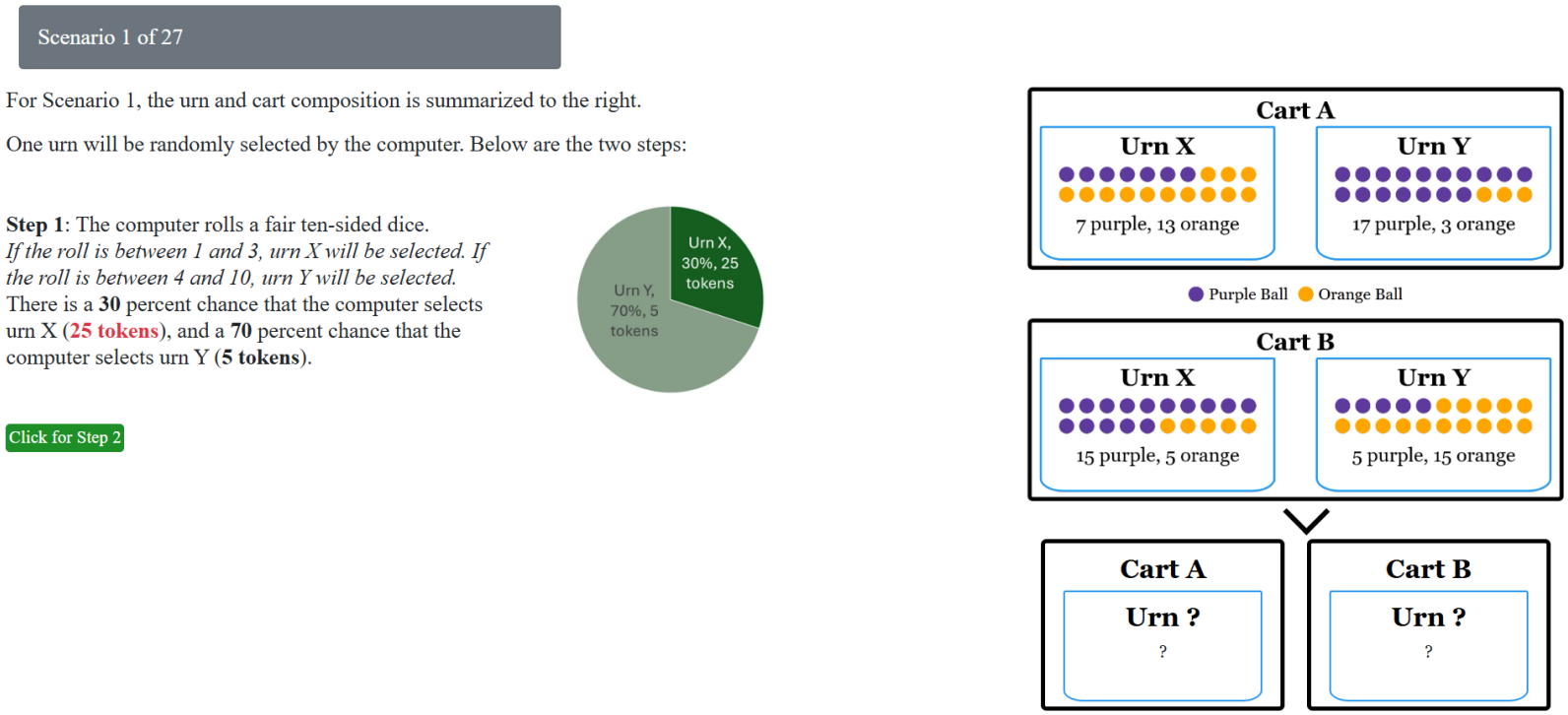


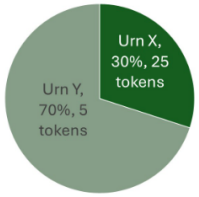
Figure D-9: Part 1 Example Task

Scenario 1 of 27

For Scenario 1, the urn and cart composition is summarized to the right.

One urn will be randomly selected by the computer. Below are the two steps:

Step 1: The computer rolls a fair ten-sided dice.
If the roll is between 1 and 3, urn X will be selected. If the roll is between 4 and 10, urn Y will be selected.
There is a **30** percent chance that the computer selects urn X (**25 tokens**), and a **70** percent chance that the computer selects urn Y (**5 tokens**).



Step 2: The computer uses a fair coin to determine from which cart (A or B) the urn is selected. If heads, the urn from cart A will be selected. If tails, the urn from cart B will be selected. There is a **50** percent chance that the computer selects cart A, and a **50** percent chance that the computer selects cart B.

The selected urn is labeled "Urn ?" from "Cart ???". The computer will draw a ball from the selected urn.

