How People Think Others Update Beliefs: An Experiment*

Marina Agranov† and Polina Detkova‡

March 18, 2026

Abstract

We study how people think *others* update their beliefs when confronted with new evidence. We find support for the Martingale property: when two individuals share the same prior, one believes that new evidence cannot systematically shift the other's beliefs in either direction. When priors differ, however, people expect information to move others' beliefs toward their own prior. Yet these expected adjustments respond too weakly to changes in information quality relative to Bayesian predictions. We identify the main source of this insensitivity and discuss its implications.

1 Introduction

The crucial element of many strategic settings is forming beliefs about how *other* players update their beliefs upon encountering new information. This step is imperative for choosing optimal actions.[1] Forming beliefs about how others update is necessary in games with asymmetric information (Spence, 1973), coordination games (Morris and Shin, 2002), social learning settings (Bikhchandani et al., 2024), and global games (Carlsson and van Damme, 1993) among others. In all these environments, one's actions depend on how one expects new evidence to affect other players' beliefs, which, in turn, affect other players' actions.

Much is known about how people update their own beliefs in response to new evidence (we survey this literature in Section 1.1). At the same time, little is known about how people think others update their beliefs. This is the focus of our paper. We study how second-order beliefs respond to information—that is, how individuals believe others revise their beliefs when confronted with new evidence. As discussed above, understanding this process is essential for many strategic interactions, making it a particularly suitable subject for experimental investigation. We exploit the advantages of controlled laboratory experiments to provide some of the first empirical

*Agranov gratefully acknowledges the support of NSF grant SES-2214040. We thank Odilon Camara, Duarte Goncalves, Daniel Gotlieb, Navin Kartik, Kirby Nielsen, Jacopo Perego, Leeat Yariv, and Sevgi Yuksel for helpful comments and suggestions.

†Caltech and NBER. Email: marina.agranov@gmail.com.

‡Royal Holloway. Email: pdetkova@gmail.com.

[1] Indeed, in equilibrium, people are expected to predict correctly others' revised beliefs and their actions and best-respond to it.

evidence on this core component of strategic behavior, isolating it from the confounding factors that typically arise in strategic settings.

We conduct a series of experiments and empirically document how people think others revise their beliefs and how that relates to people's own belief updating process. Our study examines participants' genuine, home-grown beliefs about various factual statements—some neutral and rooted in general knowledge, while others politically charged. To simplify the exposition, throughout the paper, we refer to two players, Anne and Bob. Anne is tasked with predicting Bob's beliefs. In all treatments, Anne knows Bob's prior and the accuracy of the information structure from which Bob receives his signals. However, in some treatments, Anne directly observes the signal realization Bob receives, while in others, she does not. In the former case, Anne predicts *Bob's conditional posterior*, while in the latter she predicts *Bob's expected posterior*.

The treatment variation described above reflects two distinct but common scenarios. In the first, imagine a friend sends you a link to a news article. You want to understand how this information has changed your friend's original opinion—that is, you aim to infer his conditional posterior based on your knowledge of his initial beliefs, the reliability of the news source, and the specific content of the news. In the second scenario, you contemplate advising your friend to read a particular news source in the future. Beforehand, you try to predict how this might influence his beliefs on average, without knowing the exact piece of news he will encounter. This corresponds to predicting his average posterior given his prior and the accuracy of the source, but not the signal realization.[2]

We evaluate our experimental results using two benchmarks. The first benchmark is theoretical and builds on the recent paper by Kartik et al. (2021), which derives predictions grounded in principles of Bayesian updating. We detail these predictions in the next paragraph. The second benchmark is behavioral: we compare Anne's predicted conditional and expected posteriors for Bob with Bob's elicited posteriors. This comparison enables us to quantify the extent to which Anne accurately internalizes Bob's belief-updating process in response to information.

According to the theory, Anne's beliefs about Bob's conditional posterior should be independent of her own prior beliefs; it should only depend on Bob's prior and signal precision. The predictions about Bob's expected posterior are more nuanced. If Anne and Bob share the same prior, Bob's expected posterior should be the same as Bob's and Anne's priors. This prediction is a fundamental property of Bayesian updating: beliefs are Martingale, meaning that new information cannot systematically alter beliefs in any direction. However, if Anne and Bob have different priors, Kartik et al. (2021) show that Anne expects that any new information will shift Bob's expected posterior closer to her own, with the gap between the two narrowing as the signal becomes more accurate. This theoretical result is known as information validates prior (IVP).[3] Motivated

[2]We use the terms information structure accuracy, signal precision, and information quality interchangeably.

[3]These results hold in settings that satisfy standard ordering assumptions: the priors must be likelihood-ratio ordered, and the signal structures from which Bob draws new evidence must satisfy the monotone likelihood-ratio property. When the state is binary, as in our experiment, these assumptions are nonrestrictive. Moreover, for the binary state, the IVP property is equivalent to the result obtained in Francetich and Kreps (2014), according to which conditional on the event being true, the expected posterior is bigger than the prior. However, as discussed in Kartik et al.

by these theoretical predictions, our experiment features variation in the distance between Anne's and Bob's priors and information structures with different accuracies.

To build intuition for these properties, consider an extreme example with two information sources: one is uninformative and generates signals at random, while the other is fully informative and always produces signals that match the true state. Suppose Anne and Bob hold different priors. How do these two information sources affect Bob's average posterior from Anne's perspective? If Bob samples from the uninformative source, his average posterior will remain equal to his prior, since the signals convey no information about the state. In contrast, if he samples from the fully informative source, his average posterior will align exactly with Anne's prior. This is because Anne believes that signal frequencies are determined entirely by her own prior, while the signal realizations reveal the state perfectly. The IVP property generalizes this intuition: As the information structure becomes more precise, Bob's average posterior moves closer to Anne's prior. When Anne and Bob share the same prior, the situation changes. In this case, Anne's prediction of Bob's average posterior is not affected by the quality of the information source and always equals their common prior, as summarized by the Martingale property.

