Belief Formation and Polarization under Biased Information Sources*

Feng Zhu[†] Mengxing Wei[‡] Xueying Lyu [§]

August 1, 2025

Abstract

In the digital era, fragmented information disseminated by social media and algorithmic recommendations increasingly shapes public opinion, raising significant concerns about filter bubbles and polarization. This paper investigates the impacts of algorithmically induced information biases on belief updating as well as learning efficiency and polarization, compounded by intrinsic cognitive biases. Using a controlled laboratory experiment in which participants sequentially update their beliefs based on signals drawn from potentially biased sources, we document a persistent pattern of surprise-driven learning: contrary to the well-established confirmation bias, participants systematically overreact to signals that challenge their prior beliefs. When information originates from biased sources, both polarization and learning inefficiency increase. However, mere awareness of source bias is insufficient to mitigate these adverse effects. While awareness dampens the tendency to overreact to disconfirming information, it does not improve learning efficiency and even exacerbates polarization.

Keywords: Belief Updating; Polarization; Lab Experiment

[*] We thank Xue Xu and Shengbo Zhu for their helpful comments. We thank participants at the Workshop on Artificial Intelligence and Behavioral Economics hosted by Wuhan University. Zhu thanks the National Natural Science Foundation of China (No.72103102) and Research funds from LEBPS (No.ZX20220101) for financial support. Wei thanks the National Natural Science Foundation of China (No. 72003100). Lyu thanks the National Natural Science Foundation of China (No.72203111) for financial support.

[†] Center for Transnationals' Studies & Laboratory for Economic Behaviors and Policy Simulation, School of Economics, Nankai University, Tianjin, China; email: zf@nankai.edu.cn.

[‡] Corresponding author. School of Economics, Transnationals' Studies & Laboratory for Economic Behaviors and Policy Simulation, Nankai University, Tianjin, China; email: mengxing.wei@hotmail.com.

[§] Corresponding author. Center for Transnationals' Studies & Laboratory for Economic Behaviors and Policy Simulation, School of Economics, Nankai University, 94 Weijin Road, Nankai District, Tianjin, 300071, China; email: xueying@nankai.edu.cn.

1 Introduction

In an era marked by rapid technological advancement and an unprecedented surge in information, digital platforms, such as TikTok and Youtube, profoundly influence billions of individuals globally by reshaping the dissemination and reception of information. As individuals are increasingly exposed to exploding volumes of fragmented content, they form and update their opinions on unfolding events at an accelerating pace. Information is increasingly disseminated in fragmented forms via social media and algorithmic recommendations, subtly shaping public opinion and attitudes. This phenomenon, closely associated with the emergence of *filter bubbles* and opinion *polarization*, has garnered substantial attention (Pariser, 2011; Bakshy et al., 2015; Sunstein, 2018; Allcott et al., 2020). In particular, a growing body of literature expresses concerns that algorithmic biases systematically reinforce users' prior beliefs, as media outlets and recommendation algorithms increasingly tailor content to consumer preferences, thereby creating filter bubbles and exacerbating political and social polarization (Mullainathan, 2002; Allcott and Gentzkow, 2017; Bowen et al., 2023; Qian and Jain, 2024).

Nevertheless, the precise impact of filter bubbles on polarization remains an open question. While many scholars argue that these bubbles significantly drive polarization (Santos et al., 2021; Mosleh et al., 2021a,b), other studies identify only limited effects of algorithmic personalization (Hosseinmardi et al., 2021, 2024). Interestingly, exposure to opposing views can sometimes backfire, further intensifying polarization rather than mitigating it (Bail et al., 2018). Beyond algorithmically induced biases, inherent cognitive biases in human information processing complicate this issue further, especially given their potential interactions. Although previous research has separately examined algorithmic and cognitive biases (e.g., Levendusky, 2013; Martin and Yurukoglu, 2017; Bail et al., 2018; Guess et al., 2021), studies explicitly exploring their interaction remain scarce (Faia et al., 2024; Acemoglu et al., 2025).

Our study addresses this gap by systematically investigating how individuals update beliefs when exposed to sequences of potentially biased signals. Utilizing a unified theoretical framework and a controlled laboratory experiment, we explore how biased information sources influence belief updating, particularly in how individuals interpret and weigh new information in relation to their prior beliefs, and consequently affect learning outcomes such as learning efficiency and belief polarization. Additionally, we assess whether awareness of source bias acts as an effective mitigation strategy. To accomplish these goals, we conduct a controlled laboratory experiment comprising five treatments, wherein subjects predict the composition (number of red balls) of a virtual ambiguous urn containing 99 balls. In the first period, subjects receive private random draws and make initial predictions; subsequently, in periods 2–11, they update predictions based on subsets of another subject's initial draws. In the baseline treatment, information sources are unbiased and randomly selected from all other subjects in the same session. In contrast, biased treatments restrict information sources to subjects whose initial predictions are congruent (or incongruent) with the receiver's own. These biased treatments further vary by whether subjects are explicitly informed about source bias: in "awareness" treatments, subjects are notified that subsequent information in periods 2–11 originates from congruent (or incongruent) sources.

To analyze how participants interpret new signals, we develop a measure termed *interpreted red balls*, based on the previous and updated predictions. Namely, it is the implied signal that a rational agent would observe in order to make the updated prediction. Comparing interpreted and observed signals, we find that, in the baseline treatment where signals are unbiased, participants systematically overreact to surprising signals[1], reflecting a surprise-driven updating bias. This pattern opposes traditional confirmation bias but in line with the findings by Charness et al. (2021) and Kieren et al. (2020). Despite this surprise-driven updating bias, as unbiased signals accumulate, both learning inefficiency and belief polarization diminish over time.

When subjects unknowingly receive signals from biased sources, this overreaction to surprising signals persists, yet unnoticed biases systematically distort belief updating, elevating both learning inefficiency and polarization compared to baseline treatment. In the meantime, informing subjects about source bias reduces overreaction, prompting more literal interpretations of surprising signals, especially from incongruent sources. However, this awareness exacerbates polarization without significantly affecting average learning efficiency. Thus, our results indicate that awareness of source bias alone may not effectively mitigate polarization or enhance learning inefficiency.

[1] Signals are considered surprising if its color-majority differs from that of the previous prediction. See Section 4.2 for detailed discussions.

Our paper speaks to several strands of literature. First, our study contributes to the literature on biased belief updating, encompassing both the widely documented confirmation bias, the tendency to disproportionately seek, interpret, or weigh information in favor of existing beliefs (Wason, 1960; Levy and Razin, 2019; Williams, 2024), and a related but distinct phenomenon in which individuals seek or overreact to information contradicting their prior expectations (Charness et al., 2021; Kieren et al., 2020), which we term surprise-driven learning bias. While confirmation bias typically emerges in contexts involving motivated reasoning, such as self-related (e.g., ego or self-image) beliefs or deeply held ideological positions, surprise-driven learning bias is more likely to appear in relatively neutral settings, where prior beliefs are less entrenched or personally significant (Nickerson, 1998; Coutts, 2019; Charness et al., 2021). Our findings reveal that without any knowledge about source bias, participants overreact to signals that contradict their prior beliefs, suggesting supportive evidence of surprise-driven learning rather than confirmation bias. Furthermore, when informed of the source bias, participants interpret new signals more cautiously and the surprise-driven learning bias is substantially attenuated. Our experiment, centered on short-term beliefs in a controlled inference task with modest personal stakes, aligns closely with this neutral domain. Accordingly, our findings of overreaction to surprising signals do not contradict the established evidence on confirmation bias; instead, they highlight the nuanced role that motivation and belief strength play in shaping biased updating behavior.

Our study also aligns closely with the literature on social learning, which examines how individuals derive knowledge from observing others. Classic social learning models (e.g., Banerjee, 1992; Bikhchandani et al., 1992) demonstrate that sequential observational learning can lead to rational herding or informational cascades, where agents disregard their private information and imitate predecessors, resulting in inefficient outcomes. These theoretical predictions have been extensively tested in experimental studies (see, e.g., Anderson and Holt, 1997; Weizsäcker, 2010; Bikhchandani et al., 2024). In our setting, participants are able to observe a subset of signals of others, rather than directly observing their actions. Consistent with Grimm and Mengel (2020), we find that, as signals accumulate, information cascade does not occur, and the aggregated information from unbiased sources alleviates learning inefficiency, though it does not fully eliminate it.

More broadly, our paper speaks to the literature on media bias and its interaction with confirmation bias, social learning, and polarization. Persistent media biases, driven by consumer preferences and owner interests, are well-documented (Groseclose and Milyo, 2005; Mullainathan and Shleifer, 2005; Gentzkow and Shapiro, 2010). A growing body of literature examines how media bias affects beliefs and polarization (DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017). Such biases are not restricted to traditional media, and digital platforms amplify polarization through algorithmic personalization and recommendation systems (Flaxman et al., 2016). Proposed interventions include promoting cross-partisan exposure (Levy, 2021), facilitating probabilistic rather than binary judgments (Guilbeault et al., 2021), fostering economic self-interest that biases agents toward their initial positions, and adjusting ideological tolerance thresholds[2] (Axelrod et al., 2021), as well as manipulating network formation dynamics (Santos et al., 2021). Nevertheless, the efficacy of these interventions remains unclear. For instance, Bail et al. (2018) find that exposure to opposing views from opposing political party on social media can intensify political polarization, while Levy (2021) finds that exposure to counter-attitudinal information reduces affective polarization without significantly altering political beliefs. Our experimental findings suggest that biased information leads to increased polarization, compared with information from unbiased sources. However, merely making individuals aware about the source bias is insufficient to mitigate the emerging polarization.

In summary, our study contributes to the existing literature in three significant ways. First, we propose a unified framework designed to systematically investigate how inherent cognitive biases interact with exogenous information source biases in belief updating and social learning. A notable strength of our laboratory design lies in its capability to exogenously manipulate both information source biases and the awareness of such biases. This enables us to evaluate a practical remedy, raising subjects' awareness about the biased nature of information sources, and provides valuable insights relevant to regulating algorithm-driven recommendations prevalent in the contemporary digital landscape.

[2] In the Attraction-Repulsion Model in Axelrod et al. (2021), tolerance is the level of ideological differences that agents find attractive rather than repulsive.

Second, our findings present novel evidence that contrasts traditional views of confirmation bias. Instead of neglecting signals that conflict with pre-existing beliefs, participants in our experiment overreact, assigning disproportionately high weight to surprising, belief-challenging signals. This observation aligns with recent findings by Charness and Dave (2017) and Kieren et al. (2020), highlighting a phenomenon wherein certain individuals actively seek out and overemphasize information contradicting their prior beliefs — a behavior we term "surprise-driven learning". Crucially, this learning pattern proves robust and persistent across our experimental data, but its intensity is moderated when subjects become aware of the source bias, particularly when signals originate from belief-incongruent sources.

Finally, our analysis reveals nuanced outcomes regarding awareness as a mitigation strategy for biased information sources. Simply informing individuals of source bias proves insufficient to reduce the negative impacts on learning efficiency and social polarization. In fact, polarization intensifies, while the overall effect on learning efficiency remains limited. This is because the attenuation of overreaction to surprising signals is complicated by nature and context-dependent: although awareness mitigates individuals' tendency to overreact to surprising signals, such moderation does not consistently translate into improved accuracy in belief formation. This is because, in some cases, overreaction may partially offset the distortions caused by biased sources. This paradox highlights a key insight—awareness of bias moderates behavioral responses but does not reliably enhance learning outcomes.

The remainder of the paper proceeds as follows. Section 2 presents a conceptual framework and develops hypotheses; Section 3 details the experimental design and procedures; Section 4 presents our results; and Section 5 concludes.

2 Theoretical Framework and Hypotheses

Formally, consider an ambiguous urn with N balls, either red or blue, with a proportion $p \in (0, 1)$ of red balls. Each agent $i \in \mathcal{I} \equiv \{1, \ldots, I\}$ first draws n balls with replacement, where n is an odd number. We call this draw a *private sample*, in which the agent observes $s_i \in \{0, \ldots, n\}$ red balls. Let $M_R(s_i) = \mathbb{1}\{s_i > \frac{n}{2}\}$ indicate whether her private sample says "red-majority". It is obvious that $s_1, \dots, s_I$ are identical and independent random variables following Binomial(n, p).

After observing her private sample, agent i observes a sequence of T samples from the urn, which we call *partial signals* $\tilde{s_{it}}$ $(t = 1, \dots, T)$. Each $\tilde{s_{it}}$ is a subsample of size m drawn without replacement from another agent j's private sample, namely $\tilde{s}_{it}$ $\sim$Hypergeometric(n, s_j, m). Depending on treatment, these partial signals $\tilde{s_{it}}$ could be either from an unbiased source, or from a biased source. In the unbiased case, the source of $\tilde{s_{it}}$ is random and s_j is from $\mathcal{I} \setminus \{i\}$ with equal probability. In this case, since the random source s_j follows Binomial(n, p), we have agent i's partial signal $\tilde{s_{it}} \overset{\text{iid}}{\sim}$ Binomial(m, p) for $t \in \{1, \dots, T\}$. By contrast, in biased treatments, the source agent j is chosen based on whether j's initial majority matches (or contradicts) $M_R(s_i)$, the color majority of agent i's own private signal. This induces positive (or negative) correlation between s_i and $\tilde{s_{it}}$. In other words, $\tilde{s_{it}}$ is *filtered* before being observed by the agent, by $M_R(\tilde{s_{it}}) = 1$ or $M_R(\tilde{s_{it}}) = 0$. In this case, $\tilde{s}_{it}$ $\sim$Hypergeometric(n, s_j, m), where s_j either always comes from $S_1 = \{s_j : M_R(s_j) = 1\}$ or $S_0 = \{s_j : M_R(s_j) = 0\}$.

Our primary interest lies in understanding the influence of biased signals on learning, specifically when subjects simply use sample proportion to estimate the population.[3] That is, subjects neglect the potential correlation of signals, pool counts, and make the following naive estimates: (1) $\hat{p}_i^0 = \frac{s_i}{n}$, based solely on the private signal initially received; (2) $\hat{p}_i^{\text{new}} = \frac{\sum_{t=1}^T \tilde{s_{it}}}{mT}$, based solely on the partial signal received afterwards; and (3)$\hat{p}_i^{\text{both}} = \frac{s_i + \sum_{t=1}^T \tilde{s_{it}}}{n+mT}$, combining both signals. It is worth noting that in all treatments, subjects differ only in their partial signals $\tilde{s}_{it}$, while their private signals come from the same Binomial distribution. We therefore concentrate on $\hat{p}_i^{\text{new}}$ in the following analyses, as it isolates the persistent effects of biased information sources.[4] For notation simplicity, we omit both the superscript and subscript, and simply write partial signals and predictions as $\tilde{s}_t$ and $\tilde{p}$ hereafter.

First, we consider whether predictions converge to the true state as signals accumulate at

[3] Using sample proportions as estimators is statistically justified, aligning with the maximum likelihood estimator (MLE). From a Bayesian perspective, it also corresponds to the maximum a posteriori (MAP) estimator under an uninformative prior. A recent literature in cognitive psychology finds that people often naively assume that the sample is unbiased and use sample properties to estimate population analogs (Fiedler and Juslin, 2006; Juslin et al., 2007). In Section 4.1, we demonstrate that subjects' predictions generally conform to this principle.

[4] $\hat{p}_i^0$ corresponds to the case where agents receive multiple private signals and form predictions based solely on these observations; similarly, $\hat{p}_i^{\text{both}}$ reflects the scenario where agents incorporate sequences of both private and partial signals. However, neither case directly pertains to the primary setting of interest in our study.

the aggregate level, namely, whether the estimator is unbiased. Not surprisingly, unbiased information leads to unbiased prediction. On the other hand, signals from red-majority (blue-majority) sources tend to lead to an overestimation (underestimation) of the true proportion. Interestingly, when aggregating across different color-majority of private (initial) signals, this leads to an ex-ante unbiased prediction when signals are from congruent sources, but a biased prediction when signals are from incongruent sources. The following hypothesis summarizes these results.

Hypothesis 1. *Predictions converge to the true state when signals are drawn from an unbiased or congruent information source, but not when they originate from an incongruent source.*

Proof sketch: if the information source is unbiased, that is, if $\tilde{s}_t$ is drawn from a random source without being filtered, then the estimator is unbiased:

$$\mathbb{E}[\hat{p}] = \frac{\mathbb{E}[\sum_{t=1}^{T} \tilde{s}_t]}{mT} = \frac{T\,\mathbb{E}[\tilde{s}]}{mT} = \frac{mpT}{mT} = p$$

In contrast, when the information source is biased, whether congruent or incongruent, subsequent signals are always from a truncated population, depending on the color-majority of the initial private signal. As a result, predictions based on signals from red-majority sources $(M = 1)$ tend to overestimate the true state, while those from blue-majority sources $(M = 0)$ tend to underestimate it. More formally, we have

$$\mathbb{E}[\hat{p}|M = 0] < p < \mathbb{E}[\hat{p}|M = 1],$$

which follows from the fact that $\mathrm{Cov}(\tilde{s}_i, M) > 0$. Specifically,

$$\begin{aligned}\mathrm{Cov}(\tilde{s}_i, M) &= \mathbb{E}[\tilde{s}_i M] - \mathbb{E}[\tilde{s}] \times \mathbb{E}[M] = \Pr(M = 1)\,\mathbb{E}[\tilde{s}|M = 1] - mp \times \Pr(M = 1) \\ &= \Pr(M = 1)\,(\mathbb{E}[\tilde{s}|M = 1] - mp) > 0,\end{aligned}$$

which implies $\mathbb{E}[\hat{p}|M = 1] > p$, and the inequality for $M = 0$ follows analogously.