Click for Computer to Draw a Ball

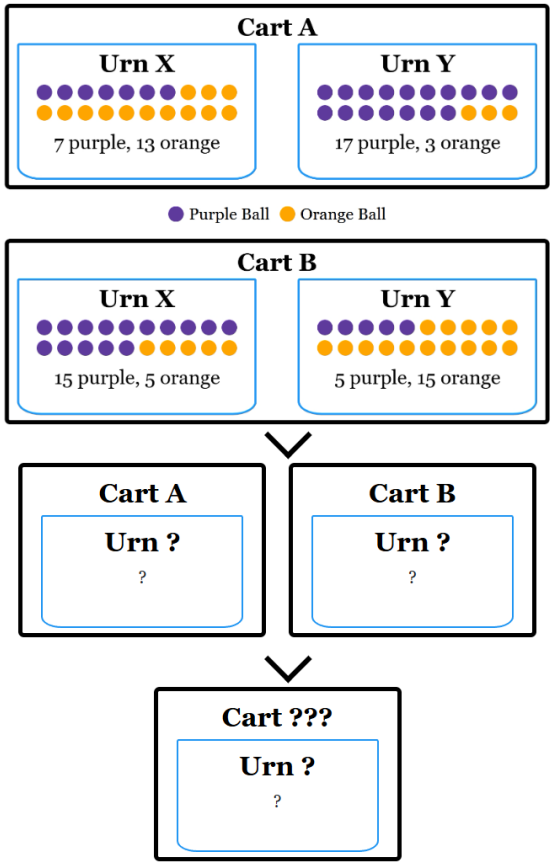


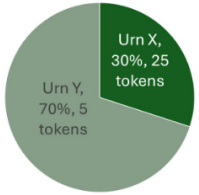
Figure D-10: Part 1 Example Task

Scenario 1 of 27

For Scenario 1, the urn and cart composition is summarized to the right.

One urn will be randomly selected by the computer. Below are the two steps:

Step 1: The computer rolls a fair ten-sided dice.
If the roll is between 1 and 3, urn X will be selected. If the roll is between 4 and 10, urn Y will be selected.
There is a **30** percent chance that the computer selects urn X (**25 tokens**), and a **70** percent chance that the computer selects urn Y (**5 tokens**).



Step 2: The computer uses a fair coin to determine from which cart (A or B) the urn is selected. If heads, the urn from cart A will be selected. If tails, the urn from cart B will be selected. There is a **50** percent chance that the computer selects cart A, and a **50** percent chance that the computer selects cart B.

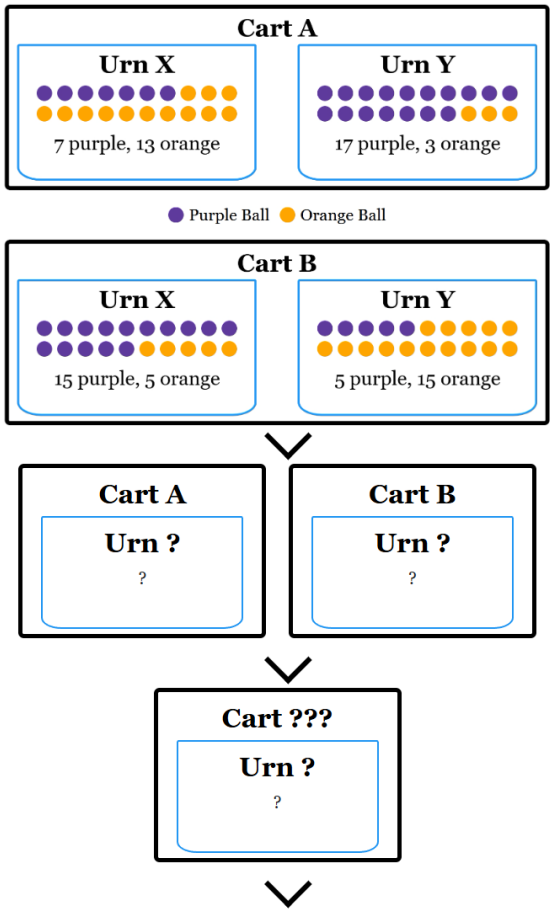
The selected urn is labeled "Urn ?" from "Cart ???". The computer will draw a ball from the selected urn.

You are asked to provide your best guess of the percentage chance (0-100) that the computer has selected **Urn X** from cart A or B.

You can **secure the largest chance of winning the prize by truthfully reporting your best** guess. Please click the green button for the prize details.

0 10 20 30 40 50 60 70 80 90 100

Please enter your guess of the percentage chance that the computer has selected Urn X (0-100):



The ball randomly drawn from the selected urn is a **Purple ball (●)** .

Rules Calculator

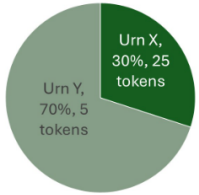
Figure D-11: Part 1 Example Task

Scenario 1 of 27

For Scenario 1, the urn and cart composition is summarized to the right.

One urn will be randomly selected by the computer. Below are the two steps:

Step 1: The computer rolls a fair ten-sided dice.
If the roll is between 1 and 3, urn X will be selected. If the roll is between 4 and 10, urn Y will be selected.
There is a **30** percent chance that the computer selects urn X (**25 tokens**), and a **70** percent chance that the computer selects urn Y (**5 tokens**).



Step 2: The computer uses a fair coin to determine from which cart (A or B) the urn is selected. If heads, the urn from cart A will be selected. If tails, the urn from cart B will be selected. There is a **50** percent chance that the computer selects cart A, and a **50** percent chance that the computer selects cart B.