Our empirical results show strong support for the Martingale property: regardless of signal precision, when Anne and Bob share the same prior, Anne believes that Bob's expected posterior will be equal to his prior.[4] Regarding the IVP property, we find partial support. In line with the IVP, Anne believes that *any* new evidence will decrease the disagreement between them, shifting Bob's expected posterior closer to her prior. This is true for all statements, including the politically charged ones. However, more precise information structures only marginally enhance this effect. We find a few small exceptions to this for neutral statements depending on Anne's prior and her relation to Bob's prior. Otherwise, Anne expects Bob's beliefs to be fairly rigid and not responsive to the quality of information he samples from, diverging from what Bayesian theory predicts.

To understand the weak precision effect—that is, why Anne's beliefs about Bob's expected posteriors are less responsive to information quality than predicted—we conduct both reduced-form and structural analyses of the two components that jointly determine expected posteriors. The first is Anne's beliefs about Bob's conditional posteriors, and the second is Anne's beliefs about signal frequencies. We compare both components to Bayesian benchmarks and, in the case of conditional posteriors, also to Bob's actual conditional posteriors observed in the experiment, which correspond to Anne's own posteriors

We find that Anne's beliefs about Bob's conditional posteriors mirror the way she updates her own beliefs, except that she expects Bob to underinfer from his prior more strongly than she does herself- and more strongly than he actually does.[5] This underinference is stronger for

(2021), neither of the two results (Kartik et al. (2021) and Francetich and Kreps (2014)) nests each other in a more general setting beyond binary signals.

[4]These results echo the support for the Martingale property documented in Danz et al. (2024) albeit in a very different setup.

[5]While projection bias has been documented in other contexts (Loewenstein et al., 2002; Danz et al., 2024; Madarasz, 2016), our findings provide one of the first evidence of this phenomenon in the context of belief updating.

politically charged statements.[6] Moreover, we present a novel finding indicating that both Bob's actual corner beliefs and Anne's beliefs about Bob's corner beliefs (i.e., a prior of 0% or 100%) are not as rigid and degenerate, as previously thought; Bob is willing to revise these beliefs when confronted with contradictory evidence and Anne anticipates Bob to do so. Overall, these patterns lead to Bob's conditional posteriors being less sensitivity to changes in Anne's prior and remaining relatively similar regardless of the quality of the information he consumes — this constitutes the first "flattening" effect.[7]

The second element affecting Anne's beliefs about Bob's expected posterior is the signal frequencies which define how Bob's conditional posteriors are weighted in the expected value. Our data shows that Anne places substantial weight on her own prior when forming beliefs about signal frequencies, while also incorporating Bob's prior. This implies that Anne expects signal frequencies to be less responsive to her own prior and the quality of information Bob consumes relative to the Bayesian prediction, which constitutes the second "flattening" effect.

Together, the two flattening effects produce the weak precision pattern documented in the aggregate results: Bob's expected posteriors remain relatively stagnant, showing limited responsiveness to the quality of information he consumes. This finding is important as it underscores the limited impact of information in altering perceptions of others' beliefs, and consequently influencing behavior. Given that information plays a vital role in economic settings, particularly as a tool for policy interventions, this result challenges the presumption of its efficacy in driving collective action.

In our analysis, we use a combination of reduced-form analysis and structural estimations. Many of the results discussed above are evident from the raw data without the need for a behavioral model. However, the structural approach provides a parsimonious framework to capture observed patterns and allows us to conduct counterfactual exercises, which often require extrapolation beyond the parameters directly observed in the experimental data. In one such exercise, we compare the magnitudes of the two "flattening" effects. We find that Bob's non-responsiveness to information quality relative to the Bayesian benchmark is primarily due to a lack of sensitivity in his conditional posteriors to the quality of the information, rather than a flattening in signal frequencies.

At a high level, our experimental results reveal three main patterns. First, individuals expect others to update their beliefs in broadly the same way they update their own, suggesting that people do not construct a fundamentally different model for second-order beliefs. Second, people partially incorporate others' prior beliefs, even when those priors contradict their own, but the adjustment is limited. Third, information has only a modest effect in bridging the beliefs of individuals who initially disagree. Taken together, the latter two findings suggest that disagreements

[6] Our results regarding political statements are in line with the studies that demonstrate under-updating of motivated beliefs compared to neutral ones (Möbius et al., 2022).

[7] This effect may be even more pronounced, given that Anne forms expectations about Bob's posterior without observing the signal realization, as Aina et al. (2023) show that contingent belief updating leads to even flatter beliefs than conditional updating.

are persistent and that information alone is insufficient to eliminate them.

Our findings extend beyond the settings discussed at the beginning of the introduction and offer broader insights into societal polarization. A substantial body of research in Political Science and Economics has focused on the drivers of polarization in the United States, which has deepened over recent decades, and explored potential solutions to mitigate it (McCarthy, 2019; McCarthy et al., 2006). The mere abundance of news sources, many of which exhibit some degree of political bias, does not alleviate polarization (DellaVigna and Kaplan, 2007; Martin and Yurukoglu, 2017; Azzimonti and Fernandes, 2023). Individuals tend to select information sources aligned with their pre-existing beliefs (Garrett, 2009; Stroud, 2010), reinforcing their prior convictions when consuming such content, in line with the martingale property of beliefs. A natural question arises: what if individuals were exposed to news from opposing political perspectives, such as Democrats reading Republican-aligned news? According to the IVP property, polarized groups may believe that exposure to such information would help bridge the gap between them and reduce division, particularly when the news sources are highly accurate. However, our results challenge this approach, showing that while exposure to different viewpoints does shift expectations about others' beliefs, the change is modest, unaffected by the quality of the news source. Thus, people appear to doubt that others would substantially revise their opinions, potentially discouraging attempts to share information to change their views.