The discussion above pertains to a given agent whose color-majority $M \in \{0, 1\}$ is fixed based on the private signal observed. The ex ante distribution of M across agents depends on the true state p. As a result, the unconditional expectation $\mathbb{E}[\hat{p}]$ varies across different types of biased sources. If signals are drawn from congruent sources, namely color-majority that matches the color-majority of the agent themselves, then by the law of total expectation, we have $\mathbb{E}[\hat{p}] = p$. Although each agent's belief is biased conditional on M, these biases cancel out at the population level, yielding an unbiased aggregate prediction. In contrast, when signals are drawn from incongruent sources, the law of total expectation no longer applies. Therefore, $\mathbb{E}[\hat{p}] \neq p$ almost surely. □

Next, we consider two measures of learning efficiency, which we refer interchangeably as learning accuracy: mean squared error $MSE = \mathbb{E}[(\hat{p} - p)^2]$ and mean absolute error $MAE = \mathbb{E}[|\hat{p} - p|]$, under the following three cases: 1) signals are from unbiased and random information sources, 2) signals are from congruent information sources, and 3) signals are from incongruent information sources. We can obtain the following results.

Hypothesis 2. *Learning efficiency decreases under biased information sources, whether congruent or incongruent, than under unbiased random source.*

The above results mirror the ranking observed in Hypothesis 1 but offer a distinct insights. Hypothesis 1 focuses on information aggregation at the population level, where positive and negative errors can cancel out each other. Under such circumstances, congruent information, despite being biased, can still produce an accurate aggregate prediction of the true state, similar to that from an unbiased random information source. In contrast, Hypothesis 2 emphasizes individual-level learning accuracy or belief bias, where it is necessary to account for the magnitude of errors (e.g., through squared or absolute deviations), thus preventing directional errors from offsetting one another. The results indicate that, consistent with Hypothesis 1, the unbiased random information source yields the highest learning efficiency. However, in this individual-level context, congruent sources no longer match the efficiency of unbiased sources.

In addition to learning efficiency, another focus of our study is opinion polarization, measured by the dispersion (variance) of predictions. Although a biased (e.g. congruent) information source can sometimes still lead to an accurate average prediction at the aggregate level, systematically biased information shifts individual predictions (i.e. conditional expectations) away from the true state, making the distribution of opinions more dispersed. We summarize the hypothesis regarding polarization as follows.

Hypothesis 3. *Polarization increases under biased information sources, whether congruent or incongruent, than under unbiased random source.*

Proof. Since statements in Hypotheses 2 and 3 are closely related, we provide a unified proof. We first briefly summarize the setup and notation, followed by detailed calculations and comparisons of variance, mean squared error, and mean absolute error.

Setup

Each agent receives a private signal $s_1, \ldots, s_I \overset{\text{iid}}{\sim} \text{Binomial}(n, p)$. Let $s_{i_0} \sim \text{Binomial}(n, p)$ denote the private signal of the agent of interest, which partition the set of signals $\{s_1, \ldots, s_I\}$ into two groups: $S_1 = \{s_j : s_j > \frac{n}{2}\}$ and $S_0 = \{s_j : s_j \leq \frac{n}{2}\}$. Subsequent partial signals s_j are selected according to the following rules, under three distinct scenarios:

1. Unbiased: s_j is chosen uniformly from $S_1 \cup S_0$.
2. Congruent: If $s_{i_0} \leq \frac{n}{2}$, then s_j is uniformly chosen from S_0; otherwise, from S_1.
3. Incongruent: If $s_{i_0} \leq \frac{n}{2}$, s_j is uniformly chosen from S_1; otherwise, from S_0.

Given the selected source s_j, partial signals $\tilde{s}_1, \ldots, \tilde{s}_T \overset{\text{iid}}{\sim} \text{Hypergeometric}(N, s_j, m)$ are drawn, leading to the estimator defined as $\frac{1}{Tm} \sum_{t=1}^{T} \tilde{s}_t$. To distinguish among these three scenarios, we denote their respective estimators as $\hat{p}_{\text{U}}$, $\hat{p}_{\text{C}}$, and $\hat{p}_{\text{I}}$.

Variance

In the Unbiased scenario, $\tilde{s}_t \overset{\text{iid}}{\sim} \text{Binomial}(m, p)$, thus:

$$\text{Var}(\hat{p}_{\text{U}}) = \frac{1}{(Tm)^2} \text{Var}(\sum \tilde{s}_t) = \frac{1}{T^2 m^2} (Tmp(1-p)) = \frac{p(1-p)}{Tm}. \tag{1}$$

In the Congruent scenario, we first calculate the conditional variance

$$\text{Var}(\tilde{s}_t | s_j) = m \frac{s_j}{n} (1 - \frac{s_j}{n}) \frac{n-m}{n-1}.$$

Applying the law of total variance (on the sum $\tilde{s} = \sum_{t=1}^{T} \tilde{s}_t$), we obtain

$$\begin{aligned}\text{Var}(\tilde{s}) &= \mathbb{E}[\text{Var}(\tilde{s}|s_j)] + \text{Var}(\mathbb{E}[\tilde{s}|s_j]) \\ &= \mathbb{E}[Tm\frac{s_j}{n}(1-\frac{s_j}{n})\frac{n-m}{n-1}] + \text{Var}(Tm\frac{s_j}{n}).\end{aligned}$$

Since $s_j \sim \text{Binomial}(n, p)$, we have $\mathbb{E}[s_j] = np$, $\text{Var}(s_j) = np(1-p)$, and $\mathbb{E}[\frac{s_j}{n}(1-\frac{s_j}{n})] = p(1-p)(1-\frac{1}{n})$. Direct computation yields:

$$\mathbb{E}[\text{Var}(\tilde{s}|s_j)] = Tm\frac{n-m}{n-1}p(1-p)(1-\frac{1}{n}) = Tm\frac{n-m}{n}p(1-p),$$

and

$$\text{Var}(\mathbb{E}[\tilde{s}|s_j]) = \text{Var}(Tm\frac{s_j}{n}) = T^2m^2\frac{\text{Var}(s_j)}{n^2} = T^2m^2\frac{p(1-p)}{n}.$$

Putting together, we have

$$\begin{aligned}\text{Var}(\hat{p}_{\text{C}}) &= \text{Var}(\frac{\tilde{s}}{Tm}) = \frac{1}{T^2m^2}\text{Var}(\tilde{s}) \\ &= \frac{1}{T^2m^2}\left(Tm\frac{n-m}{n}p(1-p) + T^2m^2\frac{p(1-p)}{n}\right) \\ &= \frac{p(1-p)}{nmT}(n-m+mT) = \frac{p(1-p)}{T}\left(\frac{1}{m}+\frac{T-1}{n}\right). \qquad (2)\end{aligned}$$

Comparing Equation (1) with Equation (2), it immediately follows that $\text{Var}(\hat{p}_{\text{U}}) < \text{Var}(\hat{p}_{\text{C}})$.

For the incongruent scenario, similar calculations lead to:

$$\begin{aligned}\text{Var}(\hat{p}_{\text{I}}) = \frac{1}{(Tm)^2}\text{Var}(\tilde{s}) &= \frac{1}{(Tm)^2}[T\,\mathbb{E}[\text{Var}(\tilde{s}_t|s_j)] + T^2\,\text{Var}(\mathbb{E}[\tilde{s}_t|s_j])] \\ &= \frac{1}{mT}\,\mathbb{E}[\frac{n-m}{n-1}\frac{s_j}{n}(1-\frac{s_j}{n})] + \text{Var}(\frac{s_j}{n}).\end{aligned}$$

Therefore,

$$
\begin{aligned}
\text{Var}(\hat{p}_{\text{I}}) - \text{Var}(\hat{p}_{\text{U}}) &= \frac{1}{mT}\,\mathbb{E}[\frac{n-m}{n-1}\frac{s_j}{n}(1-\frac{s_j}{n})] + \text{Var}(\frac{s_j}{n}) - \frac{p(1-p)}{mT} \\
&= \text{Var}(\frac{s_j}{n}) + \frac{1}{mT}\left[\frac{n-m}{n-1}\,\mathbb{E}[\frac{s_j}{n}(1-\frac{s_j}{n})] - p(1-p)\right] \\
&= \text{Var}(\frac{s_j}{n}) + \frac{1}{mT}\left[\frac{n-m}{n-1}\left(p(1-p) - \text{Var}(\frac{s_j}{n})\right) - p(1-p)\right] \qquad (3) \\
&= \text{Var}(\frac{s_j}{n})\left(1 - \frac{n-m}{(n-1)Tm}\right) - \frac{p(1-p)}{Tm}\frac{m-1}{n-1} \\
&= \frac{p(1-p)}{n}\left(1 - \frac{n-m}{(n-1)Tm}\right) - \frac{p(1-p)}{Tm}\frac{m-1}{n-1} \\
&= \frac{p(1-p)(T-1)}{nT} > 0,
\end{aligned}
$$

where Equation (3) use the identity that $\mathbb{E}[\frac{s_j}{n}(1-\frac{s_j}{n})] = p(1-p) - \text{Var}(\frac{s_j}{n})$, which is true because $\mathbb{E}[\frac{s_j}{n}] = p$ and $\text{Var}(\frac{s_j}{n}) = \mathbb{E}[(\frac{s_j}{n})^2] - \mathbb{E}[\frac{s_j}{n}]^2$. This completes the proof that $\text{Var}(\hat{p}_{\text{U}}) < \text{Var}(\hat{p}_{\text{I}})$

Mean Square Error (MSE)

Both $\hat{p}_{\text{U}}$ and $\hat{p}_{\text{C}}$ are unbiased estimators, implying their MSEs equal their variances. Consequently, $\text{Var}(\hat{p}_{\text{U}}) < \text{Var}(\hat{p}_{\text{C}}) \iff \text{MSE}(\hat{p}_{\text{U}}) < \text{MSE}(\hat{p}_{\text{C}})$.

For the Incongruent estimator, since bias is present, we have:

$$\text{MSE}(\hat{p}_{\text{I}}) = \text{Var}(\hat{p}_{\text{I}}) + (\mathbb{E}[\hat{p}_{\text{I}}] - p)^2 \geq \text{Var}(\hat{p}_{\text{I}}) > \text{Var}(\hat{p}_{\text{U}}) = \text{MSE}(\hat{p}_{\text{U}}).$$

Mean Absolute Error (MAE)

We employ a normal approximation via the central limit theorem, obtaining:

$$\hat{p}_{\text{U}} \sim \mathcal{N}(p, \sigma_U^2),\, \hat{p}_{\text{C}} \sim \mathcal{N}(p, \sigma_C^2) \text{ and } \hat{p}_{\text{I}} \sim \mathcal{N}(p+\mu, \sigma_I^2),$$

with $\mu = \mathbb{E}[\hat{p}_{\text{I}}] - p$, and inequalities $\sigma_U^2 < \sigma_C^2$, $\sigma_U^2 < \sigma_I^2$.

For the comparison of MAE between $\hat{p}_{\text{U}}$ and $\hat{p}_{\text{C}}$, notice that $\hat{p}_{\text{U}} - p$ and $\hat{p}_{\text{U}} - p$ follow a normal distribution with zero mean, we have

$$\text{MAE}(\hat{p}_{\text{U}}) = \mathbb{E}[|\hat{p}_{\text{U}} - p|] = \sigma_U\sqrt{\frac{2}{\pi}},$$

and similarly, $\mathrm{MAE}(\hat{p}_{\mathrm{C}}) = \sigma_C\sqrt{\frac{2}{\pi}}$, thus clearly $\mathrm{MAE}(\hat{p}_{\mathrm{C}}) > \mathrm{MAE}(\hat{p}_{\mathrm{U}})$.

For $\hat{p}_{\mathrm{I}}$, using the similar technique, we have $\hat{p}_{\mathrm{C}} \sim \mathcal{N}(\mu, \sigma_I)$. This gives out

$$\mathrm{MAE}(\hat{p}_{\mathrm{I}}) = \mathbb{E}[|\hat{p}_{\mathrm{I}} - p|] = \sigma_I\sqrt{\frac{2}{\pi}}e^{-\frac{\mu^2}{2\sigma_I^2}} + \mu\left[2\Phi\left(\frac{\mu}{\sigma_I} - 1\right)\right] \equiv g(\mu, \sigma_I),$$

where $\Phi(\cdot)$ denotes the standard normal CDF. It is easy to see that $g(\mu, \sigma_I)$ is an even function in μ, making it without loss of generality to assume $\mu \geq 0$. Taking partial derivatives, one can show that $\frac{\partial g(\mu,\sigma_I)}{\partial \mu}$ and $\frac{\partial g(\mu,\sigma_I)}{\partial \sigma_I}$ are both positive. Consequently,

$$\mathrm{MAE}(\hat{p}_{\mathrm{I}}) = g(\mu, \sigma_I) > g(0, \sigma_I) > g(0, \sigma_U) = \mathrm{MAE}(\hat{p}_{\mathrm{U}}),$$

concluding the proof. □

In addition to examining learning outcomes such as accuracy and polarization, we also investigate how individuals update their beliefs in response to new information. Even in the absence of source bias, belief updating often departs from the rational benchmark due to cognitive limitations. Two well-documented deviations are confirmation bias, where individuals underweigh information that contradicts their prior beliefs, and surprise-driven updating, where individuals overreact to unexpected signals. Both patterns have empirical support in the literature (e.g., Charness and Dave, 2017; Levy and Razin, 2019; Kieren et al., 2020), but it remains unclear which tendency will dominate in our experimental setting, where individuals receive signals sequentially and infer an underlying state. To address this ambiguity, we formulate the following competing hypotheses:

Hypothesis 4a. *Individuals exhibit confirmation bias during belief updating.*

Hypothesis 4b. *Individuals exhibit surprise-driven updating bias in the learning process.*

Lastly, we consider the effect of agent awareness of source bias. Intuitively, if subjects are sufficiently sophisticated, awareness of source bias enables them to correct for the systematic deviation from the truth induced by the biased source, thereby debiasing the estimator. At the same time, awareness about the bias include more information available to the agent, hence resulting in an estimator with both lower opinion dispersion (variance) and lower learning inefficiency (MSE), according to the Rao-Blackwell theorem.[5]

Hypothesis 5. *When the information source is biased, making individuals aware of the bias improves learning efficiency and reduces opinion polarization.*

3 Experimental Design

3.1 Tasks

The experiment consisted of two parts. In Part 1, subjects completed the main task — the urn inference task — three times. In Part 2, subjects completed the Monty Hall task, designed to assess adherence to Bayesian updating, and a choice list task (Holt and Laury, 2002) to elicit risk preferences.[6]

In each **urn inference task**, there was a virtual ambiguous urn containing 99 balls, either red or blue. Subjects did not know the color composition and were required to estimate the number of red balls based on 11 sequential draws. These draws occurred in two stages.

In the first period, the computer randomly drew 5 balls from the urn with replacement, providing each subject a *private signal*. After observing their private signal (namely, the color of their own five colored balls), subjects privately reported their prediction about the number of red balls. In each of periods 2–11, subjects observed a *partial signal*, consisting of 3 balls randomly selected without replacement from another subject's private signal in period 1. For each period, the identity of the subject providing this partial signal was randomly selected from one of three pools, depending on treatment: (1) all other subjects, (2) subjects with congruent initial predictions (agreeing on red-majority or blue-majority status), or (3) subjects with incongruent initial predictions (disagreeing on red-majority or blue-majority status).[7]

[5] Specifically, let X be data available, $\delta(X)$ is any square-integrable estimator of a parameter θ, and $T(X)$ is any statistic, then the Rao-Blackwellized estimator $\delta^*(T) = \mathbb{E}[\delta(X)|T(X)]$ have both lower variance and MSE than the original estimator $\delta(X)$. One can prove these results using law of total variance, law of total expectation, and Jensen's inequality.

[6] Materials such as experimental instructions, screenshots of experimental interfaces and quiz questions are available in the Supplementary Materials.

[7] In the experiment, we mitigated potential confounding effects from group identity by framing the task as follows: first-period predictions of the number of red balls were classified into two intervals, $[0, 49]$ and $[50, 99]$.

After observing each *partial signal* (namely, the color of the three balls), subjects privately reported their updated prediction about the number of red balls.[8] All predictions across the 11 periods were incentivized following the quadratic scoring rule: payoff $= 150 - 0.015 \times$ $(\#\text{actual red balls} - \#\text{predicted red balls})^2$, with one randomly selected period determining the payoff for each urn.

Additionally, at the end of period 11, subjects reported the maximum perceived discrepancy between their final estimate and the actual number of red balls, reflecting their uncertainty about their predictions. This response, along with the post-experiment survey, received a flat payment.