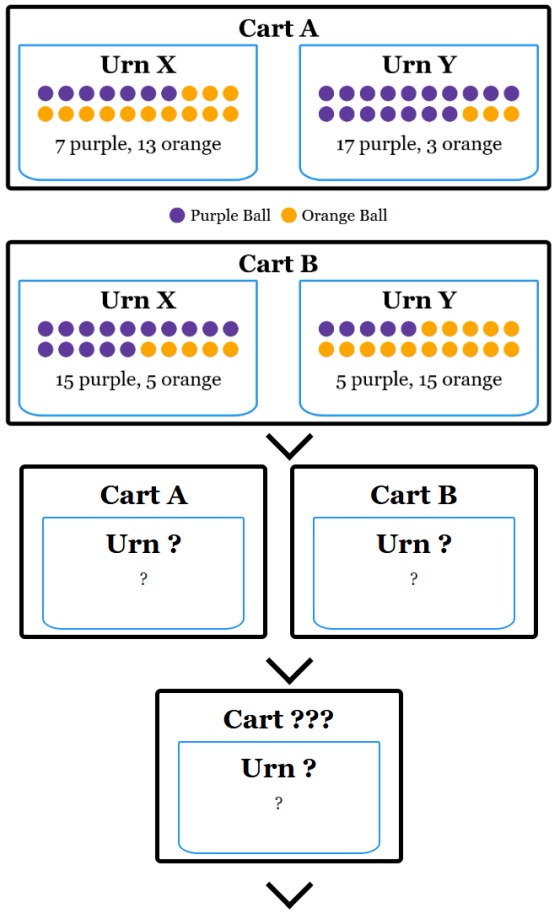
The selected urn is labeled "Urn ?" from "Cart ???". The computer will draw a ball from the selected urn.

You are asked to provide your best guess of the percentage chance (0-100) that the computer has selected **Urn X** from cart A or B.

You can **secure the largest chance of winning the prize by truthfully reporting your best** guess. Please click the green button for the prize details.

0 10 20 30 40 50 60 70 80 90 100

Your guess of the percentage chance that the computer has selected urn X: 76.7%



The ball randomly drawn from the selected urn is a **Purple ball (●)** .

Rules Calculator

Figure D-12: Part 1 Example Task

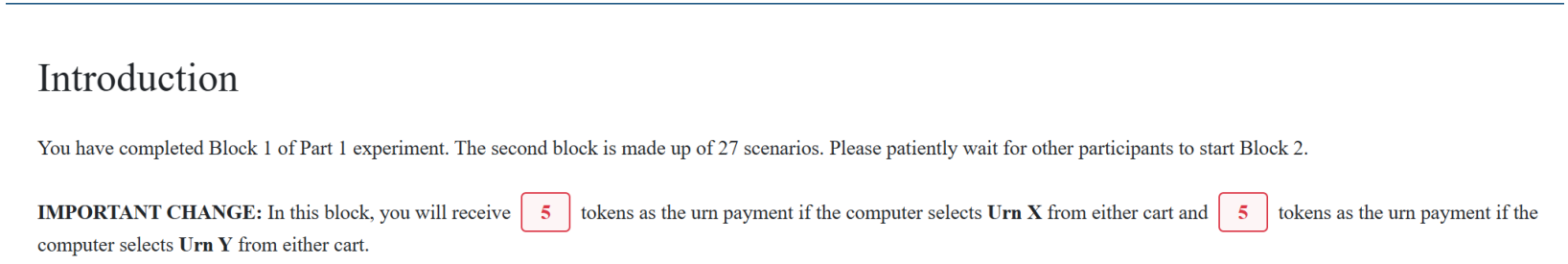


Figure D-13: Part 1 Block 2 Overview

Notes: Overview screen at the start of Block 2. The page initially has no Next button. Once all participants have arrived, the experimenter advances the session. The Next button then appears, and participants proceed to the first task.