1.1 Connection to the Literature

Our experiment is based on the theoretical results derived in Kartik et al. (2021) and heavily uses the insights from experimental literature that studies first-order beliefs. In recent decades, we have learned a lot about how people update their beliefs upon encountering new evidence. Benjamin (2019) provides an excellent and comprehensive review of empirical research from both Economics and Psychology, identifying consistent patterns and notable deviations from Bayesian theory. While some findings support Bayesian predictions, others highlight systematic discrepancies. Recent contributions to the field include Esponda et al. (2023), Augenblick et al. (2024), Ba et al. (2023), Gneezy et al. (2023), Enke and Graeber (2023), and Agranov and Reshidi (2024) among others. By and large, this literature finds that while belief revisions generally follow the direction predicted by Bayesian theory, the magnitude of these revisions often deviates from the expected levels.

Most studies in this branch of literature employ neutral contexts and induce participants' priors to establish a controlled baseline for initial beliefs. An exception is the recent study by Thaler (2024), which elicits participants' genuine beliefs on politically charged topics such as crime, climate change, gun control, and racial discrimination. This study finds that individuals distort new information in favor of their pre-existing views, consistent with motivated reasoning mechanisms. Its design elegantly differentiates this explanation from Bayesian updating motives. Like Thaler (2024), we use genuine beliefs in our experiment but pursue a different research question focusing on how people think others revise their genuine beliefs upon receiving new

information.

Our paper contributes to a growing experimental literature that studies higher-order beliefs and higher-order rationality. Most of this literature focuses on strategic settings: higher-order beliefs play an important role in these settings as they affect what actions players take.[8] For instance, Manski and Neri (2013) elicit the subjects' first- and second-order beliefs in the Hide-and-Seek game and examine the coherence between these beliefs and actions. The results show remarkable consistency: observed choices are optimal given first-order beliefs in 89% of the time and in 75% of the time given second-order beliefs. Healy (2024) elicits participants' preferences over game outcomes, their strategies, as well as first- and second-order beliefs in a series of classical games, including Prisoners' Dilemma and the Centipede game. The data reveals heterogeneity in participants' preferences, which are not captured by game payoffs, but this heterogeneity only partially explains the gap between participants' beliefs and their own actions. Kneeland (2015) studies the Ring Games and demonstrates that over 70% of players are both rational and believe in others' rationality, though this decreases for higher-order beliefs. Friedenberg and Kneeland (2024) extend this work to distinguish between players who have limited reasoning abilities and those who can reason iteratively but have limited belief in others' rationality and find that over 60% of participants engage in strategic reasoning beyond basic rationality. Calford and Chakraborty (2023) show that the discrepancies in one's belief about an opponent and one's beliefs about others' beliefs about that opponent affect deviations from subgame perfection in a sequential social dilemma. Szkup and Trevino (2020) infer how people think others update their beliefs in a coordination game with incomplete information, i.e., the global game.[9] Thaler (2025) elicits beliefs of senders about the motivated reasoning of receivers and demonstrates that they adjust their message accordingly to these beliefs.

Some of the papers discussed above infer second-order beliefs from participants' actions and participants' beliefs about others' actions, while others elicit second-order beliefs directly. Our paper uses the latter approach and elicits second-order beliefs directly without relying on inference techniques. However, different from this literature, we deliberately focus on a non-strategic environment, which provides a clean free-from-strategic-considerations play-field to document how people think others revise their beliefs in response to new information. In the similar spirit, Evdokimov and Garfagnini (2022) investigate higher-order beliefs in a three-player game where participants receive either private or public signals about the state. Player 1 reports his beliefs, Player 2 reports second-order beliefs, and Player 3 reports third-order beliefs. The authors find that belief updating is slower with private information, and higher-order learning often fails. In

[8] There are several excellent surveys of belief elicitation in experiments including Trevino and Schotter (2014), Charness et al. (2014), Schlag et al. (2015), and Healy and Leo (2024). The survey of Trevino and Schotter (2014) provides a detailed discussion of elicitation methods used to recover second-order beliefs.

[9] Another class of games in which second-order beliefs are crucial is psychological games. In these games players' payoffs depend not only on material payoffs but also on the first-, second-, and possible higher-order beliefs about one's opponent. For instance, Dufwenberg and Gneezy (2000) elicit players' first- and second-order beliefs in the Lost Wallet game, Charness and Dufwenberg (2006) do so in the Trust Game, and Agranov et al. (2024) do so in an extended version of the sender-receiver game. This literature is more distantly related to our current study.

contrast, we test key Bayesian properties—specifically, the Martingale and IVP properties—and focus on how second-order beliefs respond to new information.

Finally, our paper contributes to a growing literature on the perception of biases in others. Danz et al. (2024) study how individuals project their own information onto others and anticipate—yet underestimate—others' projection onto them, as predicted by the behavioral model of Madarasz (2016). Their results strongly support the projection equilibrium model and, among other findings, document adherence to the Martingale property in a very different setting. Fedyk (2024) finds that individuals exhibit sophistication about others' present bias while remaining largely naive about their own—a phenomenon known in psychology as the "bias blind spot" (Wang and Jeon, 2020; Pronin et al., 2002, 2004). Related to this, we show that people overestimate the compression of beliefs in others, linking our results to the broader literature on misperceptions about others (see Bursztyn and Yang, 2022, for a review). Trujano-Ochoa (2024) studies the extent to which people account for others' biases in belief updating and information acquisition.[10]

The rest of the paper is structured as follows. We present the conceptual framework in Section 2. Section 3 describes our experimental design and the experimental procedures. Section 4 presents main results on Martingale and IVP properties. Sections 5 and 6 unpack the aggregate results and study how they emerge. Section 7 offers some conclusions.