The urn inference task was repeated three times in Part 1, each time with a different color composition (55, 44, and 36 red balls).[9] The order of these compositions followed a Latin square design, with each treatment's three sessions randomly assigned to the three different orders. Feedback regarding actual urn compositions was withheld until the experiment concluded.

In the **Monty Hall task**, subjects were informed that one of three doors concealed a prize of 5 Yuan, while the other two concealed 1 Yuan each, with equal probability. After subjects selected one door, the computer revealed one of the unchosen doors containing 1 Yuan. Subjects then chose to either stick with their initial door or switch to the remaining unopened door. Payments in both the Monty Hall task and the risk-choice task in Part 2 depended on subjects' choices.

3.2 Treatments

The experiment included a baseline and a 2 × 2 factorial design. In the *baseline* treatment, the partial signals observed in period 2–11 were drawn from another subject randomly selected from the entire session, irrespective of congruence in their initial predictions in period 1.

In the factorial design, treatments varied along two dimensions: the source of the partial

In the congruent (incongruent) treatments, partial signals were from subjects whose predictions fell within the same (different) interval. See Section B in the Supplementary Materials for further details.

[8] When participants made predictions in periods 2–11, information on their own previous predictions and signals was displayed on the decision screen.

[9] We varied the ratio to test its effect. Section A in the Supplementary Materials shows that our main results are robust across these ratios.

signal (Congruent or Incongruent) and awareness of information bias (Aware or Unaware). In the Congruent treatments, subjects observed partial signals from others whose initial predictions were similar (both predicting red-majority or blue-majority). To avoid confounding group-identity effects, instructions avoided using terms like “group” or “team”, instead describing signals as originating from subjects with first predictions in the same numerical range ($[0, 50)$ or $[50, 99]$). In the Incongruent treatments, partial signals were drawn from subjects with differing initial predictions. Thus, both Congruent and Incongruent treatments provided systematically biased signals compared to the baseline.

The second treatment dimension concerned subjects’ awareness of the biased sources. In Aware treatments, subjects were informed explicitly about the congruence or incongruence of signal sources. In Unaware treatments, subjects were not informed about this bias, although signals remained biased, coming from congruent or incongruent sources. Table 1 summarizes these treatments and lists the number of sessions and subjects per treatment.

Table 1: Summary of Treatment Design

Treatment	Source	Aware of Source	# Sessions	# Subjects
Baseline	Random	Aware[a]	3	72
CU	Congruent	Unaware	3	72
IU	Incongruent	Unaware	3	72
CA	Congruent	Aware	3	72
IA	Incongruent	Aware	3	72

[a] In Baseline, subjects were informed that the balls in Periods 2–11 were from a random participant (other than themselves) in the session.

3.3 Procedures

The experiment took place at the Laboratory for Economic Behaviors and Policy Simulation (LEBPS) at Nankai University during February and March 2023. Participants were recruited from the lab’s standing subject pool, which consists of undergraduate and graduate students from various disciplines at Nankai University. No subject participated in the experiment more than once.

Upon arrival, subjects were randomly assigned to computer terminals by drawing a card from a pile of numbered cards. After being seated, they received written instructions for Part

1 of the experiment. The experimenter read the instructions aloud, after which subjects completed a set of comprehension questions. Only after all participants had answered all questions correctly did the experiment proceed. Instructions for Part 2 were distributed upon the completion of Part 1. Following Part 2, participants completed a questionnaire collecting demographic information (e.g., age, gender, past experiment experience, statistics knowledge), after which they are displayed with their final earnings on screen and received payment privately on site. The experiment was computerized and implemented using oTree (Chen et al., 2016).

In total, we conducted 15 experimental sessions involving 360 subjects. The average duration of a session was approximately 90 minutes, and subjects earned, on average, 76.6 Chinese Yuan (roughly 11.1 US dollars), including a 10 Yuan participation fee. Summary statistics for participant characteristics, including means and standard deviations, are reported in Table 2.

Table 2: Subject Characteristics by Treatment

Characteristics	**Treatment**				
	Baseline	CU	IU	CA	IA
Age	22.18 (1.97)	22.13 (1.88)	22.01 (1.87)	21.79 (1.54)	21.99 (1.66)
Male	0.25 (0.44)	0.38 (0.49)	0.38 (0.49)	0.41** (0.5)	0.35 (0.48)
Experiment Experience	0.65 (0.48)	0.56 (0.5)	0.56 (0.5)	0.65 (0.48)	0.61 (0.49)
Statistics Knowledge	0.74 (0.44)	0.68 (0.47)	0.64 (0.48)	0.7 (0.46)	0.82 (0.39)
Bayesian Rationality	0.1 (0.3)	0.12 (0.33)	0.1 (0.3)	0.15 (0.36)	0.08 (0.28)
Risk Preference	6.75 (1.68)	7.26 (1.81)	7.07 (1.7)	6.82 (1.78)	6.83 (1.76)
Lack of Confidence	22.1 (14.13)	21.21 (12.21)	22.62 (12.92)	20.51 (14.02)	20.85 (13.11)

Notes: The table reports the mean (standard deviation) of each variable across treatments, based on 72 subjects. *Male*, *Experimental Experience*, and *Statistics Knowledge* represent the self-reported fractions of male participants, participants who have previously participated in other experiments, and participants who have studied a probability or statistics-related course, respectively. *Bayesian Rationality* denotes the fraction of participants who correctly answered the incentivized Monty Hall task. *Risk Preference* refers to the switching point (ranging from 1 to 10) in the multiple price list task. *Lack of Confidence* measures the average perceived maximum difference between the participant's final prediction in the 11th period and the actual number of red balls in the urn across three urns. All characteristics do not exhibit statistically significant differences between the baseline and other treatments at the 5% level, except for the fraction of male participants in the CA treatment.

Notably, our experimental design diverges from classic social learning experiments in two aspects. First, our subjects observe sequences of partial signals from others rather than observing their actions. This design choice is motivated by two primary reasons. Firstly, it better represents real-world scenarios in which individuals are influenced primarily by information (news feed, tweets, video clips etc.), rather than observable actions, from diverse sources. Secondly, as demonstrated by Agranov et al. (2022), direct observation of signals enhances information aggregation when signals originate from random and unbiased sources, enabling clearer analysis of the effects introduced by biased sources and subjects' awareness of such biases.

A second key distinction is the use of a non-binary urn inference task with non-binary signals. This element is indispensable for our study because a binary signal setting would preclude understanding how identical signals might be interpreted differently when originating from distinct sources. In our setup, rational agents observing a partially informative signal — for instance, two blue balls and one red ball — from another subject who previously predicted more red balls[10], can infer that the remaining unrevealed balls are likely red. Hence, the actual interpretation becomes richer (two blue balls out of a total of three revealed balls), capturing nuances unattainable in a binary framework. Unlike binary-state settings where subjects must report probabilistic beliefs requiring complex Bayesian reasoning (see, e.g., Zou and Xu, 2023), our non-binary prediction task simplifies cognitive demands, offering a clearer and more intuitive decision-making context for participants.

4 Results

4.1 Descriptive results

We first analyze how subjects formed beliefs and made their first predictions based on their private signals (five balls) across all treatments. According to our assumption in Section 2, without prior information, subjects would estimate the urn's composition by multiplying the observed proportion of red balls in their sample by 99. Figure 1 illustrates the relationship between observed and predicted proportions.

The figure shows that, on average, subjects predict a higher number of red balls in the urn when observing more red balls in their private signal. However, their predictions systematically deviate from direct proportionality, shifting toward the midpoint (50). This adjustment is especially pronounced in extreme cases (samples consisting entirely of red or blue balls) and can be explained by risk aversion and behavioral attenuation (Enke and Graeber, 2023; Enke et al., 2024).

[10] Though predictions in period 1 are not directly observed, agents could still deduce others' period-1 prediction if they are aware that other players made congruent or incongruent predictions with themselves.

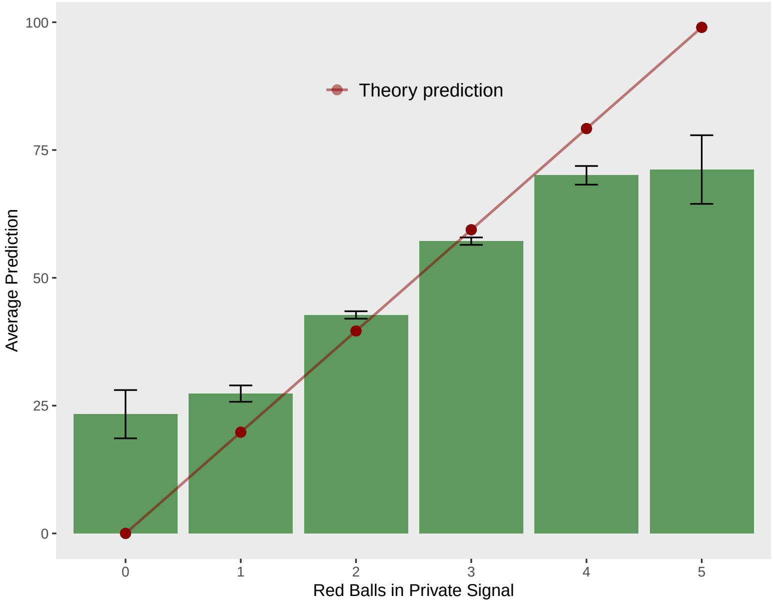


Figure 1: Average Prediction by Signals in Period 1

Note: This figure plots the average number of red balls predicted by subjects against the actual number observed in 5 balls drawn in the first period. Error bars indicate 95% confidence intervals. The red line depicts theoretical predictions from a *proportionist*, an individual who infers the population composition based on the sample proportion.

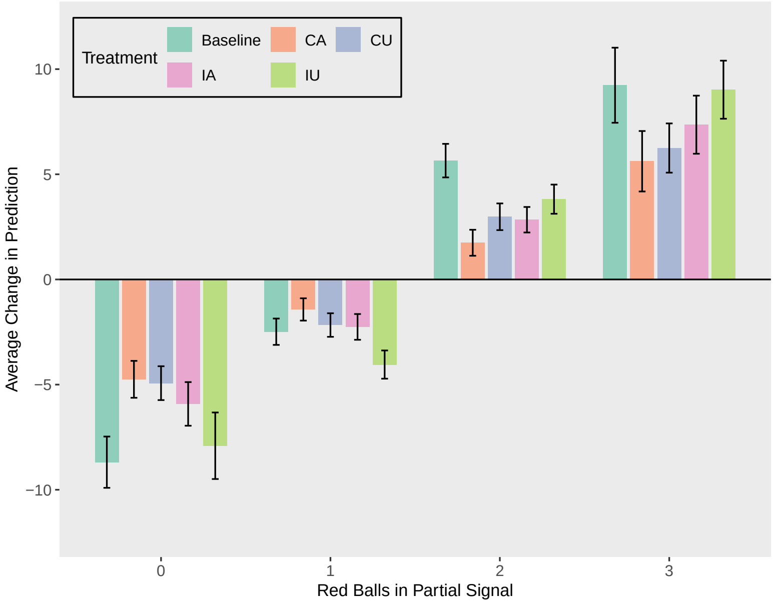


Figure 2: Average Prediction Adjustments by Signals in Periods 2–11

Notes: This figure shows the average change in predictions from the previous period following the observation of a new signal (a draw of three balls) in Periods 2–11 across treatments. Error bars indicate 95% confidence intervals.

Next, Figure 2 depicts how subjects adjusted their predictions over periods 2–11 in response to partial signals (three balls) from other subjects. Subjects decreased their predictions upon observing a blue-majority signal (0 or 1 red ball) and increased them following a red-majority signal (2 or 3 red balls). Moreover, the magnitude of adjustment grew with signal extremity — more pronounced adjustments occurred with stronger signals. Among all treatments, participants in the Baseline treatment exhibit the strongest reactions to new partial signals, followed by those in the IU treatment. In contrast, reactions in the CA and CU treatments are substantially attenuated — approximately half the magnitude observed in the Baseline treatment. This attenuation arises for two reasons. First, signals from congruent sources, compared to those from incongruent sources, are more likely to align with participants' pre-existing beliefs. This signal alignment reinforces previous predictions and therefore results in smaller belief adjustments. Second, awareness of the biased source also moderates reactions to signals: when participants are informed about the biased information sources, they exhibit greater skepticism and adjust their beliefs more cautiously, dampening their reactions to new signals. These two facts together lead to the strong reactions to new partial signals in the IU treatment.

Figure 3 plots participants' average predictions across treatments for different urn compositions, offering an overview of learning accuracy at the aggregate level. As shown in the figure, and broadly consistent with Hypothesis 1, predictions tend to converge toward the true state (i.e., the number of red balls, presented as red dashed horizontal lines) when signals are drawn from an unbiased information source (Baseline) or a congruent biased source (CU), but not when signals come from an incongruent biased source (IU). Introducing awareness of source bias reduces the overall prediction error, particularly under incongruent information, but the magnitude of this improvement is not always significant.

We now look at the paper's central question: how do learning inefficiency (the difference between subjects' predictions and the first-best benchmark[11]) and polarization (measured as the standard deviation of predictions (Fryer Jr et al., 2019; Santos et al., 2021) within a session) evolve across periods and treatments? Figure 4a illustrates the evolution of learning

[11] Because each session involves a finite number of participants and a finite draw of 5 balls from the urn, the benchmark prediction is calculated based on the complete set of signals observed by all participants in the same session during the first period.

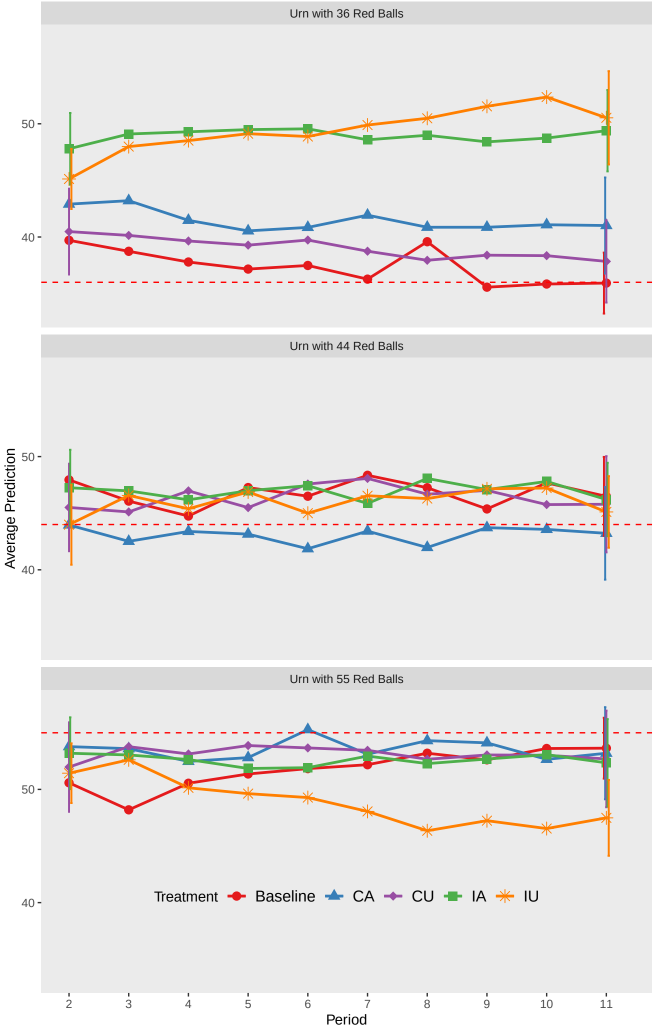


Figure 3: Average Prediction Dynamics by Urn in Periods 2–11

Notes: This figure presents the average predictions in Periods 2–11 across treatments for different urn compositions. The three rows correspond to urns containing 36, 44, and 55 red balls, respectively. The horizontal dashed red line denotes the true state in each case. Error bars indicate 95% confidence intervals.

inefficiency. Initially (period 1), all treatments exhibit similar levels of inefficiency due to the absence of biased partial signals. However, as partial signals accumulate (in periods 2–11), learning inefficiency gradually declines only in the baseline treatment. By the final period, learning inefficiency in the baseline treatment where signals are from random sources is significantly lower than those in the other four treatments, where information was biased. Among these four treatments, learning inefficiency remains at a relatively high level, irrespective of subjects' awareness of the biased source.