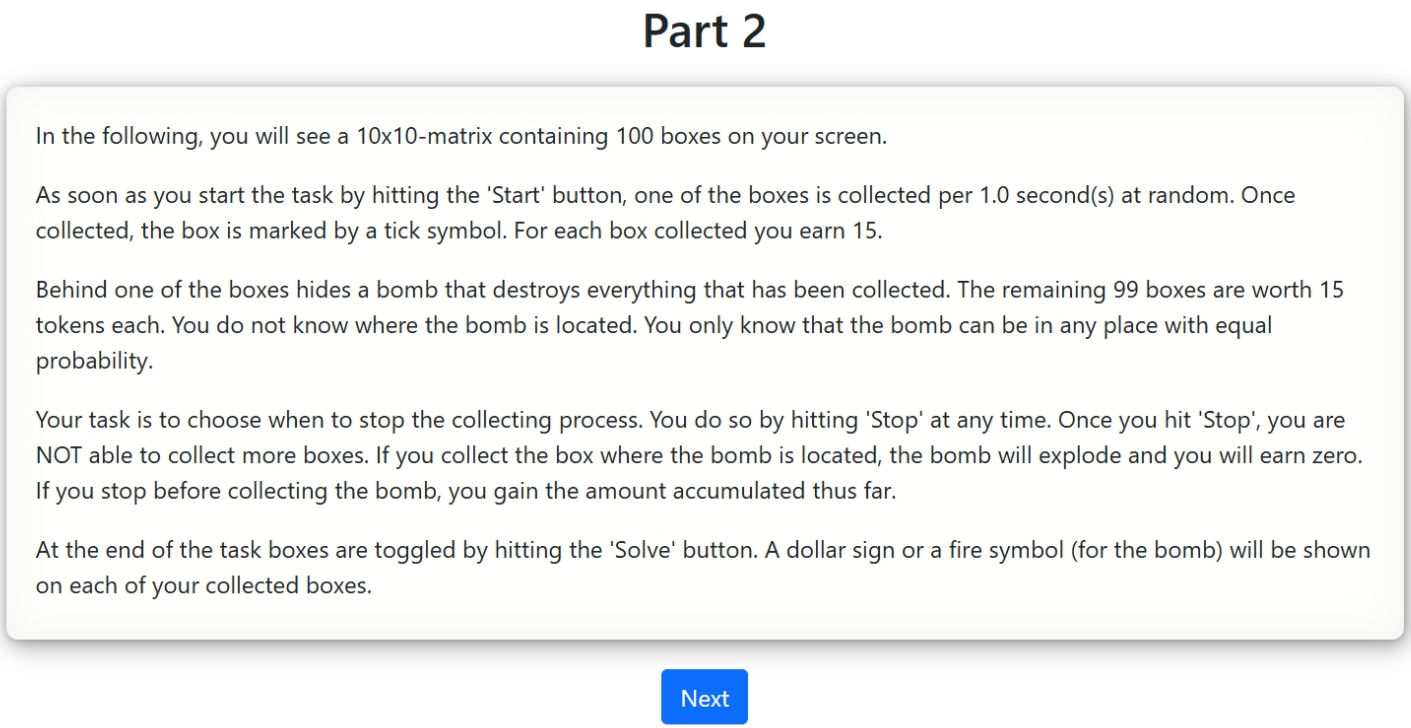


Figure D-14: BRET Task Instructions

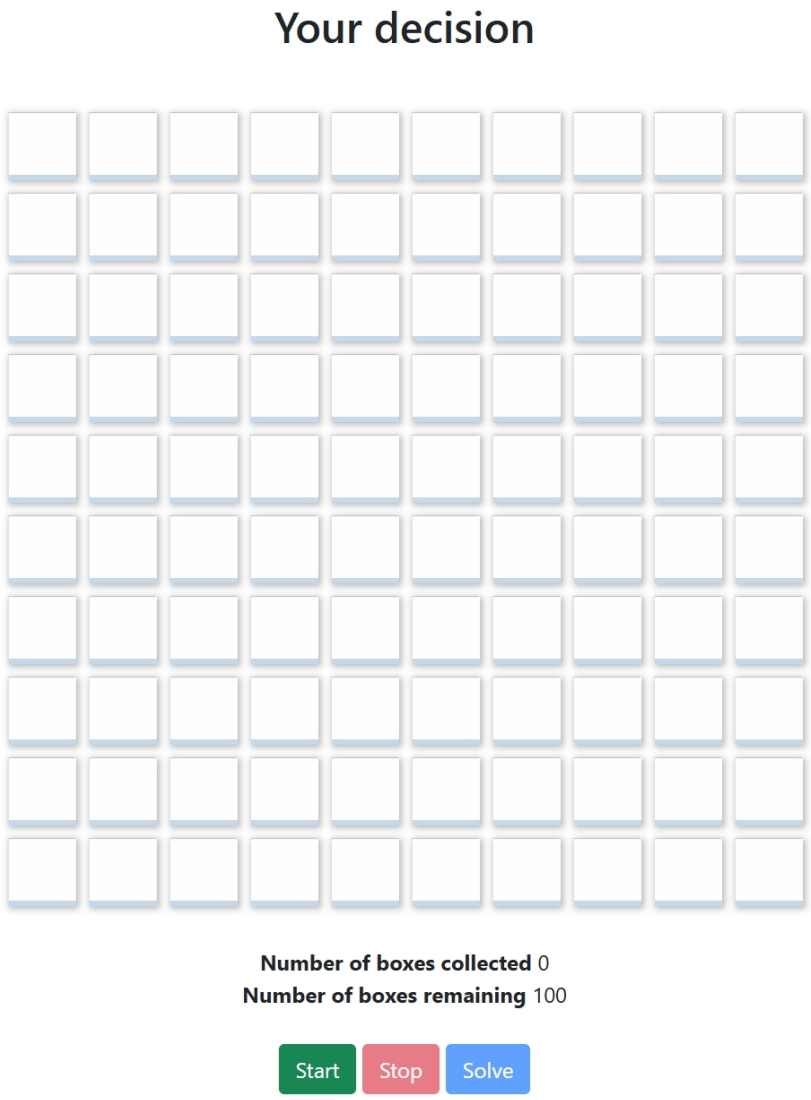


Figure D-15: BRET Task Interface

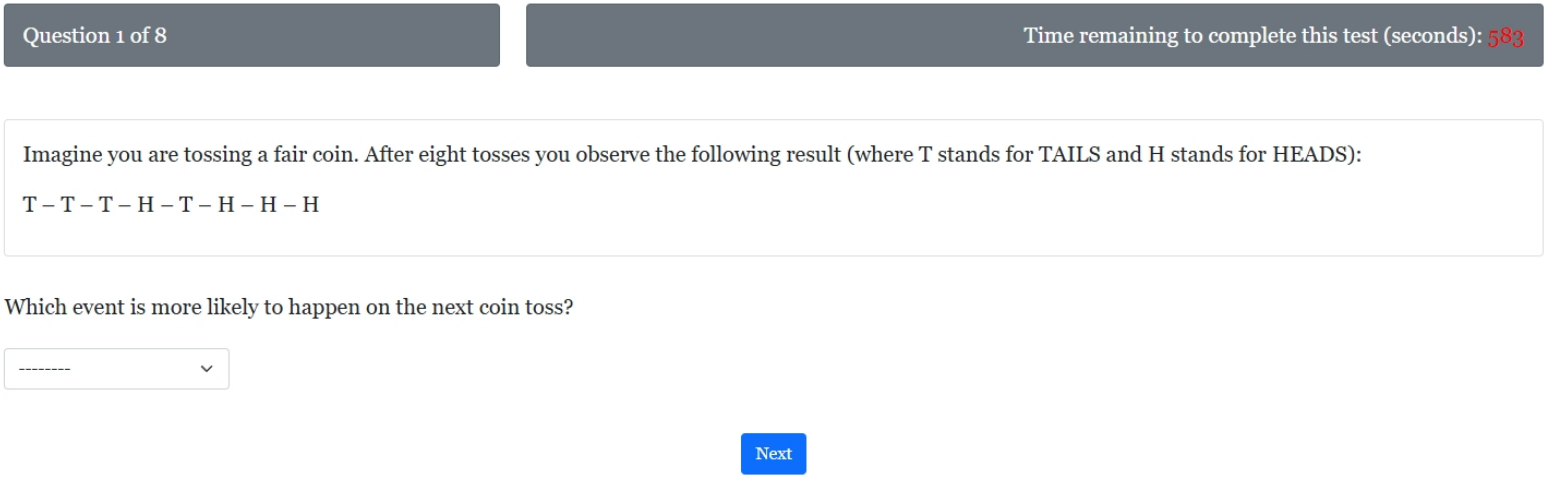


Figure D-16: Cognitive Test 1

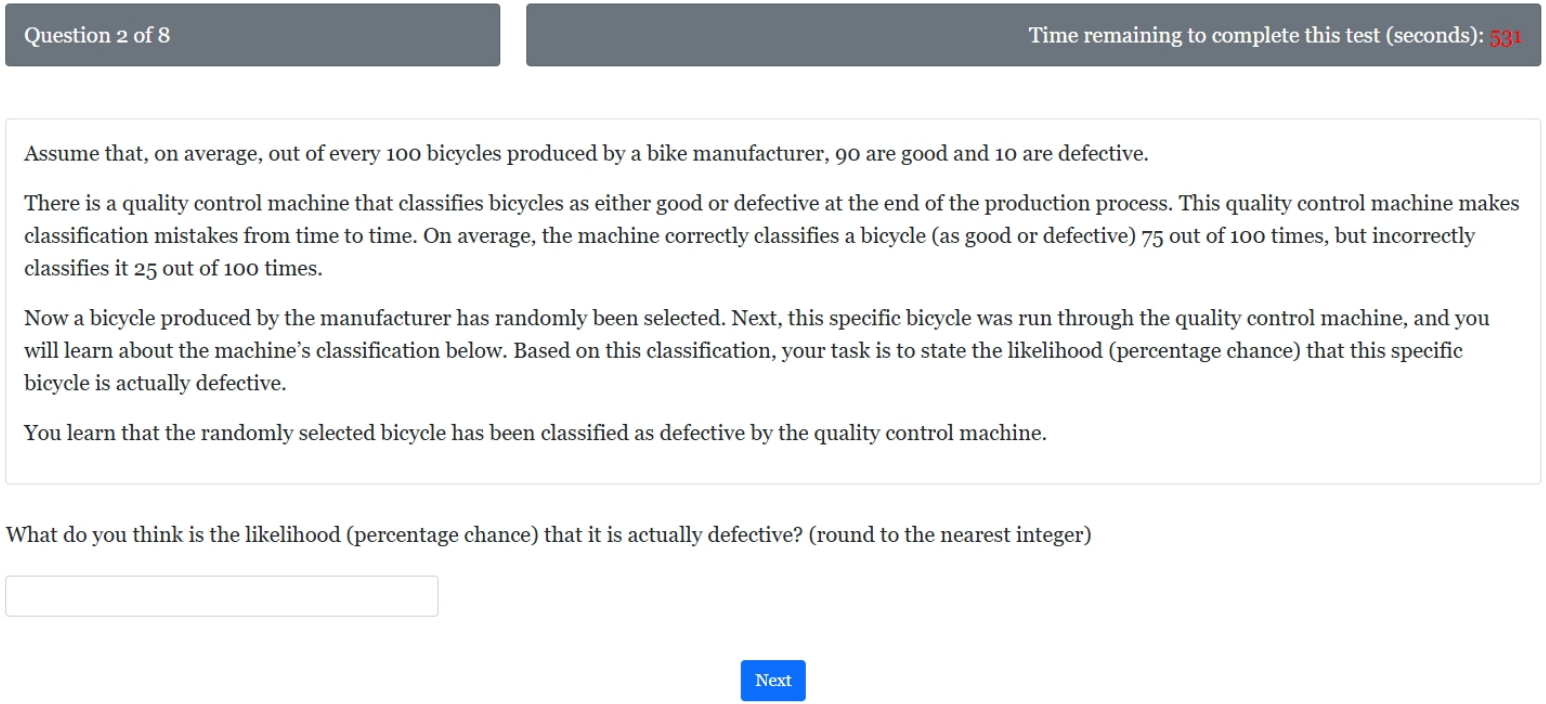


Figure D-17: Cognitive Test 2

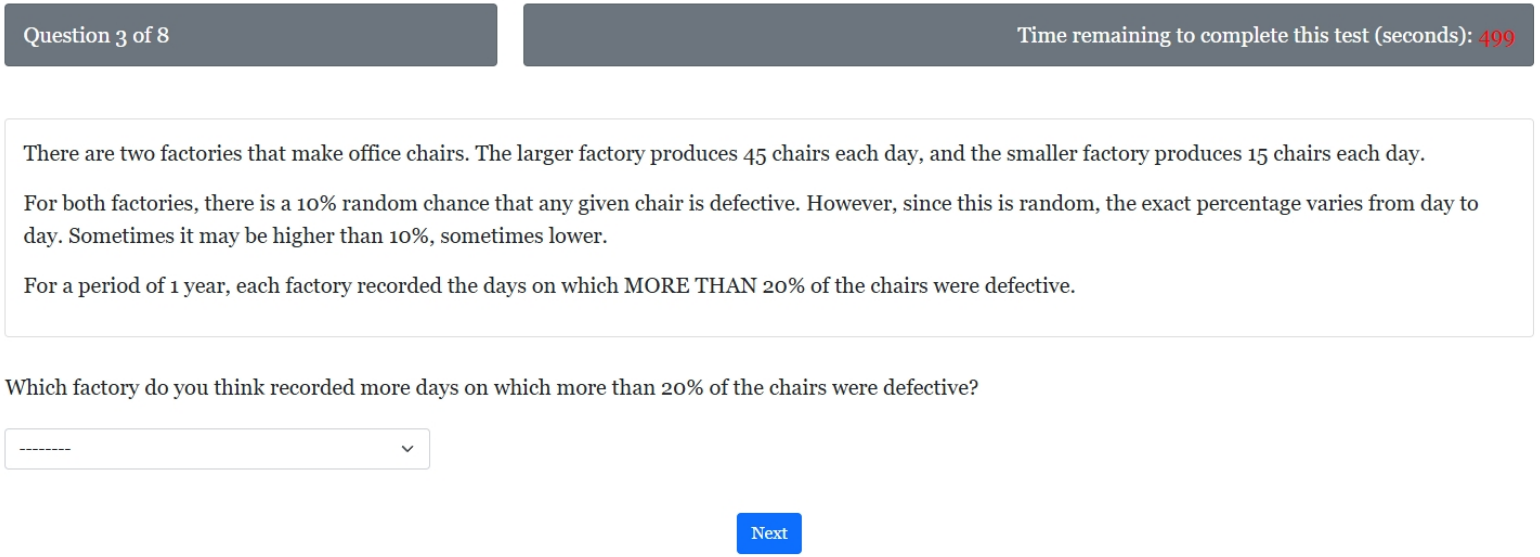


Figure D-18: Cognitive Test 3

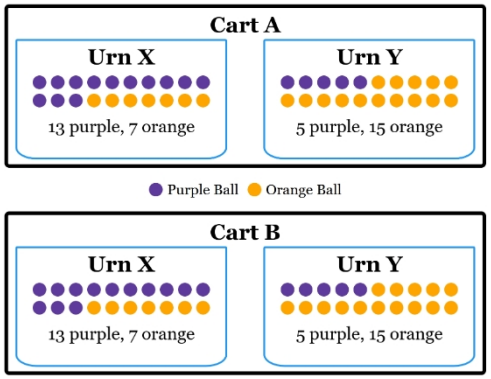
Question 4 of 8 | Time remaining to complete this test (seconds): 391

There are two carts. Each cart has two urns: Urn X and Urn Y. Each urn has in total 20 orange and purple balls. The computer follows a two-step procedure to select an urn: We secretly roll a fair ten-sided die and flip a fair coin. An urn is selected based on the outcomes. The compositions and selection procedure are as follows:

Step 1: The computer rolls a fair ten-sided die number from 1 to 10. If the die roll is between 1 and 5, urn X is selected. If the die roll is between 6 and 10, urn Y is selected.

Step 2: The computer rolls a fair six-sided die, with letter A on 3 sides and letter B on 3 sides. If the die roll is A, cart A is selected. If the die roll is B, cart B is selected.

Carts and Urns



Next, we drew one ball at random from the selected urn. The ball randomly drawn from the selected urn is a Purple ball.

What do you think is the likelihood (percentage chance) that the selected urn is Urn X? (round to the nearest integer)

Next

Figure D-19: Cognitive Test 4

Question 5 of 8 | Time remaining to complete this test (seconds): 355

You have a budget of 180 points. You can either keep it or use it to buy a company.

Bob is selling his company. The VALUE of Bob's company to him is either 20 or 120 points, but you do not know which. There is a 50% chance it is worth 20 points to him and a 50% chance it is worth 120 points to him.

Bob's company has a higher value to you than to Bob. If you acquire his company, it will pay you 1.5 times its value to Bob. Therefore, if the value of the company turns out to be 20 points for Bob, it would be worth 30 points for you. If the value of the company turns out to be 120 points for Bob, it would be worth 180 points for you.