2 Conceptual Framework

Consider a standard belief-updating task with a binary state $\omega \in \{0,1\}$. There are two decision-makers, Anne and Bob, who may have the same or different priors about the state. We denote by a_0 and b_0 the prior of Anne and Bob, respectively. These priors indicate the probability that the state is $\omega = 1$ according to each of the two decision-makers. The benchmark results discussed in this section follow Kartik et al. (2021) and treat the initial prior beliefs as dogmatic: Anne does not revise her own prior upon learning that Bob's prior may differ, i.e., $a_0 \neq b_0$. In this sense, we abstract from the origins of these initial beliefs and assume that Anne gains no information about the state from observing Bob's prior. In Section 6.2, we relax this assumption and explore the possibility that Anne updates her prior after observing Bob's, drawing on the model of social exchange developed by Yuksel and Oprea (2022).[11]

Bob receives a partially informative signal s and updates his beliefs about the state. We denote by b_s Bob's posterior belief after observing signal s and refer to it as *Bob's conditional posterior*. The signals are also binary and have accuracy θ. The accuracy of a signal indicates the likelihood that the signal matches the state conditional on the state, i.e., $\theta = \Pr[s = \omega|\omega]$.

[10] This work is still in progress. The preliminary draft we have access to suggests that, on average, people expect others to update similarly to themselves, but exhibit a significantly lower willingness to pay when others' strategies are implemented on their behalf. This latter finding is consistent with expecting others to update more conservatively than oneself.

[11] The current setup can also be reinterpreted as one based on a common prior, with Bob and Anne holding different interim beliefs due to differences in the information they have received. The qualitative results discussed in this section continue to hold under this interpretation, as long as Anne's and Bob's interim beliefs differ.

Anne knows both Bob's prior and signal accuracy, and attempts to predict Bob's average posterior after he receives a signal. The challenge arises because Anne does not observe the signal Bob receives; instead, she must weigh Bob's conditional posteriors according to the likelihood of the signals. We call this object *Bob's expected posterior*, and denote it by $\mathbb{E}^{\mathrm{A}}[b]$. The superscript A stresses that this is Anne's beliefs about Bob's expected posterior. If Anne is Bayesian and expects Bob to be Bayesian, then

$$\mathbb{E}^{\mathrm{A}}[b] = \Pr[s=1] \cdot b_{s=1} + \Pr[s=0] \cdot b_{s=0} \tag{1}$$

where

$$b_{s=1} = \frac{b_0\theta}{b_0\theta + (1-b_0)(1-\theta)} \quad , \quad b_{s=0} = \frac{b_0(1-\theta)}{b_0(1-\theta) + (1-b_0)\theta} \, ,$$
$$\Pr[s=1] = a_0\theta + (1-a_0)(1-\theta) \text{ , and } \Pr[s=0] = 1 - \Pr[s=1].$$

In words, Anne expects Bob's prior b_0 to influence how Bob updates his beliefs for a given signal, while her own prior a_0 to determine the signal probabilities. It is easy to see that when Anne and Bob share the same prior, we recover the fundamental property of Bayesian updating: beliefs are *Martingale*, i.e., information cannot systematically bias beliefs in any direction. This means that Anne's belief about Bob's expected posterior, which is the same as her own posterior should be equal to her and his prior.

When Anne and Bob have different priors the situation changes and *any* information is predicted to move Bob's expected posterior closer to Anne's prior. The recent paper by Kartik et al. (2021) shows that Anne expects a Blackwell more informative signal to bring Bob's expected posterior closer to her own prior. To translate this to our setting, say, Bob has access to two information structures that only differ in signal accuracy, $1 > \theta_1 > \theta_2 > \frac{1}{2}$. Then, Anne expects that both signal structures will move Bob's average posterior closer to her prior compared to what Bob's original prior was. Moreover, she anticipates that the structure with more precise signals θ_1, will result in a larger shift and a smaller final disagreement between Anne's prior and Bob's expected posterior.

This result is known as the *Information Validates Prior (IVP)* property. As we argued in the introduction, its significance is broad, spanning many strategic settings studied in Economics and Political Science. It is precisely this result that we set out to investigate empirically in our paper.

3 Experimental Design

Given our interest in how participants think others update their beliefs when they may have potentially different priors we chose to work with genuine, home-grown beliefs participants have about various facts. In the next section, we discuss the advantages and disadvantages of using this method compared to induced beliefs.

Specifically, we used twelve factual statements in the experiment. Each statement is either true or false. Participants know that the experimenter knows whether the statement is true or false, but naturally may hold different beliefs about the probability that a statement is true. Here

are two examples of such statements[12]:

- In 2023, the United States spent more than 10% of the federal budget on foreign aid.
- Rhino horn is made up of keratin - the same protein which forms the basis of our hair and nails.

Treatments. The experiment consists of three main treatments. Treatment T0 is the benchmark treatment, in which we document how people update their own beliefs in response to new information. The purpose of treatment T1 is to study how Anne thinks Bob updates his beliefs when she knows Bob's signal, i.e., Anne's beliefs about Bob's conditional posteriors. Finally, the purpose of treatment T2 is to study Anne's beliefs about Bob's expected posterior, i.e., the situation in which Anne does not observe Bob's signal.