Similarly, Figure 4b shows polarization dynamics. Initially (period 1), polarization levels are comparable across treatments. Over subsequent periods, polarization decreases only in the

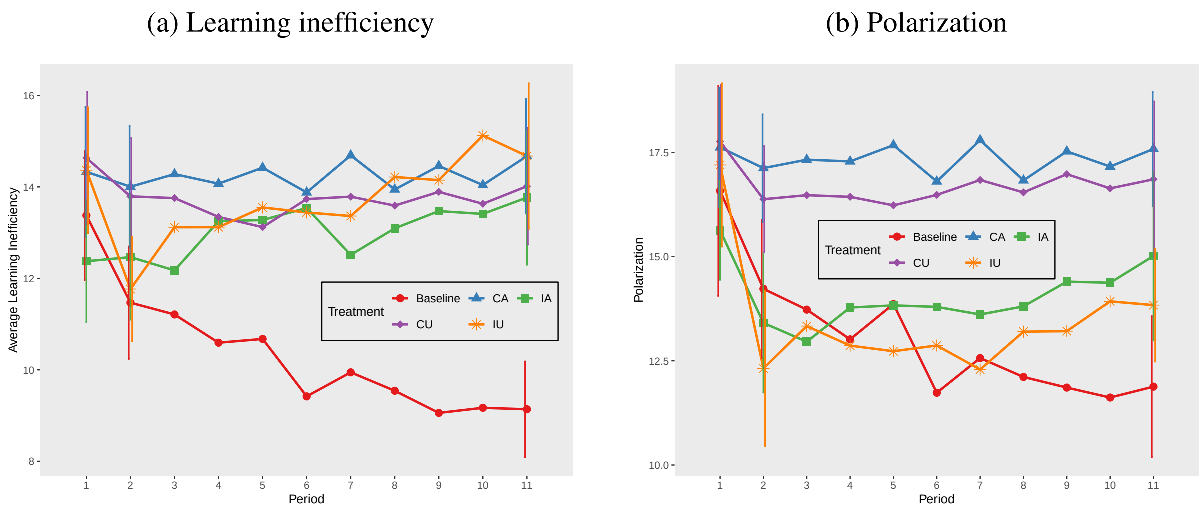


Figure 4: Learning inefficiency and Polarization Dynamics in different Treatments

Notes: Figure (a) plots the dynamics of learning inefficiency, measured as the absolute difference between subjects' predictions and the first-best benchmark, and figure (b) plots the dynamics of polarization, measured as the standard deviation of subjects' predictions within the same session, over 11 periods across treatments. Error bars indicate 95% confidence intervals.

baseline treatment. In contrast, biased treatments display persistent polarization. Notably, the first partial signal (period 2) reduces polarization in incongruent treatments, resembling the baseline. In contrast, polarization in congruent treatments remains stable in all period 2–11. Overall, incongruent treatments (IA, IU) exhibit lower polarization than congruent treatments (CA, CU) but still exceed baseline levels (Wilcoxon rank sum test, $p < 0.01$)[12]. Interestingly, subjects unaware of the bias exhibit lower polarization compared to subjects aware of the bias (IA v.s. IU $p = 0.06$, CA v.s. CU $p < 0.01$).

4.2 Surprising versus Confirming Signals

We begin by examining how participants update their beliefs in the Baseline Treatment, where partial signals received in period 2–11 are randomly drawn and unbiased. Even without source bias, individuals often exhibit a cognitive tendency known as *confirmation bias*. Considering this, we analyze how newly observed signals are processed relative to participants' beliefs in the previous period and examine the resulting impacts on learning inefficiency and belief polarization. Specifically, we categorize signals as *confirming* if they align with participants'

[12] All p-values reported in the main text are obtained from two-sided Wilcoxon rank-sum tests, unless explicitly stated otherwise.

previous predictions (e.g., a participant previously predicted a red-majority and observes a new signal suggesting the same). Conversely, signals that contradict previous predictions (e.g., a participant previously predicted a red-majority but observes a signal indicating a blue-majority) are defined as *surprising*.

To quantify how participants respond to confirming and surprising signals, we devise an outcome variable, *Interpreted Red Balls* (I_R), which is the implied signal that a proportionist[13] would observe in order to make the updated prediction. More formally, assuming that participants' prediction was p_{t-1} in the last period, and predicted p_t after observing a partial signal of 3 balls and denoting N_{t-1} to be the total number of balls observed up to (and including) period $t-1$, then the Interpreted Red Balls I_R satisfies:

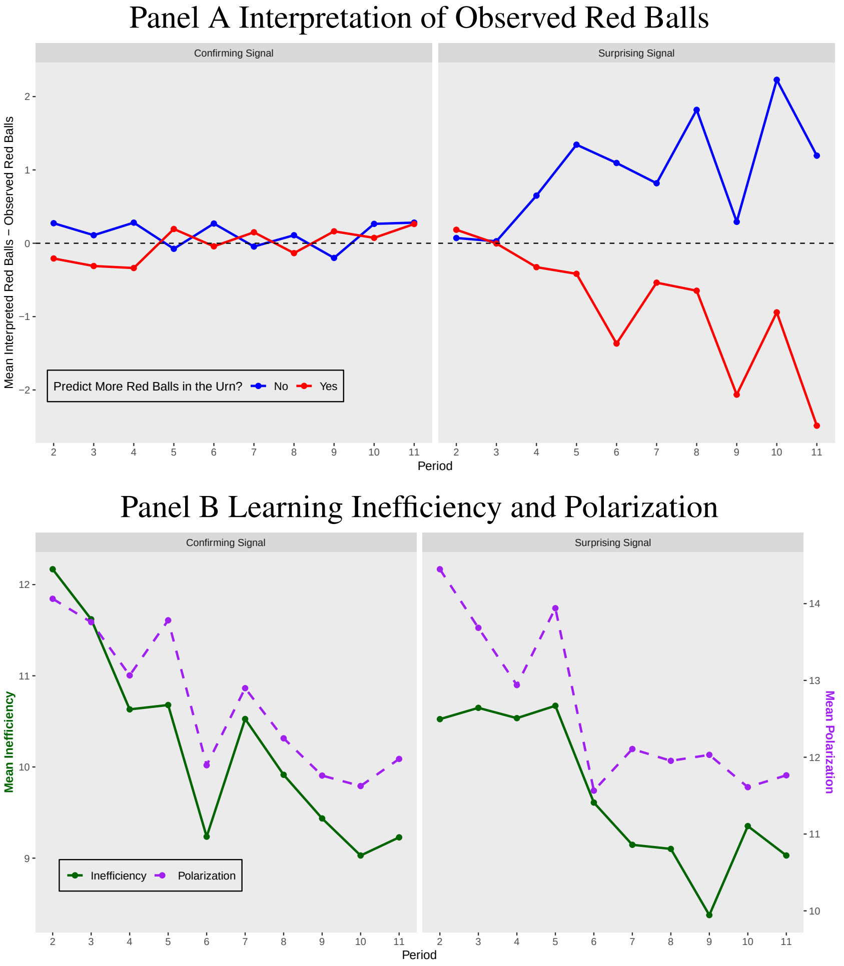
$$p_t = \frac{\overbrace{\frac{p_{t-1}}{99} N_{t-1} + I_R}^{\text{Total red balls}}}{\underbrace{N_{t-1} + 3}_{\text{Total balls observed}}} \times 99,$$

which implies $I_R = \frac{p_t}{99}(N_{t-1}+3) - \frac{p_{t-1}}{99}N_{t-1}$.

Panel A in Figure 5 illustrates the difference between the number of red balls participants interpret from each signal and the actual number observed, separately for confirming and surprising signals in Baseline Treatment. Since our primary focus is how participants respond to new signals after forming a belief based on existing private information, our plot focuses on Period 2–11. This applies to all remaining analyses hereafter, unless explicitly specified otherwise. Under confirming signals (left panel), participants' interpretations remain closely aligned with the actual data, hence the discrepancy fluctuate around zero, suggesting their belief updating is approximately proportionist. However, under surprising signals (right panel), participants exhibit systematic overreaction to such signals: those previously leaning toward red tend to over-interpret blue signals, while those previously favoring blue tend to over-interpret red signals. These deviations grow over time, reflecting systematic distortions in how participants process surprising information.

Panel B of Figure 5 demonstrates how surprising-driven belief updating bias affects learn-

[13] We use the word *proportionist* to refer to someone who predict the population based on (cumulative) sample proportion.


Panel A Interpretation of Observed Red Balls

Panel B Learning Inefficiency and Polarization

Figure 5: Interpretation of Observed balls, learning inefficiency, and polarization in Baseline Treatment

Notes: Panel A shows the dynamics of the average gap between the number of red balls observed and the interpreted number (inferred from predictions in Periods $t-1$ and t). The left subplot corresponds to confirming signals (aligned with predictions in $t-1$); the right subplot to surprising signals (contradicting predictions in $t-1$). Blue and red lines indicate the majority color in participants' predictions in Period $t-1$. Panel B plots the dynamics of learning inefficiency (green, solid lines) and polarization (purple, dashed lines) for confirming (left) and surprising (right) signals.

ing inefficiency and polarization. We define learning inefficiency (illustrated in the figure with green solid lines) by calculating the deviations of reported beliefs from the benchmark (first best). Because participants in a session only received a finite draw of 5 balls from the urn, we calculate the benchmark prediction based on the full set of signals visible to all participants in the same session and urn in the first period. The gap between each participant's prediction and benchmark prediction is defined as learning inefficiency. The figure reveals that, regardless of whether signals are confirming or not, learning inefficiency declines over time, reflecting effective aggregation of unbiased information. Polarization, measured as the standard deviation of participants' predictions within each session-urn-period cell and marked in the figure with purple dashed lines, exhibits a similar pattern. Despite early heterogeneity, participants' disagreement on the state gradually narrowed as more unbiased signals are accumulated, regardless whether signals are confirming or surprising.

To sum up, graphs in Figure 5 indicate that while surprising signals tend to elicit stronger, and often biased, individual responses in the short run, the random nature of signal draws ensures that such biases do not systematically accumulate. Since surprising signals occur in both directions (namely, observing "blue" while believing "red", or vice versa) across periods, individual overreactions tend to offset one another over time. As a result, even though ball interpretation is systematically biased when signals are surprising, both learning inefficiency and belief polarization converge toward those observed under confirming signals, demonstrating that reasonably efficient collective learning can still emerge even in the presence of persistent individual-level distortions. Below we summarize these results relating to Hypothesis 4.

Result 1. *Participants do not exhibit confirmation bias; instead, they overreact to surprising signals, reflecting a surprise-driven updating bias. However, as unbiased signals accumulate, both learning inefficiency and belief polarization diminish over time, despite this bias.*

At first glance, our findings may appear surprising, as participants exhibit a surprise-driven learning bias rather than the more frequently documented confirmation bias. It is worth noting that confirmation bias predominantly arises in contexts involving strongly held, personally motivated, or identity-related beliefs, where individuals typically discount or underweight contradictory information to maintain psychological consistency or self-enhancement (Nickerson, 1998). Conversely, surprise-driven learning bias is more prevalent in scenarios characterized by relatively neutral, short-term beliefs, where participants often attribute disproportionate significance to unexpected signals, leading them to over-adjust their beliefs in response to disconfirming evidence (Kuhnen, 2015; Coutts, 2019; Charness et al., 2021). In our experiment, participants engage in an incentivized urn inference task, forming short-lived beliefs without substantial personal or ideological attachment. Therefore, the observed pattern of overreaction to surprising signals aligns closely with existing literature.

Next, we quantify how surprising signals affect belief updating, with the following regression model:

$$Y_{ist} = \alpha_0 + \alpha_1 \text{ Surprising}_{ist} + \theta X_i + \lambda_t + \varepsilon_{ist} \tag{4}$$

where Y_{ist} denotes one of three outcome variables: (1) *signal interpretation*, defined as the difference between the number of red balls participant i interprets from the signal and the actual number observed in period t; (2) *learning inefficiency*, defined as the absolute deviation between a participant's prediction and the benchmark prediction, which is calculated based on the full set of signals visible to all participants in the same session and urn in the first period; and (3) *belief polarization*, measured as the standard deviation of predictions within each session-urn-period cell. For signal interpretation, we estimate separate regressions for participants with red-leaning and blue-leaning predictions in the last period to allow for asymmetric responses to surprising signals. For learning inefficiency and belief polarization, we estimate this equation using the full sample in Baseline Treatment. The key independent variable $Surprising_{ist}$ is a dummy equal to 1 if the signal contradicts the participant's previous belief and 0 otherwise.

The vector of controls X_i includes demographic variables such as age (*Age*) and a male dummy (*Male*), as well as a dummy for whether the subject has participated in similar experiments before (*Experiment experience*) and whether they have taken statistics-related courses (*Statistics knowledge*). We also account for cognitive and behavioral traits including performance on the Monty Hall task (*Bayesian rationality*), the switching point in the risk choice list (*Risk preference*), and the participant's reported expectation of their maximum prediction error (*Lack of confidence*). We also include period fixed effects, λ_t, to account for common learning dynamics over time. Standard errors are clustered at the participant level.

Table 3 reports OLS regression results[14] examining how participants interpret signals when faced with surprising information. The dependent variable is the difference between the number of red balls participants interpret from each signal and the actual number observed. Columns (1)–(3) focus on participants who predicted more red balls in the previous period, while Columns (4)–(6) focus on those who previously predicted more blue balls. Across all specifications, the coefficient on the *Surprising signal* dummy is large and statistically significant. Participants

[14] Unless otherwise noted, all regression analyses in this paper are estimated using OLS.

Table 3: Interpretation of observed balls with random signals

	Dependent Variable: Interpreted red balls - Observed red balls					
VARIABLES	Predict more red balls in the previous period			Predict more blue balls in the previous period		
	(1)	(2)	(3)	(4)	(5)	(6)
Surprising signal	-0.842***	-0.809***	-0.831***	0.910***	0.916***	0.866***
	(0.273)	(0.275)	(0.280)	(0.286)	(0.280)	(0.273)
Age		0.0717	0.0690		-0.0400	-0.0406
		(0.0993)	(0.0972)		(0.0331)	(0.0324)
Male		0.396	0.392		-0.159	-0.165
		(0.259)	(0.258)		(0.186)	(0.189)
Experiment experience		-0.114	-0.0755		0.103	0.115
		(0.249)	(0.243)		(0.187)	(0.187)
Statistics knowledge		-0.639	-0.653		0.533**	0.513**
		(0.505)	(0.496)		(0.210)	(0.208)
Bayesian rationality		-0.584	-0.553		0.506	0.520
		(0.415)	(0.409)		(0.375)	(0.382)
Risk preference		0.00690	0.00943		0.0503	0.0538
		(0.0534)	(0.0526)		(0.0580)	(0.0587)
Lack of confidence		-0.0300***	-0.0297***		0.0274***	0.0272***
		(0.00912)	(0.00889)		(0.00719)	(0.00726)
Observations	919	919	919	1,241	1,241	1,241
R-squared	0.031	0.083	0.107	0.040	0.093	0.110
Period FEs	NO	NO	YES	NO	NO	YES

Notes: The first three columns report the results when participants predict more red balls in the previous period, while the last three columns report the results when participants predict more blue balls in the previous period. Columns (1) and (4) report the results without individual characteristics, columns (2) and (5) report the results controlling for several individual characteristics, and columns (3) and (6) report the results further controlling for period fixed effects. Observations are from the Baseline. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

with red-leaning priors tend to underinterpret red signals, reporting roughly 0.8 fewer red balls than observed, while those with blue-leaning priors tend to overinterpret them by about 0.9. This pattern indicates that surprising signals systematically distort interpretation: individuals overweigh information that contradicts their existing beliefs.

Somewhat unexpectedly, participants with stronger statistical training exhibit even greater distortions in response to surprising signals. Rather than mitigating misinterpretation, higher cognitive sophistication appears to amplify overreaction to surprising signals. In addition, individuals who report lower confidence (i.e., a higher expected maximum prediction error) also exhibit significantly greater distortion in signal interpretation. This suggests that lack of confidence does not necessarily promote caution or accuracy; instead, it may reflect greater susceptibility to misleading or surprising information.

Table 4 examines the effects of surprising signals on learning inefficiency and belief po-

Table 4: Learning efficiency and polarization with random signals

VARIABLES	Dependent Variable: abs(Predict - Best guess)			Dependent Variable: sd(Predict)		
	(1)	(2)	(3)	(4)	(5)	(6)
Surprising signal	0.146	0.142	-0.548	0.643***	0.636***	-0.116
	(0.416)	(0.402)	(0.433)	(0.0825)	(0.0835)	(0.0884)
Age		0.132	0.140		0.0252	0.0338
		(0.205)	(0.205)		(0.0291)	(0.0292)
Male		1.048	1.025		0.0254	-0.000587
		(1.334)	(1.342)		(0.123)	(0.117)
Experiment experience		0.428	0.442		0.126	0.141
		(1.178)	(1.180)		(0.143)	(0.139)
Statistics knowledge		1.199	1.184		-0.0629	-0.0787
		(1.477)	(1.481)		(0.143)	(0.135)
Bayesian rationality		0.891	0.879		-0.191	-0.203
		(2.488)	(2.498)		(0.178)	(0.178)
Risk preference		0.180	0.177		-0.0744**	-0.0779**
		(0.266)	(0.268)		(0.0345)	(0.0345)
Lack of confidence		0.0495*	0.0501*		-0.00609	-0.00548
		(0.0291)	(0.0293)		(0.00479)	(0.00472)
Observations	2,376	2,376	2,376	2,376	2,376	2,376
R-squared	0.000	0.019	0.042	0.017	0.022	0.335
Period FEs	NO	NO	YES	NO	NO	YES

Notes: The first three columns report the results for learning efficiency, while the last three columns report the results for belief polarization. Columns (1) and (4) report the results without individual characteristics, columns (2) and (5) report the results controlling for several individual characteristics, and columns (3) and (6) report the results further controlling for period fixed effects. Observations are from the Baseline. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

larization in Baseline Treatment. The results show that surprising signals have limited effects on both learning inefficiency and polarization. Across Columns (1) to (3), the coefficients on surprising signals for learning inefficiency (measured by the absolute difference between predictions and best guesses) are small and statistically insignificant, suggesting that surprising signals do not systematically improve or impair prediction accuracy.