The realized value is determined randomly by the computer, and you will not know the value until after you've made your decision.

You can make a PRICE offer to Bob of up to 180 points. Your earnings will be determined as follows:

- If you offer a **PRICE** that is **at least as high as** Bob's realized **VALUE**, Bob will accept your offer, and your earnings will be **Earnings** = **(Your budget)** + **1.5** * **(Bob's VALUE)** – **(the PRICE you offered)**
- If you offer a **PRICE** that is **less than** Bob's realized **VALUE**, you will not acquire his company and your profits will be **Earnings** = **Your budget**

What is the optimal bid for Bob's company? (to maximize your earnings)

Next

Figure D-20: Cognitive Test 5

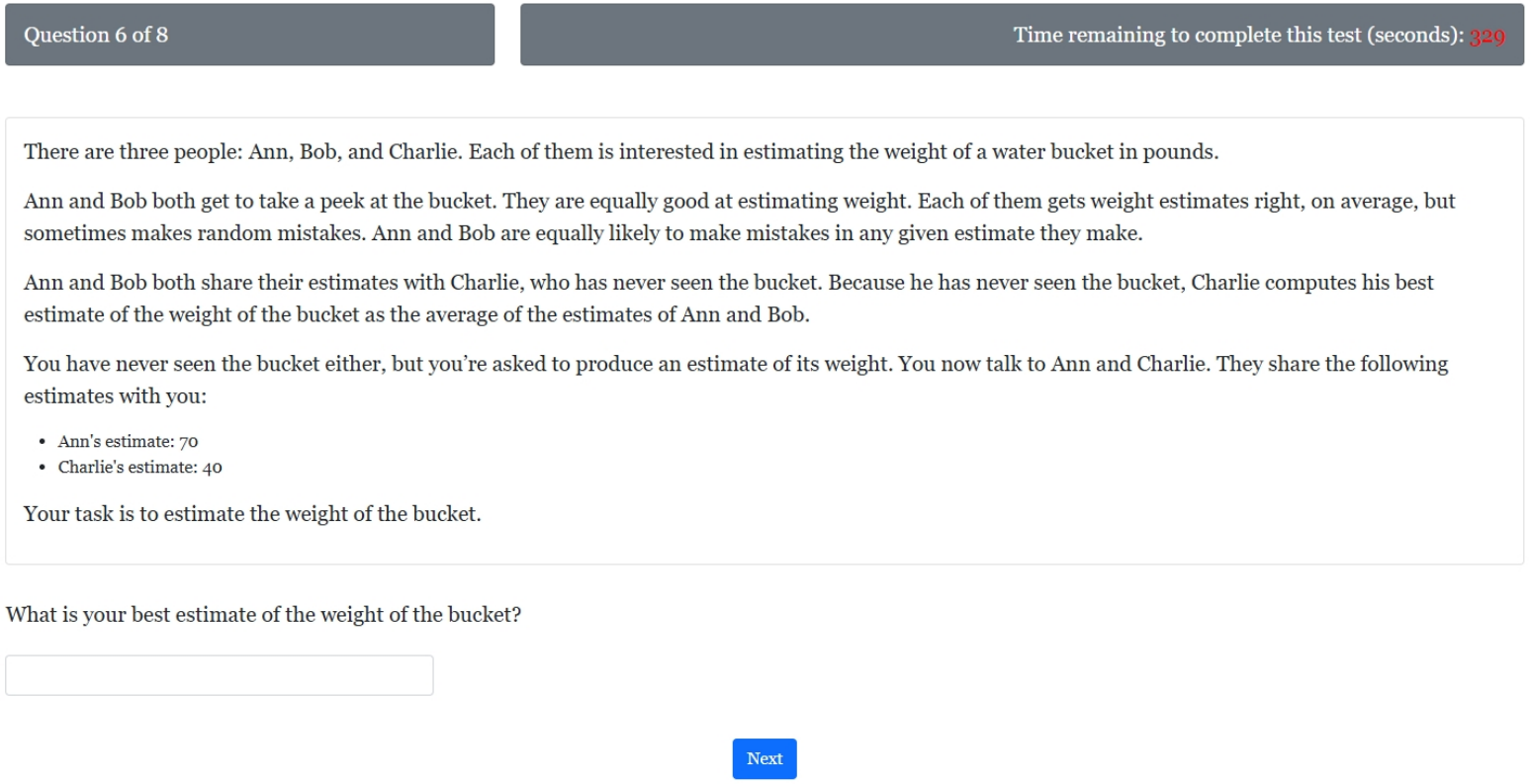


Figure D-21: Cognitive Test 6

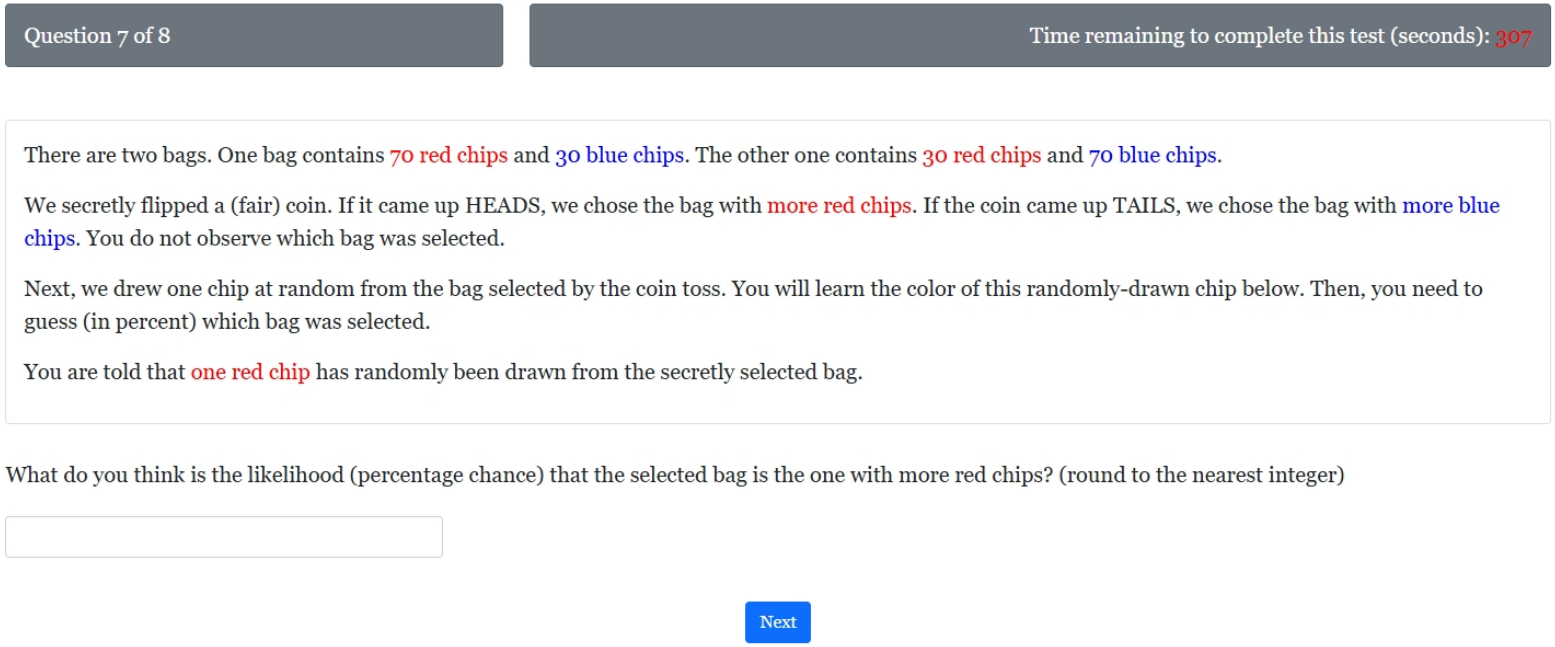


Figure D-22: Cognitive Test 7

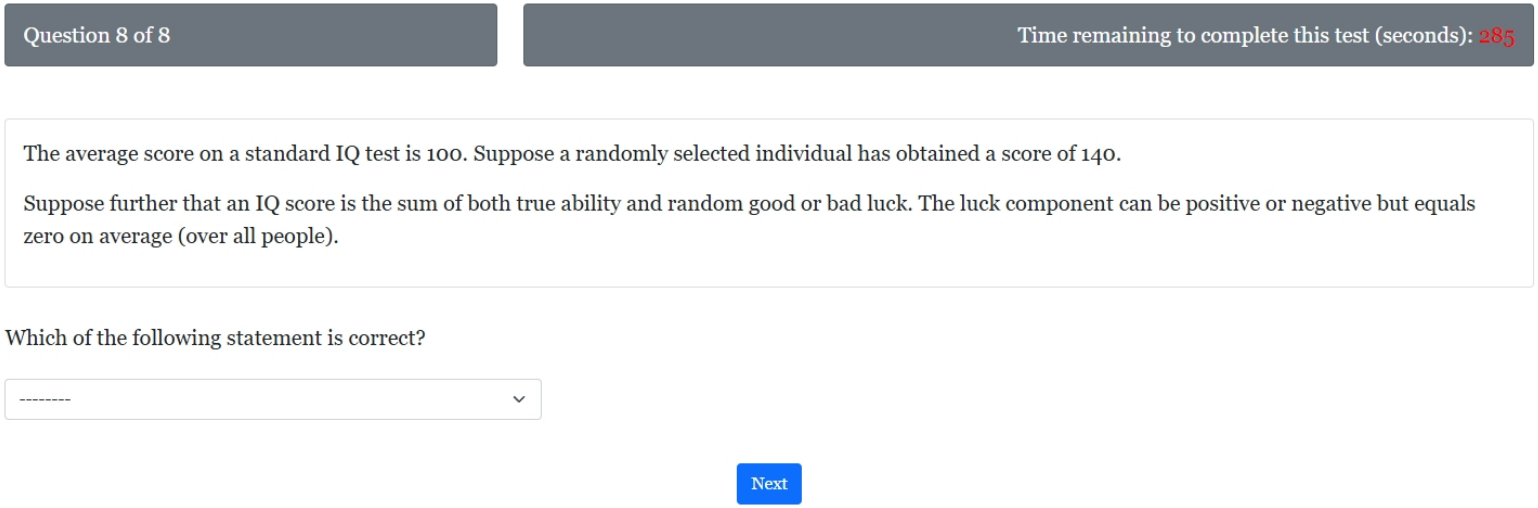


Figure D-23: Cognitive Test 8