Structure of the experiment. Each treatment consists of three parts. Participants receive the instructions for the next part after they complete the previous one. Instructions before each part include a comprehension quiz to check participants' understanding and focus their attention on the main features of the experiment. The instructions and the screenshots are presented in the Online Appendix.

Part 1 consists of 6 rounds and is the same in all treatments. In each round, we present participants with one of the statements and elicit their priors about the chance that the statement is true. We then provide participants with a partially informative signal about the correctness of the statement and elicit their posterior about the chance that the statement is true. Signals are generated from two signal structures: a more precise one with $\theta_1 = 0.90$ and a less precise one with $\theta_2 = 0.65$. One of these two structures is randomly selected in each round, and a participant knows signal accuracy when she makes her choices. We will use these two signal structures to investigate the IVP property which requires comparing the more- and the less-precise information structures. Participants receive no feedback at the end of each round in Part 1. After completing a round, they move on to the next one and are shown the next statement.

Part 2 consists of 6 rounds as well and is different in each treatment. In T0, Part 2 is the same as Part 1. That is, participants go through another set of 6 statements, report their priors, observe signals, and report their posteriors. A key reason for collecting extensive data on participants' own belief updating is to calculate participants' payments in treatments T1 and T2. This requires observing posteriors for each statement across different signal realizations within various signal structures, which is what we do in T0. We conducted T0 a few days prior to the other treatments to ensure this data would be available for payment calculations.

Part 2 in the remaining two treatments is slightly different. In each round, participants start by observing a statement and reporting their prior. Then, they are matched with past participants

[12] Figures 7 and 8 in the Appendix present all statements and the visualization used alongside the statements.

from T0 and observe the past participants' prior for the same statement and signal accuracy.[13]

In T1, each current participant is matched with a past participant who reported a specific prior. The current participant then observes the signal realization received by the matched past participant and is asked to report that participant's posterior belief conditional on the signal.

In T2, each current participant is matched with a group of past participants, all of whom reported the same prior. In contrast to T1, participants in T2 do not observe the signals received by the past participants. Instead, their task is to report their belief about the average posterior of this group of past participants with the pre-specified prior.

The last part, Part 3, consisted of just one round and was administered only to participants who reported a corner prior for one of the statements. If such an event happened, then one of the questions for which a corner prior was reported was chosen and a participant was offered a choice between a very risky bet and a safe payment of \$10. The risky bet pays \$11 in case the reported prior is correct and \$0 if it is wrong. The goal of this final (surprise) round was to gauge how much faith people have in their corner beliefs when they report them. Risking losing \$10 makes sense only if one has little doubt in the reported belief.

At the end of the experiment, participants answered a few unincentivized questions about the difficulty of the experiment. In addition, following McGranaghan et al. (2024), every 3 rounds, we presented participants with an unincentivized visual brain break to reduce fatigue.

To sum up, data from T0 and from Part 1 in T1 and T2 provide observations of Bob's (and Anne's) actual conditional posteriors. Data from Part 2 in T1 provide observations of Anne's beliefs about Bob's conditional posteriors, while data from Part 2 in T2 provide observations of Anne's beliefs about Bob's expected posteriors.

Order of statements. Since all rounds are the same in Parts 1 and 2 in T0, the order of statements was randomized across participants in this treatment. For T1 and T2, we split the statements into two batches (batch A consists of statements 1 to 6 and batch B consists of statements 7 to 12). We, then, conducted two versions of each treatment: T1A and T2A used batch A in Part 1 and batch B in Part 2, and T1B and T2B used batch B in Part 1 and batch A in Part 2. Within each part, the order of statements was randomized across participants.[14]

Parameters. Testing the IVP and the Martingale properties requires a variation in Anne and Bob's priors, capturing both similar and distinct priors between the two and spanning a wide range of possible priors. To do so, we match T1 and T2 participants with T0 participants with six pre-selected priors, $b_0 \in \{0.10, 0.20, 0.60, 0.70, 0.90, 1.00\}$.[15] For each prior b_0, we selected two

[13] As we describe in the parameters paragraph, we pre-selected six priors for Bob and matched current participants with past participants who reported these priors. Thus, current participants are likely to be matched with different past participants in each round in this part of the experiment.

[14] This design mitigates the concern that some of the patterns we find in the data are driven by specific statements people saw in one part of the experiment.

[15] Figures 7 and 8 in Appendix present the distribution of priors elicited from participants in T0. For each question, one prior was used to serve as the past participant's prior in T1 and T2. We indicate the selected prior for each question

statements that had a sufficient number of participants reporting such a prior in T0 and providing us with posterior beliefs of past participants (participants in T0) for each signal realization and each signal accuracy. As described above, such data is necessary for computing the payments of participants in T1 and T2.[16]

Table 1: Design

Treatment		Part 1 6 rounds	Part 2 6 rounds	Part 3 at most 1 round	Nb participants
T0	elicit observe report	own prior signal acc., signal own posterior	own prior signal acc., signal own posterior	risky bet	201
T1	elicit observe report	own prior signal acc., signal own posterior	own prior other's prior, signal acc., signal other's conditional posterior	risky bet	198
T2	elicit observe report	own prior signal acc., signal own posterior	own prior others' prior, signal acc. others' expected posterior	risky bet	202

Subject pool. The experiments were conducted on the Prolific platform in 2024 with roughly 200 participants in each treatment, for a total of 603 participants. We recruited participants between the ages of 21 and 65, who live in the United States, specify English as their first language, and have a high (90+) approval rating on Prolific. For each treatment, an equal number of men and women were recruited.