In terms of belief polarization (Columns 4 to 6), the effect of surprising signals appears large and positive when period fixed effects are not controlled (e.g., 0.643 in Column 4, significant at the 1% level). This is consistent with earlier findings from Figure 5, where polarization appears higher under surprising signals. However, since the signals are randomly assigned and thus not systematically aligned with participants' existing beliefs, this initial increase reflects temporary overreactions rather than persistent belief divergence.

Once period fixed effects are included (Column 6), the coefficient on surprising signals becomes smaller (–0.116) and statistically insignificant, indicating that the initial polarization effect is largely driven by period-specific variation. This suggests that over time, the effects of surprising signals average out, and participants do not become systematically more polarized under random information. In other words, while surprising signals may temporarily elevate polarization, they do not produce lasting divergence compared to confirming signals.

Among the control variables, risk preference exhibits a significant negative association with belief polarization in Columns (5) to (6), indicating that greater risk aversion is linked to lower belief divergence. This pattern suggests that risk-averse individuals tend to update their beliefs more cautiously in response to new information, potentially anchoring more strongly to prior beliefs. Consequently, groups composed of more risk-averse members are less likely to become polarized, possibly due to a general reluctance to overreact to ambiguous or extreme signals.

4.3 Biased Sources

Now we examine how participants update their beliefs when exposed to signals originating from biased sources by analyzing CU and IU treatments: in CU, participants receive signals from others who made congruent predictions in terms of the majority color of the urn in the first period; in IU, participants are shown signals from individuals who had predicted the incongruent outcome in terms of the majority color in the first period. As in Section 4.2, we first focus on belief updating, that is, how participants aggregate new information (incoming partial signals) with existing beliefs, after which we also examine how learning inefficiency and polarization evolve over time.

Figure 6 presents the patterns of signal interpretation under these different conditions. In all cases, participants interpret confirming signals (left Panels) — those aligned with their beliefs in the previous period — in a manner closely consistent with the actual data. This pattern holds regardless of whether the signals originate from biased sources. When participants unknowingly receive signals from biased sources (in CU and IU), that is, always from participants they made congruent or incongruent predictions in the first period, their interpretation of surprising signals — those that contradict their beliefs in the previous period — is significantly distorted. This behavior closely resembles the pattern observed under random signals (Panel Baseline),

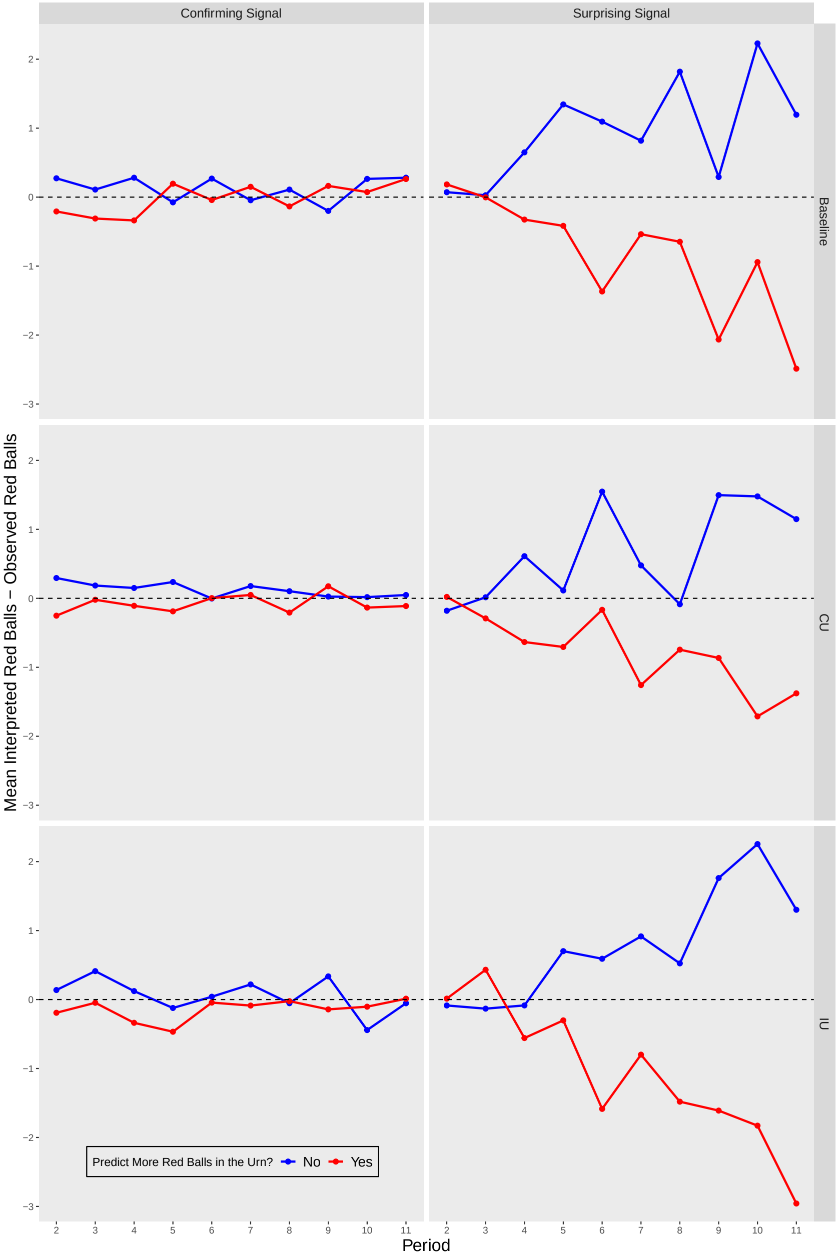


Figure 6: The impact of biased source on interpretation of observed balls

Notes: The figure shows the dynamics of the average gap between the number of red balls observed and the interpreted number (inferred from predictions in Periods $t-1$ and t) for confirming (left) and surprising (right) signals. Blue and red lines indicate the majority color in participants' predictions in Period $t-1$. The three rows correspond to the Baseline, CU and IU treatments, respectively.

indicating that when participants are unaware of the bias in the source, they treat the signals as if they were unbiased and process them accordingly.

Figure 7 illustrates how these patterns of interpretation subsequently affect belief formation — specifically, in terms of learning inefficiency and belief polarization. In the case of confirming signals (left Panels), as we have pointed out, participants interpreted the signal at its

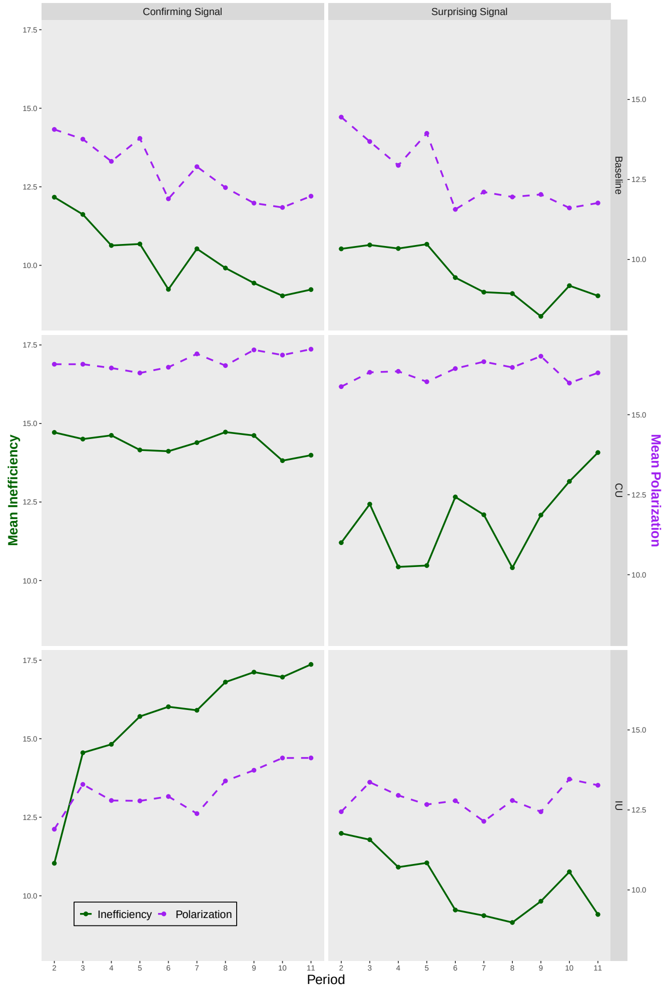


Figure 7: The impact of biased source on learning inefficiency and polarization

Notes: The figure shows the dynamics of learning inefficiency (green, solid lines) and polarization (purple, dashed lines) for confirming (left) and surprising (right) signals. The three rows correspond to the Baseline, CU and IU treatments, respectively.

face value, just like what they did in the Baseline Treatment. When information source is biased, the signals are drawn from systematically biased sources, thus treating them as unbiased leads to a steady accumulation of error. Over time, this learning inefficiency not only persists but even grows, especially when signals come from incongruent sources with opposing priors (Panel IU×Confirming). In such cases, participants unknowingly follow confirming signals that appear consistent with their beliefs but are actually misleading because the source is systematically biased. As this process repeats across periods, the cumulative distortion widens the gap between perceived and actual information quality, resulting in increased learning inefficiency ($p < 0.01$). A comparable pattern holds for belief polarization: although initial divergence may appear limited, individuals gradually drift further apart as they accumulate biased and confirming evidence.

Turning to surprising signals (right panels), the dynamics are more nuanced. When participants are unaware of the source's bias, they continue to interpret signals similarly to the Baseline treatment, as they overreact to surprising signals. However, such overreaction can sometimes offset the directional skew of the source because the underlying signal is biased. That is, the upward adjustment toward the signal may partially cancel out the bias embedded in it, especially early on. In these cases, occasional over-interpretation can temporarily reduce accumulated error. However, this incidental correction is not systematic, and the overall learning inefficiency remains high relative to the Baseline treatment (CU v.s. Baseline $p < 0.01$, IU v.s. Baseline $p = 0.012$).

Comparing between the two biased sources, it is worth noting that signals originating from congruent sources tend to generate greater polarization compared to those from incongruent sources ($p < 0.01$). This is because congruent sources are more likely to generate signals that appear to confirm their prior beliefs, as new information comes from like-minded individuals, and thus are more likely to reinforce existing views. This reinforcement limits belief movement across groups and contributes to growing divergence over time. As shown in Figure 7, belief polarization stays high under CU treatments. In contrast, exposure to incongruent sources tends to induce more cautious or balanced responses, resulting in comparatively lower levels of polarization. These findings suggest that not only the presence of bias, but also the direction of source alignment, plays a critical role in shaping the collective trajectory of belief updating.

In summary, consistent with Hypotheses 2 and 3, unrecognized source bias can systematically distort belief formation, as participants treat biased signals as unbiased. These distortions in belief updating, manifesting as learning inefficiency and belief polarization, tend to accumulate and intensify over time. Below we summarize these results.

Result 2. *When unaware of source bias, participants update their beliefs as if the source were unbiased. As a result, biased sources systematically distort belief formation, increasing both learning inefficiency and polarization.*

To rigorously quantify how biased sources influence belief updating, we estimate the following regression specification:

$$Y_{ist} = \beta_0 + \beta_1 \text{ Surprising}_{ist} + \beta_2 \text{ Biased}_s + \beta_3 \text{ Surprising}_{ist} \times \text{Biased}_s + \theta X_i + \lambda_t + \varepsilon_{ist} \quad (5)$$

where Biased_s indicates whether session s features signal sources with systematic directional bias; all other variables are defined as above. The key coefficient of interest is β_3, which captures whether—and to what extent—biased sources alter participants' responses to surprising signals.

Table 5 reports results from estimating Equation (5), focusing on how individuals interpret signals from biased sources. The dependent variable is the deviation between the participant's interpreted number of red balls and the actual number observed. Regressions are split by participants' prior beliefs: columns (1)–(3) report results for those believing a red-majority, and columns (4)–(6) for those believing a blue-majority.

Table 5: Interpretation of observed balls with biased signals

	Dependent Variable: Interpreted red balls - Observed red balls					
VARIABLES	Predict more red balls in the previous period			Predict more blue balls in the previous period		
	(1)	(2)	(3)	(4)	(5)	(6)
Surprising signal	-0.818***	-0.825***	-0.817***	0.893***	0.907***	0.909***
	(0.275)	(0.274)	(0.273)	(0.278)	(0.276)	(0.279)
Signal from congruent source	-0.102		-0.0500	0.0381		0.0224
	(0.113)		(0.104)	(0.0841)		(0.0809)
Surprising signal × Signal from congruent source	0.140		0.150	-0.342		-0.362
	(0.389)		(0.387)	(0.405)		(0.402)
Signal from incongruent source		-0.0372	-0.0620		-0.110	-0.0886
		(0.116)	(0.109)		(0.100)	(0.0932)
Surprising signal × Signal from incongruent source		-0.0554	-0.0374		-0.242	-0.266
		(0.367)	(0.369)		(0.341)	(0.346)
Observations	1,841	1,951	2,873	2,449	2,369	3,577
R-squared	0.083	0.081	0.072	0.085	0.081	0.072
Period FEs	YES	YES	YES	YES	YES	YES

Notes: The first three columns report the results when participants predict more red balls in the previous period, while the last three columns report the results when participants predict more blue balls in the previous period. Columns (1) and (4) report the results using observations from the Baseline and CU treatments, columns (2) and (5) report the results using observations from the Baseline and IU treatments, and columns (3) and (6) report the results using observations from the Baseline, CU and IU treatments. Individual characteristics and period fixed effects are controlled for all columns. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Across all specifications, surprising signals exert a large and statistically significant influence on belief updating. For participants who initially predicted a red majority, the coefficients are negative (approximately –0.82), indicating an overreaction to signals suggesting fewer red balls than expected. For those predicting a blue majority, the coefficients are positive (around 0.90), reflecting a similar overreaction to signals indicating more red balls. This consistent asymmetry reveals that participants assign excessive weight to surprising signals rather than discounting them. Importantly, this pattern persists even when signals originate from biased sources, suggesting that participants continue to interpret such information as if it were unbiased, closely resembling the behavior observed under random-signal conditions.

The remaining coefficients examine whether this overreaction is moderated by alignment between the participant's prior belief and that of the signal source. The main effects of source alignment — whether the signal comes from someone with similar or opposing priors — are generally small and statistically insignificant. Likewise, the interaction terms between surprising signals and source alignment lack consistent significance, indicating that participants overreact to disconfirming information regardless of the source's identity. In other words, when the identity of the source is not disclosed, individuals fail to adjust for potential bias, and belief updating remains systematically distorted.

Table 6: Learning efficiency with biased signals

VARIABLES	Dependent Variable: abs(Predict - Best guess)					
	(1)	(2)	(3)	(4)	(5)	(6)
Surprising signal				-0.462	-0.537	-0.469
				(0.429)	(0.429)	(0.426)
Signal from congruent source	3.420***		3.319***	4.113***		4.022***
	(0.800)		(0.793)	(0.846)		(0.836)
Surprising signal × Signal from congruent source				-2.137***		-2.186***
				(0.628)		(0.629)
Signal from incongruent source		3.332***	3.441***		5.785***	5.860***
		(0.738)	(0.728)		(0.982)	(0.977)
Surprising signal × Signal from incongruent source					-5.043***	-4.975***
					(0.915)	(0.909)
Observations	4,719	4,752	7,095	4,719	4,752	7,095
R-squared	0.064	0.042	0.044	0.072	0.077	0.071
Period FEs	YES	YES	YES	YES	YES	YES

Notes: Columns (1) and (4) report the results using observations from the Baseline and CU treatments, columns (2) and (5) report the results using observations from the Baseline and IU treatments, and columns (3) and (6) report the results using observations from the Baseline, CU and IU treatments. Individual characteristics and period fixed effects are controlled for all columns. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

The subsequent consequences of this interpretive asymmetry are evident in Tables 6 and

7, which explore impacts on learning accuracy and belief polarization, respectively. Table 6 examines learning inefficiency: signals from biased sources, whether aligned or opposed, consistently reduce learning accuracy, as reflected in the positive and significant coefficients on both Signal from congruent source and Signal from incongruent source. This indicates that exposure to biased information reduces predictive precision regardless of source alignment.

However, the interaction terms with *Surprising signal* introduce an important nuance. While biased signals generally impair learning, the exaggerated responses to disconfirming signals can occasionally offset some of this distortion. The significantly negative interaction coefficients (-2.14 for congruent source, -5.04 for incongruent source) suggest that such overreactions may, at times, bring participants' beliefs closer to the benchmark prediction. These corrections are not the result of accurate inference, but rather incidental byproducts of excessive responsiveness, which nonetheless illustrate moments when interpretive asymmetry unexpectedly improves learning accuracy.

Table 7: Polarization with biased signals

VARIABLES	Dependent Variable: sd(Predict)					
	(1)	(2)	(3)	(4)	(5)	(6)
Surprising signal				0.119	-0.157*	0.0126
				(0.0881)	(0.0836)	(0.0861)
Signal from congruent source	3.689***		3.679***	3.995***		3.985***
	(0.122)		(0.122)	(0.126)		(0.127)
Surprising signal × Signal from congruent source				-0.823***		-0.867***
				(0.115)		(0.115)
Signal from incongruent source		0.446***	0.444***		0.649***	0.623***
		(0.108)	(0.112)		(0.120)	(0.120)
Surprising signal × Signal from incongruent source					-0.416***	-0.369***
					(0.113)	(0.110)
Observations	4,719	4,752	7,095	4,719	4,752	7,095
R-squared	0.506	0.271	0.473	0.513	0.278	0.481
Period FEs	YES	YES	YES	YES	YES	YES

Notes: Columns (1) and (4) report the results using observations from the Baseline and CU treatments, columns (2) and (5) report the results using observations from the Baseline and IU treatments, and columns (3) and (6) report the results using observations from the Baseline, CU and IU treatments. Individual characteristics and period fixed effects are controlled for all columns. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table 7 shifts the focus to belief polarization. As anticipated, biased signals significantly increase polarization: both *Signal from congruent source* and *Signal from incongruent source* are linked to greater belief divergence. Once again, the interaction terms with *Surprising signal* are negative and statistically significant (-0.87 for congruent source, -0.37 for incongruent source), suggesting that overreaction to surprising information can partially offset this trend. When participants respond too strongly to unexpected signals, these shifts can occasionally reduce opinion dispersion, though only temporarily. Nevertheless, the overall level of polarization remains high, underscoring the limits of this moderating effect.

4.4 Awareness of Biased Sources

Next, we delve into the impact of awareness of biased source. Specifically, we are interested in whether and to what extent participants update their beliefs when they become aware that partial signals are not purely random: they either always come from those who share congruent first-period predictions or hold incongruent first-period predictions.

Figure 8 presents the patterns of signal interpretation under these different conditions. When new signals confirm predictions in the last period (left panels), the interpretation resembles that in other treatments (including both from unbiased source, and from biased source but participants are unaware of it) and is close to actual data observed. This is not the case when new signals contradicts predictions in the last period (right panels). When participants observe surprising signals and are informed that signals are from biased source, they tend to discount the informativeness of the signal and interpret it more cautiously. Overreaction to surprising signal still remain, but is mitigated significantly ($p < 0.01$).

Likewise, Figure 9 illustrates how such interpretive patterns translate into subsequent effects on belief formation in terms of learning inefficiency and belief polarization. Once participants are informed that the information source is biased, in particular when the signal comes from someone with incongruent initial belief, they begin to interpret surprising signals more cautiously. It is worth noting that this adjustment brings beliefs closer to the biased signal rather than the truth. When individuals reduce overreaction to surprising signals, they move less in the right direction, allowing the source's bias to exert a stronger cumulative influence. As a result, awareness to biased information sources increases rather than reduces learning inefficiency (CA vs CU $p = 0.026$). This dynamic reveals a paradox: awareness of bias tempers overreactions but does not guarantee better learning outcomes, as it can also eliminate the un-

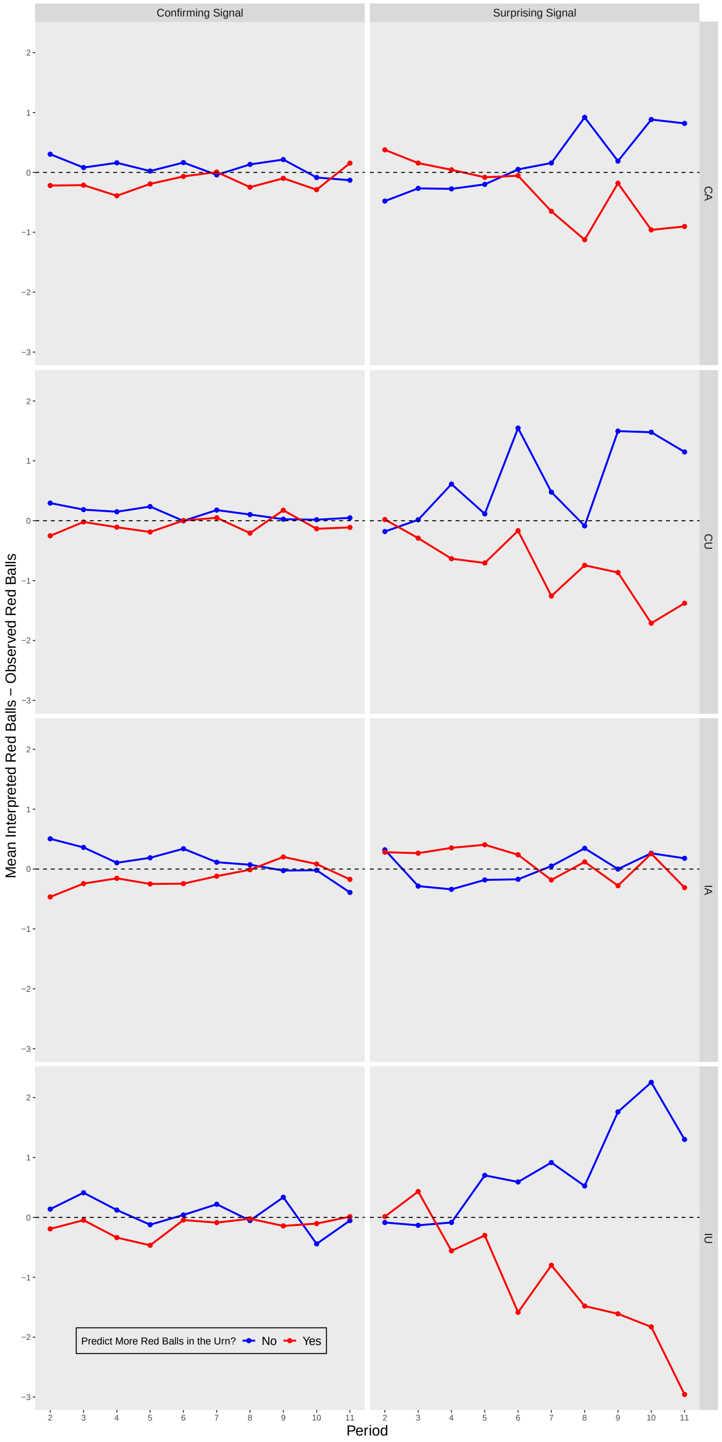


Figure 8: The impact of awareness of biased source on interpretation of observed balls

Notes: The figure shows the dynamics of the average gap between the number of red balls observed and the interpreted number (inferred from predictions in Periods $t-1$ and t) for confirming (left) and surprising (right) signals. Blue and red lines indicate the majority color in participants' predictions in Period $t-1$. The four rows correspond to the CA, CU, IA and IU treatments, respectively.

intended belief-correction benefits that overreactions sometimes produce. At the same time, awareness even intensifies polarization (CU v.s. CA $p < 0.01$, IU vs IA $p = 0.035$), as reduced

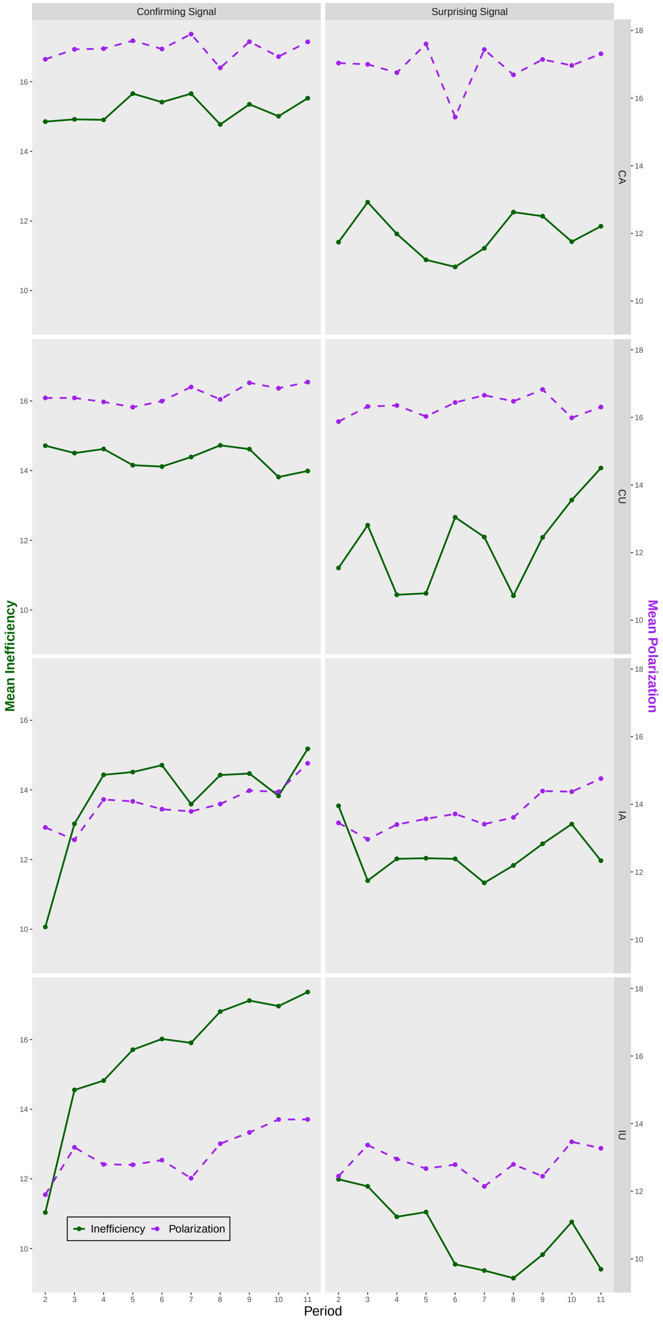


Figure 9: The impact of awareness of biased source on learning inefficiency and polarization

Notes: The figure shows the dynamics of learning inefficiency (green, solid lines) and polarization (purple, dashed lines) for confirming (left) and surprising (right) signals. The four rows correspond to the CA, CU, IA and IU treatments, respectively.

responsiveness to contradicting evidence reinforces initial belief differences.

It is worth noting that both unrecognized and recognized source bias can systematically distort belief formation, though through different mechanisms. When source bias is concealed, participants are misled by treating biased signals as unbiased. When source bias is disclosed, individuals become more cautious and exhibit reduced overreaction to surprising signals. However, this attenuation does not always lead to more accurate belief formation, as overreaction can, in some cases, partially offset the distortions introduced by biased information sources.

Below we summarize these results, which are not consistent with Hypothesis 5. A plausible explanation for this discrepancy lies in the cognitive difficulty of translating the knowledge of a bias into a concrete adjustment. While identifying a source as biased may be straightforward, the subsequent mental step — reverse-engineering the true signal — is inherently complex. This complexity may prevent individuals from properly incorporating the information, leading to the observed outcomes.

Result 3. *When participants are informed of source bias, then temper their overreaction to surprising signals and interpret such signals more at face value — particularly when signals come from belief-incongruent sources. While this adjustment intensifies belief polarization, its average effect on learning efficiency is negligible.*

To further investigate whether awareness of source identity alters the impact of surprising signals, we estimate the following specification:

$$Y_{ist} = \gamma_0 + \gamma_1 \text{ Surprising}_{ist} + \gamma_2 \text{ Know}_s + \gamma_3 \text{ Surprising}_{ist} \times \text{Know}_s + \theta X_i + \lambda_t + \varepsilon_{ist} \quad (6)$$

where Know_s indicates whether participants in session s are informed about the group identity (same or different) of the participant from whom the signal was drawn; all other variables are defined as above. The main coefficient of interest is γ_3, which captures whether this awareness moderates the interpretation of surprising signals originating from biased sources.

Table 8 presents results from estimating Equation (6). Consistent with earlier findings, surprising signals exert strong effects on belief updating across all specifications, reflecting a persistent pattern of overreaction. The interaction terms in columns (2) and (4) are statistically significant and opposite in sign to the main surprising signal effect. The coefficient is 1.029 for red-to-blue updating and -0.651 for blue-to-red, indicating that participants revise their beliefs less aggressively when they know the signal comes from someone with different priors. In contrast, no such moderating effect appears when the signal is linked to someone with similar priors: the interaction terms in columns (1) and (3) are small and not statistically significant. These results point to a conditional correction pattern: source awareness tempers overreaction, but only when the information is known to come from a participant with opposing beliefs.

Table 8: Interpretation of observed balls when knowing signal sources

VARIABLES	Dependent Variable: Interpreted red balls - Observed red balls			
	Predict more red balls in the previous period		Predict more blue balls in the previous period	
	Signal from congruent source	Signal from incongruent source	Signal from congruent source	Signal from incongruent source
	(1)	(2)	(3)	(4)
Surprising signal	-0.681**	-0.831***	0.544*	0.610***
	(0.277)	(0.260)	(0.296)	(0.216)
Know signal source	-0.0542	-0.0358	-0.0641	0.0315
	(0.115)	(0.115)	(0.0658)	(0.123)
Surprising signal	0.473	1.029***	-0.446	-0.651**
× Know signal source	(0.383)	(0.309)	(0.349)	(0.264)
Observations	1,828	2,184	2,462	2,136
R-squared	0.043	0.060	0.026	0.045
Period FEs	YES	YES	YES	YES

Notes: The first two columns report results when participants predict more red balls in the previous period, while the last two columns report the results when participants predict more blue balls in the previous period. Columns (1) and (3) report the results using observations from the CU and CA treatments, and columns (2) and (4) report the results using observations from the IU and IA treatments. Individual characteristics and period fixed effects are controlled for all columns. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

The coefficients of the interaction terms in Table 8, combined with the regression estimates in Table 5, illuminate the interplay between cognitive bias and source bias in the belief updating process. Table 5 shows that when signals are biased but the bias is not disclosed to individuals, belief updating is driven primarily by whether the signals confirm or contradict prior beliefs, regardless of the direction of the bias. In other words, individuals respond to signals based on their congruence with prior beliefs, without accounting for potential source bias or its interaction with cognitive tendencies. However, when individuals are made aware of the direction of the biased information sources, the influence of source bias, particularly its interaction with cognitive bias, becomes evident as shown in Table 8. Specifically, when signals are known to originate from sources holding opposing beliefs, the tendency to overreact to surprising signals is mitigated. This debiasing effect does not occur when signals come from sources with consistent beliefs, as summarized in Result 3.

The subsequent effects of source awareness on belief accuracy and polarization are ex-

Table 9: Learning efficiency when knowing signal sources

VARIABLES	Dependent Variable: abs(Predict - Best guess)			
	Signal from congruent source		Signal from incongruent source	
	(1)	(2)	(3)	(4)
Surprising signal		-2.499***		-5.072***
		(0.515)		(0.803)
Know signal source	0.826	1.156	-0.673	-1.957*
	(0.877)	(0.977)	(0.859)	(1.181)
Surprising signal × Know signal source		-1.112		2.923***
		(0.697)		(1.008)
Observations	4,719	4,719	4,752	4,752
R-squared	0.016	0.034	0.020	0.048
Period FEs	YES	YES	YES	YES

Notes: Columns (1) and (2) report the results using observations from the CU and CA treatments, and columns (3) and (4) report the results using observations from the IU and IA treatments. Individual characteristics and period fixed effects are controlled for all columns. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

plored in Tables 9 and 10. Table 9 shows that surprising signals consistently enhance learning efficiency by reducing prediction error. This effect holds for both congruent and incongruent sources, a larger reduction observed when the surprising signal originates from incongruent source. Source awareness further contributes to this improvement when signals come from incongruent source, indicating that knowing the signal's direction helps participants better correct for potential bias. However, this moderating effect is weaker for surprising signals. The interaction term shows that while awareness still reduces learning inefficiency, it significantly dampens the corrective force of surprising signals, cutting the effect by more than half. In contrast, when signals come from congruent source, source awareness has no meaningful impact: both the coefficients of *Know signal source* and that of the interaction term are small and statistically insignificant, suggesting that participants treat information from belief-congruent sources similarly, regardless of whether its origin is disclosed.