Participants' payments. All participants received a fixed payment upon completion: \$3 in the T0 and \$4 in the T1 and T2 treatments.[17] In addition, each participant had a 20% chance to be selected into a bonus group. For the selected participants, the computer randomly chose one of the questions from one randomly selected round for payment. The answer submitted in the chosen question determined whether the selected participant received an additional bonus of \$10. We used the standard BDM method to incentivize subjects to truthfully state their beliefs.[18] Thus, if the own posterior round was selected for bonus, whether the participant receives the payment depends on their reported posterior and on whether the statement is true or false. If, however, the other posterior round was selected for bonus, then it is the report(s) of the previously matched participant(s) that determine whether the current participant receives the bonus. In addition, in

in the figures.

[16] An alternative design would be to match participants from T1 and T2 randomly with past participants from T0. The drawback of this design is that an even larger amount of data is required for T0 to ensure that all signal realizations occur for both signal structures and all priors of past participants, some of which are naturally quite rare.

[17] These completion fees are standard, given the average time to complete each treatment.

[18] The BDM payment is theoretically an incentive-compatible method for eliciting truthful responses regardless of participants' risk attitudes (Becker et al., 1964). In addition, following Danz et al. (2021), we told participants that they had no incentive to report beliefs falsely if they wanted to maximize expected payoff in the experiment. This technique became standard in the literature as it helps participants understand the payment method and, as a result, helps the experimenter elicit participants' true beliefs.

each treatment, we randomly selected eight participants to receive an additional bonus based on their decisions in Part 3 (the corner beliefs). Treatment T0 lasted about 16 minutes and participants earned, on average, $4. Treatments T1 and T2 lasted about 20 minutes and participants earned, on average, $5.

Implementation. The experiment was approved by Caltech (IR24-1446) and preregistered on aspredicted.org (#158497).[19] The experimental software was programmed in Qualtrics. Table 1 summarizes the three treatments.

3.1 Discussion of Experimental Design

In this section, we discuss the rationale behind our key design choice of using genuine, homegrown beliefs instead of inducing beliefs in a neutral context.

To study the IVP property, one needs an environment in which participants have different beliefs. There are two ways to do that. The first approach involves inducing varying beliefs by providing participants with private signals about the state (Andreoni and Mylovanov, 2012). The second approach prescribes eliciting participants' genuine, naturally formed beliefs about certain factual events (Thaler, 2024).[20]

Both methods have their advantages and disadvantages. The primary advantage of inducing beliefs lies in the ability to control participants' beliefs. This is straightforward when inducing a common belief among all participants. However, it becomes more challenging when inducing heterogeneous beliefs, as this requires participants to update their beliefs based on the private signals they receive. Given the extensive literature documenting deviations from Bayesian updating (Benjamin, 2019), it is unclear whether an experimenter employing this approach can effectively control the induced priors.[21]

Working with genuine beliefs sidesteps this issue, as individuals naturally hold differing beliefs on various topics, including factual statements. Moreover, participants are not likely to be surprised when they learn that others have different views.

[19] We conducted two small pilots (pre-registrations #110598 and #124788) with different framings of the belief-updating task to test the software and verify standard behaviors documented in the literature. During the pilots, we identified software errors and realized that our modified framing was unclear to participants. Consequently, we reverted to the standard framing used in the literature, focusing on eliciting genuine priors rather than inducing priors. Results from these pilots are available from the authors upon request.

[20] The focus on factual events as opposed to future events that have not happened yet is dictated by the need to incentivize people to report their beliefs truthfully, which requires the experimenter to know the state—in our case, whether the statement is correct or false.

[21] The additional subtle issue with inducing heterogeneous beliefs is what Anne can infer from Bob's prior about Bob's ability to use new information. To illustrate, consider a standard environment with two urns containing balls of different colors. Both Anne and Bob know the compositions of the urns and the chance that each urn is selected; the selected urn represents the state. Each observes a private draw from the urn and forms a belief about the state. These formed beliefs could potentially serve as Anne's and Bob's priors for the investigation of the IVP and Martingale properties. Say, Bob's posterior belief is communicated to Anne. If this belief is unreasonable given the composition of the urns, then Anne will make inferences about Bob's ability to update already at the inducing-the-priors stage of the experiment, which would confound Anne's beliefs about how Bob updates his beliefs given new information.

Our approach of eliciting genuine beliefs as opposed to induced beliefs offers three additional advantages. First, it provides the enhanced external validity of the results, as they directly speak to how people adjust their natural beliefs in response to new information. Second, this approach enables us to investigate whether genuine beliefs about neutral topics—such as general knowledge statements—respond differently to new information compared to politically charged statements. Our study offers a preliminary exploration of these differences, and we hope future research will expand it and provide more comprehensive evidence. Third, it allows observing genuine corner priors —instances where participants report extreme confidence in the statement being either true or false. These cases are particularly interesting because they allow us to examine whether corner beliefs are degenerate, as theory suggests, or if they can respond to new information. This type of analysis would not be possible with induced beliefs.

Finally, we note two potential concerns associated with using genuine rather than experimentally induced beliefs. First, one might worry that participants could look up the statements online and thus know whether they are true or false. However, our data suggest that this was not a major issue in our study: (a) the majority of subjects report non-corner beliefs, which would be unlikely if they had verified the statements' truthfulness online (Figures 7 and 8); and (b) fewer than 20% of participants report corner beliefs for more than a quarter of the statements (Figure 9). Second, this approach requires collecting a relatively large amount of data to capture sufficient variation in Anne's and Bob's priors needed to evaluate the IVP and Martingale properties. This consideration informed our choice of sample size for the experiment.

4 Main Results

We start with examining Anne's beliefs about Bob's expected posteriors, which is the main object of our interest. We present the evidence on Martingale property in Section 4.1 and the evidence on IVP property in Section 4.2. In the next sections we explore what drives these aggregate results. Section 5 contains reduced-form analysis of Anne's beliefs about signal frequencies and Bob's conditional posteriors. Section 6 complements this analysis with structural estimations that organize observed behavioral patterns in a coherent way and allow for counterfactual exercises. Throughout the analysis we refer to two benchmarks: the Bayesian benchmark and the Bob's actual beliefs observed in the experiment.

Approach to Data Analysis. We define Anne and Bob as having the *same priors* if their priors differ by no more than 5 percentage points, and *different priors* if the difference exceeds 5 percentage points. We further categorize the extent of their differences using the following distinctions. We call Anne's and Bob's priors *very polarized* if they differ by more than 40 percentage points, *polarized* if the difference falls between 20 and 40 percentage points, and *somewhat polarized* if the difference is between 5 and 20 percentage points.

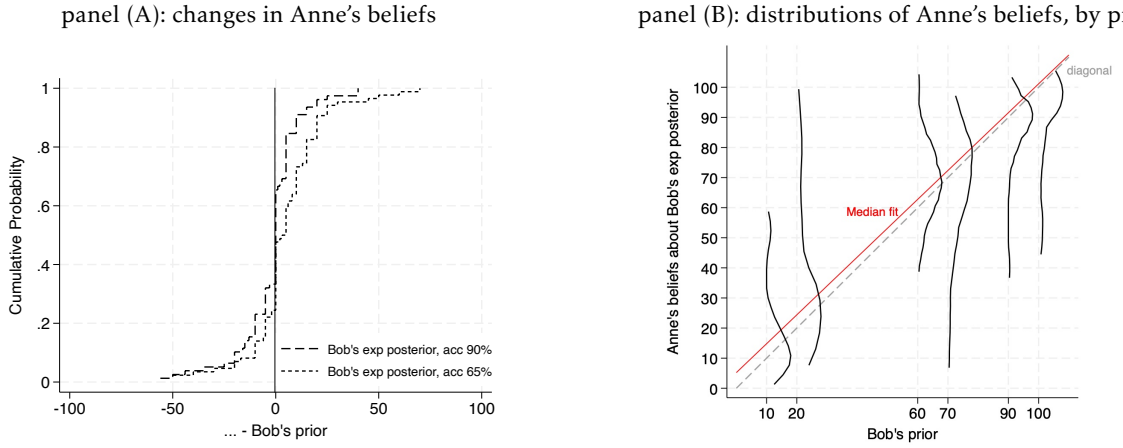
Statistical tests are performed using regressions. We regress the variable of interest (e.g., the

difference between Anne's prediction about Bob's expected posterior and Anne's prior) on a constant and an indicator for one of the treatments (e.g., the test accuracy), while clustering standard errors at the individual level to account for the inter-dependency of observations that come from the same participant. We consider treatments significantly different when the estimated treatment indicator differs from zero at the 5% level and report the corresponding p-value.

4.1 Martingale property

Does Anne expect information to systematically alter Bob's posterior when they share prior? Figure 1 provides the empirical answer to this question.

Figure 1: Anne's beliefs about Bob's expected posterior when the two share the same prior



Notes: Panel (A) shows the CDFs of the differences between Anne's beliefs about Bob's expected posterior and his prior for each signal structure. Panel (B) shows a rotated ridgeline plot of Anne's beliefs about Bob's expected posteriors by Bob's prior, with kernel densities shifted along the prior axis and overlaid with a diagonal benchmark and the median fit. Both panels focus on cases where Anne and Bob have similar priors (within 5 pp). The data come from Part 2 of T2.

Panel (A) in Figure 1 displays the CDFs of the differences between Anne's prediction of Bob's expected posterior and Bob's original prior and shows that the two CDFs are approximately symmetric around zero. The mean difference is not significantly different from zero when test accuracy is 65% ($p > 0.10$). It becomes statistically significant at 90% accuracy, but the magnitude remains small—about 5 percentage points ($p = 0.03$). Moreover, there is no significant difference between the two signal structures ($p > 0.10$).

Panel (B) in Figure 1 shows that these aggregate results continue to hold when the data are disaggregated by Bob's prior. The figure plots kernel distributions of Anne's beliefs about Bob's expected posterior separately for each prior. Under the Martingale property, we expect most of the mass to lie along the diagonal. Consistent with this prediction, Panel (B) shows that the median fit is very close to the diagonal. The regression analysis confirms these visual patterns.[22]

[22] Specifically, we estimate the median regression of Anne's beliefs about Bob's expected posterior on Bob's prior and a

Overall, these results provide strong support for Anne expecting the Martingale property to hold for Bob's beliefs when the two share similar priors.

Observation 1: Anne believes that Bob's beliefs satisfy the Martingale property, i.e., from an ex-ante perspective, information cannot alter Bob's beliefs when the two share the same prior.

4.2 IVP property

What does Anne think about Bob's expected posterior when they have different priors? The IVP property has two empirical footprints. First, Anne expects any information to be effective at bringing Bob's posterior closer to her prior relative to the original disagreement in their priors. Second, the more precise information is expected to decrease disagreements between Anne and Bob by producing larger shifts in Bob's expected posterior relative to the less precise information.

Panel (A) in Figure 2 depicts the CDFs of the absolute differences between Bob's expected posteriors and Anne's priors for the two signal structures, as well as the original difference in opinions (priors) between them.