Table 10 reveals that surprising signals reduce belief polarization, although this effect is relatively small and insignificant when the signals originate from incongruent source. However, awareness of source identity consistently increases belief polarization, with main effects ranging from 0.668 to 0.857. The significant negative interaction (-0.406 and -0.309) indicates that, while awareness of source identity generally increases polarization, this effect is attenuated when the information is surprising. Although the net effect remains polarization-increasing, the attenuation is both substantial and meaningful—approximately 36% to 61%.[15]

Table 10: Polarization when knowing signal sources

VARIABLES	Dependent Variable: sd(Predict)			
	Signal from congruent source		Signal from incongruent source	
	(1)	(2)	(3)	(4)
Surprising signal		-0.245***		-0.109
		(0.0925)		(0.0795)
Know signal source	0.668***	0.799***	0.681***	0.857***
	(0.138)	(0.149)	(0.220)	(0.233)
Surprising signal × Know signal source		-0.406***		-0.309**
		(0.120)		(0.145)
Observations	4,719	4,719	4,752	4,752
R-squared	0.069	0.082	0.216	0.219
Period FEs	YES	YES	YES	YES

Notes: Columns (1) and (2) report the results using observations from the CU and CA treatments, and columns (3) and (4) report the results using observations from the IU and IA treatments. Individual characteristics and period fixed effects are controlled for all columns. Standard errors clustered at the participant level are reported in parentheses.
*p<0.1; **p<0.05; ***p<0.01

To sum up, our estimates highlight several key findings. First, participants generally overreact to surprising signals, demonstrating a surprise-driven updating bias rather than a conformation bias. Second, when the source is biased but participants are unaware of this, they treat these signals as if they were randomly drawn. As a result, biased sources typically increase learning inefficiency and polarization. Interestingly, these negative effects are partially alleviated when participants overreact to surprising signals, which can inadvertently counterbalance the impact of source bias. Third, when the source of the signals is made aware, participants tend to react less strongly to surprising signals. While source awareness can reduce learning inefficiency particularly when signals are from belief-incongruent source, it also increases belief polarization for both congruent and incongruent sources. Again, this polarization-increasing effect is moderated in the presence of surprising signals.

[15] These values are calculated as the ratios of the interaction term coefficient to that of the main effect of knowing the signal source.

5 Concluding Remarks

Rapid technological advancements and widespread reliance on digital platforms for information consumption have reshaped how people form beliefs and opinions. In particular, algorithmic recommendation systems are often devised to deliver personalized content for higher view duration, which could potentially exacerbates political and social polarization by creating filter bubbles. At the same time, cognitive biases intrinsic to human significantly influence how individuals interpret and process new information. Understanding how these algorithmic and cognitive biases interact and jointly shape belief updating processes is crucial in addressing fundamental issues of both belief formation and polarization.

Our study employs a controlled laboratory experiment to systematically examine belief updating when individuals encounter biased signals. We identify a persistent behavioral pattern wherein individuals exhibit a surprise-driven learning bias, characterized by overreaction to signals that challenge their prior beliefs, rather than traditional confirmation bias. When subjects unknowingly receive information from biased sources, their learning inefficiency and polarization notably increase. Although informing subjects about source bias reduces their overreaction to surprising signals, this awareness alone intensifies polarization without significantly improving learning efficiency, indicating that individuals struggle to optimally adjust their beliefs despite increased skepticism.

Our results advance the understanding of belief updating biases by providing clear evidence of surprise-driven rather than confirmatory bias in a social learning context. We also highlight the complex interactions between cognitive biases and algorithmically biased information sources, revealing the limitations of simple awareness as an effective corrective strategy. From a practical standpoint, our findings offer valuable insights for policymakers and platform designers seeking to curb polarization and misinformation in the digital era. The controlled laboratory environment and the abstract experimental setting employed in this study enable precise manipulation of treatment conditions and accurate measurement of key variables, such as signals and beliefs. This design also allows us to cleanly identify causal effects while minimizing the influence of motivated reasoning (see Charness et al., 2021). Although our urn-inference task is relatively abstract compared to the complexity and diversity of real-world digital information consumption, it effectively captures key aspects of how individuals process abundant, rapidly evolving, and relatively impersonal information streams of digital content. Consequently, our results provide valuable insights into the cognitive processes underlying belief updating, the role of source bias, and the potential effectiveness of interventions aimed at mitigating biased information processing. Looking ahead, we believe future research can complement our findings with empirical studies conducted in real-world digital environments, and explore additional interventions beyond awareness, such as training individuals in bias-detecting and debiasing techniques, or introducing social feedback mechanisms that encourage open communication and deliberation, thereby promoting more balanced information processing.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2025. "AI and Social Media: A Political Economy Perspective."

Agranov, Marina, Gabriel Lopez-Moctezuma, Philipp Strack, and Omer Tamuz. 2022. "Learning through imitation: an experiment."Technical report, National Bureau of Economic Research.

Allcott, Hunt, Levi Boxell, Jacob Conway, Matthew Gentzkow, Michael Thaler, and David Yang. 2020. "Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic." *Journal of public economics* 191 104254.

Allcott, Hunt, and Matthew Gentzkow. 2017. "Social media and fake news in the 2016 election." *Journal of economic perspectives* 31 (2): 211–236.

Anderson, Lisa R, and Charles A Holt. 1997. "Information cascades in the laboratory." *The American economic review* 847–862.

Axelrod, Robert, Joshua J Daymude, and Stephanie Forrest. 2021. "Preventing extreme polarization of political attitudes." *Proceedings of the National Academy of Sciences* 118 (50): e2102139118.

Bail, C. A., L. P. Argyle, T. W. Brown et al. 2018. "Exposure to opposing views on social media can increase political polarization." *Proc Natl Acad Sci U S A* 115 (37): 9216–9221. 10.1073/pnas.1804840115.

Bakshy, E., S. Messing, and L. A. Adamic. 2015. "Political science. Exposure to ideologically diverse news and opinion on Facebook." *Science* 348 (6239): 1130–2. 10.1126/science.aaa1160.

Banerjee, Abhijit V. 1992. "A simple model of herd behavior." *The quarterly journal of economics* 107 (3): 797–817.

Bikhchandani, Sushil, David Hirshleifer, Omer Tamuz, and Ivo Welch. 2024. "Information cascades and social learning." *Journal of Economic Literature* 62 (3): 1040–1093.

Bikhchandani, Sushil, David Hirshleifer, and Ivo Welch. 1992. "A theory of fads, fashion, custom, and cultural change as informational cascades." *Journal of political Economy* 100 (5): 992–1026.

Bowen, T Renee, Danil Dmitriev, and Simone Galperti. 2023. "Learning from shared news: When abundant information leads to belief polarization." *The Quarterly Journal of Economics* 138 (2): 955–1000.

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

Charness, Gary, Ryan Oprea, and Sevgi Yuksel. 2021. "How do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment." *Journal of the European Economic Association* 19 (3): 1656–1691. 10.1093/jeea/jvaa051.

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.

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

DellaVigna, Stefano, and Ethan Kaplan. 2007. “The Fox News effect: Media bias and voting.” *The Quarterly Journal of Economics* 122 (3): 1187–1234.

Enke, Benjamin, and Thomas Graeber. 2023. “Cognitive uncertainty.” *The Quarterly Journal of Economics* 138 (4): 2021–2067.

Enke, Benjamin, Thomas Graeber, Ryan Oprea, and Jeffrey Yang. 2024. “Behavioral attenuation.”Technical report, National Bureau of Economic Research.

Faia, Ester, Andreas Fuster, Vincenzo Pezone, and Basit Zafar. 2024. “Biases in information selection and processing: Survey evidence from the pandemic.” *Review of Economics and Statistics* 106 (3): 829–847.

Fiedler, Klaus, and Peter Juslin. 2006. *Information sampling and adaptive cognition*. Cambridge University Press.

Flaxman, Seth, Sharad Goel, and Justin M Rao. 2016. “Filter bubbles, echo chambers, and online news consumption.” *Public opinion quarterly* 80 (S1): 298–320.

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.

Gentzkow, Matthew, and Jesse M Shapiro. 2010. “What drives media slant? Evidence from US daily newspapers.” *Econometrica* 78 (1): 35–71.

Grimm, Veronika, and Friederike Mengel. 2020. “Experiments on belief formation in networks.” *Journal of the European Economic Association* 18 (1): 49–82.

Groseclose, Tim, and Jeffrey Milyo. 2005. “A measure of media bias.” *The quarterly journal of economics* 120 (4): 1191–1237.

Guess, Andrew M, Pablo Barberá, Simon Munzert, and JungHwan Yang. 2021. “The consequences of online partisan media.” *Proceedings of the National Academy of Sciences* 118 (14): e2013464118.

Guilbeault, Douglas, Samuel Woolley, and Joshua Becker. 2021. “Probabilistic social learning improves the public’s judgments of news veracity.” *PLoS One* 16 (3): e0247487.

Holt, Charles A, and Susan K Laury. 2002. “Risk aversion and incentive effects.” *American economic review* 92 (5): 1644–1655.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. 2021. "Examining the consumption of radical content on YouTube." *Proceedings of the National Academy of Sciences* 118 (32): e2101967118.

Hosseinmardi, Homa, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J Watts. 2024. "Causally estimating the effect of YouTube's recommender system using counterfactual bots." *Proceedings of the national academy of sciences* 121 (8): e2313377121.

Juslin, Peter, Anders Winman, and Patrik Hansson. 2007. "The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals.." *Psychological review* 114 (3): 678.

Kieren, Pascal, Jan Müller-Dethard, and Martin Weber. 2020. "Disconfirming Information and Overreaction in Expectations." In *Proceedings of Paris December 2021 Finance Meeting EUROFIDAI-ESSEC.*

Kuhnen, Camelia M. 2015. "Asymmetric learning from financial information." *The Journal of Finance* 70 (5): 2029–2062.

Levendusky, Matthew S. 2013. "Why do partisan media polarize viewers?" *American journal of political science* 57 (3): 611–623.

Levy, Gilat, and Ronny Razin. 2019. "Echo chambers and their effects on economic and political outcomes." *Annual Review of Economics* 11 (1): 303–328.

Levy, Ro'ee. 2021. "Social media, news consumption, and polarization: Evidence from a field experiment." *American economic review* 111 (3): 831–870.

Martin, Gregory J, and Ali Yurukoglu. 2017. "Bias in cable news: Persuasion and polarization." *American Economic Review* 107 (9): 2565–2599.

Mosleh, Mohsen, Cameron Martel, Dean Eckles, and David G Rand. 2021a. "Shared partisanship dramatically increases social tie formation in a Twitter field experiment." *Proceedings of the National Academy of Sciences* 118 (7): e2022761118.

Mosleh, Mohsen, Gordon Pennycook, Antonio A Arechar, and David G Rand. 2021b. "Cognitive reflection correlates with behavior on Twitter." *Nature communications* 12 (1): 921.

Mullainathan, Sendhil. 2002. "A Memory-Based Model of Bounded Rationality." *Quarterly Journal of Economics* 117 (3): 735–774.

Mullainathan, Sendhil, and Andrei Shleifer. 2005. "The market for news." *American economic review* 95 (4): 1031–1053.

Nickerson, Raymond S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of general psychology* 2 (2): 175–220.

Pariser, Eli. 2011. *The filter bubble: What the Internet is hiding from you.* penguin UK.

Qian, Kun, and Sanjay Jain. 2024. "Digital content creation: An analysis of the impact of recommendation systems." *Management Science*.

Santos, Fernando P, Yphtach Lelkes, and Simon A Levin. 2021. "Link recommendation algorithms and dynamics of polarization in online social networks." *Proceedings of the National Academy of Sciences* 118 (50): e2102141118.

Sunstein, Cass. 2018. *Republic: Divided democracy in the age of social media.* Princeton university press.

Wason, P. C. 1960. "On the Failure to Eliminate Hypotheses in a Conceptual Task." *Quarterly Journal of Experimental Psychology* 12 129–140. 10.1080/17470216008416717.

Weizsäcker, Georg. 2010. "Do we follow others when we should? A simple test of rational expectations." *American Economic Review* 100 (5): 2340–2360.

Williams, Cole. 2024. "Echo chambers: Social learning under unobserved heterogeneity." *The Economic Journal* 134 (658): 837–855.

Zou, Wenbo, and Xue Xu. 2023. "Ingroup bias in a social learning experiment." *Experimental Economics* 26 (1): 27–54.

Supplementary Materials

A Robustness

To further examine the robustness of surprising-driven learning, Figure S1 explores whether the magnitude and direction of signal distortion vary across different urn compositions. Specifically, we compare outcomes across urns with 36, 44, and 55 red balls to assess two potential concerns. First, we test whether the strength of the bias differs between more extreme versus more balanced distributions by comparing the 36-red and 44-red urns. Second, we evaluate whether our measure—based on participants' reports of the number of red balls—introduces asymmetries depending on whether the true state is red- or blue-dominant, by comparing the 44-red and 55-red urns.

Across all panels, we plot the difference between the number of red balls participants interpret from each signal and the actual number observed, disaggregated by prior belief (i.e., whether participants predicted more red or blue balls in the previous period). In the left-hand panels, where signals confirm prior beliefs, both red- and blue-leaning participants interpret the signals with minimal distortion. In contrast, the right-hand panels reveal systematic overreaction under surprising signals: participants tend to overinterpret signals that contradict their prior belief. Both patterns hold symmetrically for both red- and blue-leaning individuals and persists across all urn types. Overall, these results reinforce our earlier findings: interpretation bias under surprising signals is robust across different distributions and does not appear to be driven by framing or reporting asymmetries. In addition, both learning inefficiency and belief polarization exhibit similar patterns across different urn compositions, reinforcing the robustness of our findings across signal environments.[1]

[1] The figures showing learning inefficiency and polarization across different urn compositions are available upon request.

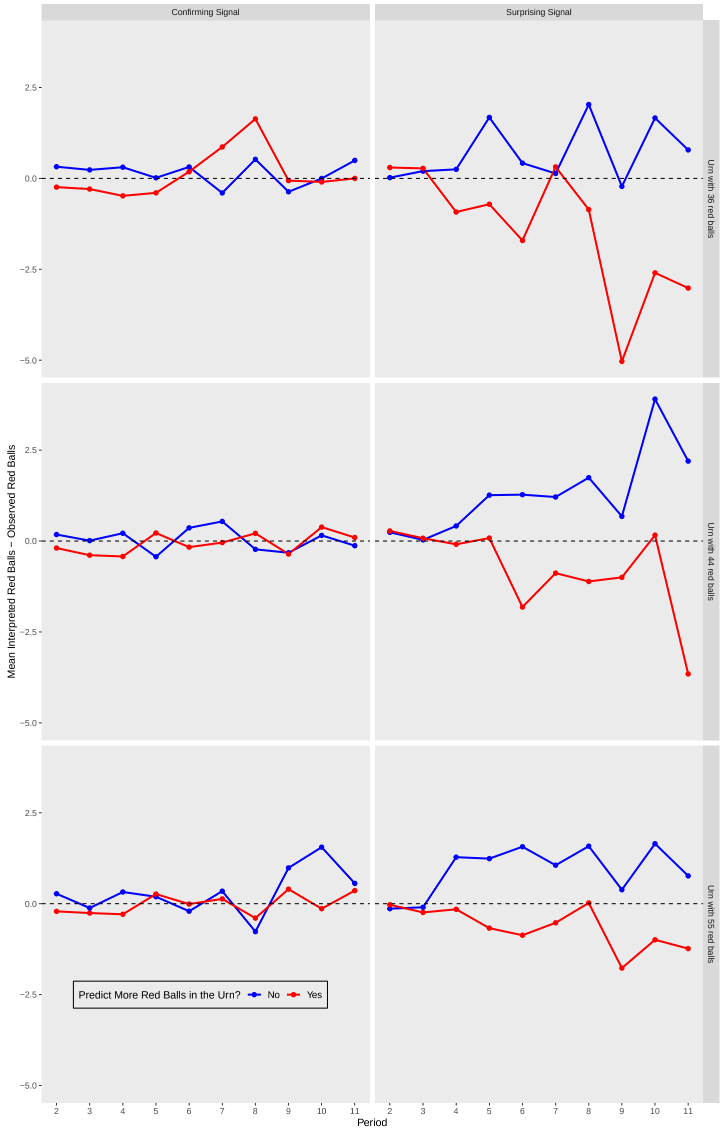


Figure S1: Interpretation of observed balls under different urn compositions

Notes: The figure shows the dynamics of the average gap between the number of red balls observed and the interpreted number (inferred from predictions in Periods $t-1$ and t). The left subplot corresponds to confirming signals (aligned with predictions in $t-1$); the right subplot to surprising signals (contradicting predictions in $t-1$). Blue and red lines indicate the majority color in participants' predictions in Period $t-1$. The three rows correspond to urns containing 36, 44, and 55 red balls, respectively.

B Instructions for the Experiment (Baseline treatment, translated into English)

You are currently participating in a decision-making experiment. Today's experiment consists of two parts, in each of which you will need to make some decisions. Your earnings depend on your decisions (and luck), so it is very important to read these instructions carefully.

During the experiment, you must not communicate with other participants in any form, nor can you use devices like mobile phones. To protect your privacy, all results of the experiment will be processed anonymously. If you have any questions, please raise your hand, and the experimenter will come to you to assist. If you violate any of the above rules, your decisions will not be analyzed, and you will receive no earnings.

Your earnings from the first part of the experiment will be calculated in points, which will be converted at the end of the experiment at a rate of **10 points = 1 RMB**. Earnings from the second part will be paid directly in RMB (Yuan). In addition to your earnings from the experiment decisions, you will receive a participation fee of 10 RMB. At the end of the experiment, your total earnings will be transferred to you via WeChat or Alipay.

Instructions for Part 1 of the Experiment

This part consists of three rounds, each following the same procedure detailed below.

In each round, there is a box containing red and blue balls, totaling 99 balls, but the exact number of each color is unknown. In each round, all participants face exactly the same box. Each round includes 11 stages, and during these stages, the number of red and blue balls remains unchanged. The specific procedure for each stage is as follows:

Stage 1: The computer randomly draws 5 balls with replacement for each participant from the box of 99 balls. After each ball is drawn, it is returned to the box. The draws for different participants are independent; thus, the proportion of red and blue balls observed by different participants may be the same or different. For example, Participant A may randomly get 3 red balls and 2 blue balls, Participant B may also randomly get 3 red and 2 blue balls, Participant C might randomly get 1 red and 4 blue balls, and so on. After seeing the 5 randomly drawn balls, **you must predict how many red balls (0-99) you think are in the box.**[2]

Stages 2-11: In each of these stages, the computer will **randomly**[3] select another participant from the remaining participants **(excluding yourself)**.[4] Then, it will **randomly select 3 balls**

[2] In Treatments CA and IA, the following sentence appears here: "Then, based on each participant's **prediction of the number of red balls in Stage 1**, the computer will divide all predictions into two intervals: one interval where the predicted number of red balls is less than 50 (i.e., the number of blue balls is 50 or more), and another interval where the predicted number of red balls is 50 or more."

[3] In Treatments CU and IU, where the information source is non-random but unknown to participants, we omit the word "randomly" in this sentence.

[4] In Treatment IA, this sentence is replaced by the following: "In each of these stages, the computer will randomly select, for each participant, one person from all other participants whose Stage 1 predictions **fall into**

without replacement from the 5 balls initially observed by that participant in Stage 1 and display these balls to you. After observing these 3 balls, **you must again predict how many red balls (0-99) you think are in the box.**

Note that the participant you are matched with in each of Stages 2-11 is randomly chosen, and may or may not be the same person each stage. You **will not know** who this participant is. Even if you are matched with the same participant across different stages, the 3 balls you observe may differ since they are randomly drawn from the original 5 balls that participant observed in Stage 1. Your screen will display the 3 balls observed in the current stage, as well as the balls observed in all previous stages (including Stage 1), and your predictions for each stage. At **the end of Stage 11**, you must answer the following question carefully: **what do you think is the maximum possible difference between your Stage 11 prediction and the actual number of red balls in the box?** (Answering this question will earn you an additional 30 points.)

Earnings Calculation: At the end of the experiment, the computer will **randomly select one stage** from each round (Stage 1-11) for payout calculation. Your earnings (in points) for each round are calculated as follows:

$$150 - 0.015 \times (\text{Actual number of red balls} - \text{Your predicted number of red balls})^2$$

Important notes:

1. The three boxes used in the three rounds (Box 1, Box 2, Box 3) each contain 99 balls (red and blue combined), but the proportion of red and blue balls differs across boxes. The sequence of the three rounds will be randomly chosen by the computer from the following three possibilities:

- Sequence 1: Box 1 → Box 2 → Box 3
- Sequence 2: Box 3 → Box 1 → Box 2

a different interval. According to the definitions of interval above, if your Stage 1 prediction for the number of red balls is less than 50, the computer will randomly select one participant who predicted 50 or more red balls in Stage 1; if your Stage 1 prediction is 50 or more, the computer will randomly select one participant who predicted fewer than 50 red balls in Stage 1.” Similar replacement also takes place in Treatment CA.

- Sequence 3: Box 2 → Box 3 → Box 1

2. After completing the first two rounds, you **will not be informed** of the actual number of red balls in these boxes or your earnings from those rounds. Only after all three rounds (and the post-experiment questionnaire) are complete will your screen display for each round: (1) the randomly selected stage for payment and your predicted number of red balls for that stage; and (2) your earnings for each round and your total earnings for the entire experiment.

Before starting the experiment, you need to correctly answer several quiz questions.

Instructions for Part 2 of the Experiment

This part consists of two tasks.

Task 1

In front of you are three doors: behind one door is a prize of 5 RMB, and behind each of the other two doors is a prize of 1 RMB. Please choose one door. After making your initial choice, the computer will provide some additional information, after which you will have the opportunity to change your choice and confirm your final decision. Your final choice determines your payoff.

Task 2

You need to make 10 decisions in a decision table. Each row corresponds to one decision, in which you must choose the option you prefer between options A and B. Note that only one of these 10 decisions will be randomly selected for payment. Each decision has an equal chance of being chosen, so please make your selections carefully.

Once the computer randomly selects one row, your earnings will be determined as follows: if you choose Option A for that decision, you will earn 4 RMB. If you choose Option B, you have a chance of earning either 12 RMB or 0 RMB. The computer will generate a random result according to the probability specified in the decision to determine whether you receive 12 RMB or 0 RMB.

C Quiz Questions (Baseline treatment, translated into English)

1. Today's experiment consists of _____ rounds, each with a virtual box. In these boxes, the total numbers of red and blue balls are _____, and the proportions of blue and red balls in these boxes are _____.

   (a) three, the same and equals 99, different

   (b) five, different, different

   (c) three, different, possibly the same or different

   (d) five, the same and equals 99, possibly the same or different

2. In Stage 1 of each round, you can see _____ balls, and the color composition of the balls you observed and those observed by other participants is _____.

   (a) three, different

   (b) five, different

   (c) three, possibly the same or different (depending on computer's draw)

   (d) five, possibly the same or different (depending on computer's draw)

3. In Stage 2–11 of each round, you can see _____ balls. The balls comes from other participants that are _____ across stage.

   (a) three, different

   (b) five, different

   (c) three, possibly the same or different (depending on computer's draw)

   (d) five, possibly the same or different (depending on computer's draw)

4. In each round, your experimental earnings are _____. The experiment will pay _____ rounds in total.

   (a) randomly selected from Stages 1–11, three

   (b) randomly selected from Stages 2–11, three

   (c) the summation of earnings in Stages 2–11, one randomly selected

   (d) the summation of earnings in Stages 1–11, one randomly selected

5. In each round, balls in Stage 1 are drawn _____, and balls in Stage 2–11 are drawn _____.

   (a) with replacement, with replacement

   (b) with replacement, without replacement

   (c) without replacement, with replacement

   (d) without replacement, without replacement

6. Suppose that in some stage, you predicted 27 red balls in the box, and the actual number of red balls is 67. (Assuming this stage is randomly selected for payment,) then you earn _____ points in this stage. (enter an non-negative integer between 0–150).

7. Suppose that in some round, you predicted 55 red balls in Stage 1 of this round, then in Stages 2–11 of this round, the 3 balls you will observe comes from participants whose prediction of red balls in Stage 1 is _____.[5]

   (a) less than 50

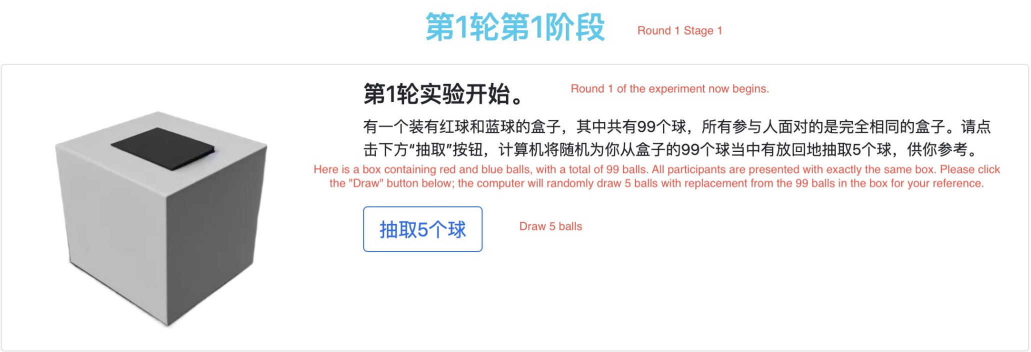
   (b) greater than or equal to 50

   (c) possibly less than 50, or greater than or equal to 50

   (d) none of the above is correct

[5] This question appears only in Treatments IA and CA.

D Screenshots of the Experimental Interface (Urn Inference Task)[6]



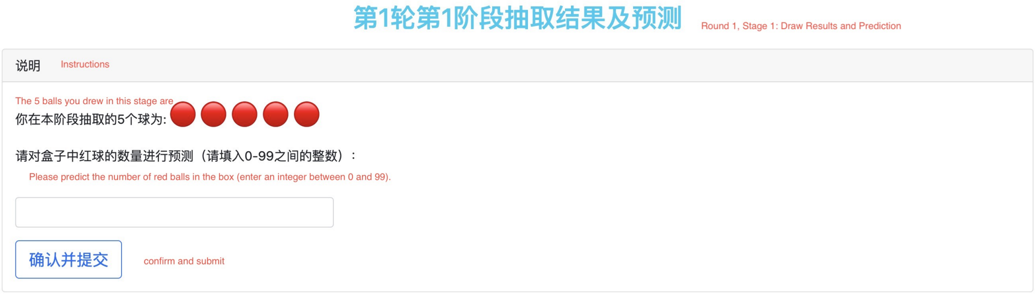
第1轮第1阶段 Round 1 Stage 1

第1轮实验开始。 Round 1 of the experiment now begins.

有一个装有红球和蓝球的盒子，其中共有99个球，所有参与人面对的是完全相同的盒子。请点击下方“抽取”按钮，计算机将随机为你从盒子的99个球当中有放回地抽取5个球，供你参考。

Here is a box containing red and blue balls, with a total of 99 balls. All participants are presented with exactly the same box. Please click the "Draw" button below; the computer will randomly draw 5 balls with replacement from the 99 balls in the box for your reference.

Figure S2: Stage 1 information



第1轮第1阶段抽取结果及预测 Round 1, Stage 1: Draw Results and Prediction

说明 Instructions

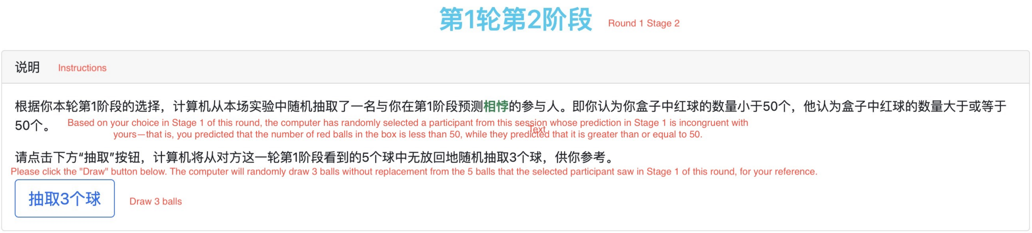
The 5 balls you drew in this stage are

你在本阶段抽取的5个球为: ● ● ● ● ●

请对盒子中红球的数量进行预测（请填入0-99之间的整数）：

Please predict the number of red balls in the box (enter an integer between 0 and 99).

Figure S3: Stage 1 draw results and prediction



第1轮第2阶段 Round 1 Stage 2

说明 Instructions

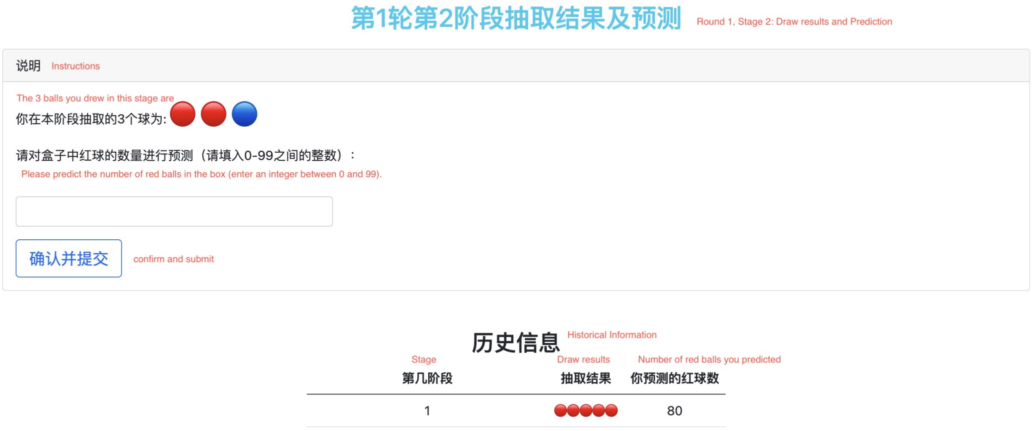
根据你本轮第1阶段的选择，计算机从本场实验中随机抽取了一名与你在第1阶段预测相悖的参与人。即你认为你盒子中红球的数量小于50个，他认为盒子中红球的数量大于或等于50个。 Based on your choice in Stage 1 of this round, the computer has randomly selected a participant from this session whose prediction in Stage 1 is incongruent with yours—that is, you predicted that the number of red balls in the box is less than 50, while they predicted that it is greater than or equal to 50.

请点击下方“抽取”按钮，计算机将从对方这一轮第1阶段看到的5个球中无放回地随机抽取3个球，供你参考。

Please click the "Draw" button below. The computer will randomly draw 3 balls without replacement from the 5 balls that the selected participant saw in Stage 1 of this round, for your reference.

Figure S4: Stages 2-11 information

[6] Screenshots for other tasks are available upon request.



第1轮第2阶段抽取结果及预测 Round 1, Stage 2: Draw results and Prediction

说明 Instructions

The 3 balls you drew in this stage are
你在本阶段抽取的3个球为: ● ● ●

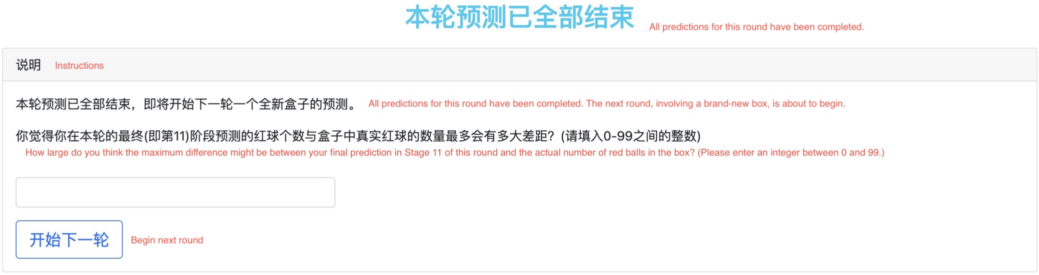
请对盒子中红球的数量进行预测（请填入0-99之间的整数）：
Please predict the number of red balls in the box (enter an integer between 0 and 99).

确认并提交 confirm and submit

历史信息 Historical Information

Stage 第几阶段	Draw results 抽取结果	Number of red balls you predicted 你预测的红球数
1	●●●●●	80

Figure S5: Stages 2-11 draw results and prediction



本轮预测已全部结束 All predictions for this round have been completed.

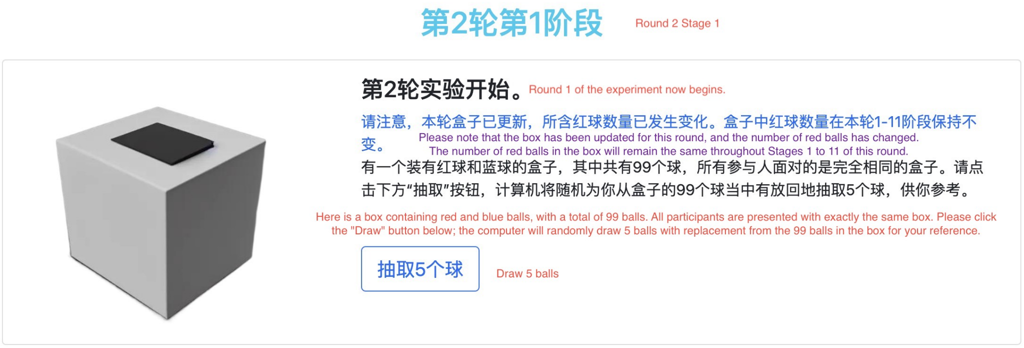
说明 Instructions

本轮预测已全部结束，即将开始下一轮一个全新盒子的预测。 All predictions for this round have been completed. The next round, involving a brand-new box, is about to begin.

你觉得你在本轮的最终(即第11)阶段预测的红球个数与盒子中真实红球的数量最多会有多大差距？(请填入0-99之间的整数)
How large do you think the maximum difference might be between your final prediction in Stage 11 of this round and the actual number of red balls in the box? (Please enter an integer between 0 and 99.)

开始下一轮 Begin next round

Figure S6: Confidence elicitation after Stage 11



第2轮第1阶段 Round 2 Stage 1

第2轮实验开始。Round 1 of the experiment now begins.

请注意，本轮盒子已更新，所含红球数量已发生变化。盒子中红球数量在本轮1-11阶段保持不变。
Please note that the box has been updated for this round, and the number of red balls has changed.
The number of red balls in the box will remain the same throughout Stages 1 to 11 of this round.

有一个装有红球和蓝球的盒子，其中共有99个球，所有参与人面对的是完全相同的盒子。请点击下方“抽取”按钮，计算机将随机为你从盒子的99个球当中有放回地抽取5个球，供你参考。

Here is a box containing red and blue balls, with a total of 99 balls. All participants are presented with exactly the same box. Please click the "Draw" button below; the computer will randomly draw 5 balls with replacement from the 99 balls in the box for your reference.

抽取5个球 Draw 5 balls

Figure S7: Information at the beginning of a new round (urn)