Consistent with the IVP prediction, any information structure shifts Bob's posteriors closer to Anne's priors (panel (A) in Figure 2). This shift is large in magnitude and statistically significant ($p < 0.01$).[23] Furthermore, more precise signals shift Bob's beliefs closer to those of Anne. In fact, the CDF curve for test accuracy 65% first-order stochastically dominates the one for test accuracy 90%. However, the difference between the two CDFs is rather small and only marginally significant ($p = 0.08$).

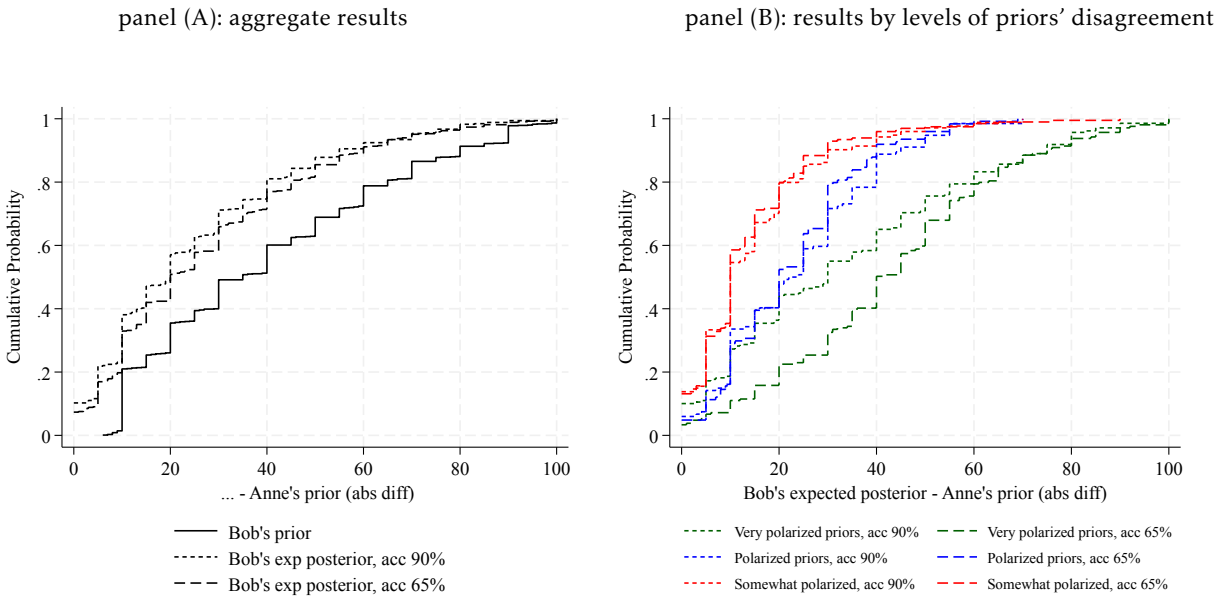
Panel (B) in Figure 2 shows that the difference between the two information structures primarily comes from cases in which Anne's and Bob's original priors are very polarized (at least 40 pp apart). Put differently, when Anne and Bob have very different initial opinions, Anne expects Bob's posterior to move closer to her prior when he learns from a more accurate source (the two green lines). The effect in this case is large in magnitude and highly significant ($p < 0.01$): the median shift is 10 pp and it is almost 20 points when Anne has extreme priors of $a_0 < 0.20$ or $a_0 > 0.80$. At the same time, contrary to the IVP property, when Anne's and Bob's priors are not that polarized, we observe no difference between the two information structures in general (the two blue and the two red lines in panel (B) of Figure 2 are very similar).

The regressions reported in Table 2 provide statistical support for the visual patterns shown in Figure 2. We construct a measure of how much closer (or further away) Bob's expected posterior is from Anne's prior relative to Bob's prior. Specifically, the dependent variable is the difference

constant, restricting attention to observations in which Anne and Bob share the same prior. To account for observations from the same participant, the standard errors are bootstrapped at the individual level. The estimated coefficients are very close to the diagonal: constant $\alpha = 6.43$ ($se = 3.30$), slope $\beta = 0.93$ ($se = 0.04$). We also estimate the OLS regression of the same kind and recover the constant $\alpha = 16.15$ ($se = 3.46$) and slope $\beta = 0.78$ ($se = 0.05$). As expected the linear fit is worse than the median fit given its sensitivity to truncation at the boundaries. Moreover, in both the median and linear specifications, the coefficients on interactions with information structure are not significantly different from zero ($p > 0.05$) confirming that the Martingale property holds for both structures.

[23] The CDFs of the differences between Bob's and Anne's priors under the two signal structures overlap, as expected given the random assignment of signal precision. For brevity, we omit this figure, but it is available upon request.

Figure 2: Anne's beliefs about Bob's expected posteriors when the two have different priors



Notes: Panel (A) depicts the CDFs of the absolute differences between Bob's prior and Anne's beliefs about Bob's expected posterior as well as Bob's and Anne's priors. Panel (B) displays the former difference, broken down by levels of priors' disagreement between Anne and Bob. We focus on cases where Anne and Bob have different priors. The data is from Part 2 of treatment T2.

Table 2: Anne's beliefs about Bob's expected posterior when the two have different priors

	Dependent Variable = $\lvert\mathbb{E}^A[b]-a_0\rvert-\lvert b_0-a_0\rvert$		
	reg (1)	reg(2)	reg (3)
Indicator for acc 90%	-0.05** (0.015)	0.23** (0.02)	-0.06** (0.02)
Indicator for acc 90% × $\lvert b_0-a_0\rvert$		-0.67** (0.05)	
Indicator for political statement			0.02 (0.02)
Indicator for political statement × acc 90%			0.07** (0.03)
Constant	-0.12** (0.01)	-0.12** (0.01)	-0.12** (0.01)
Nb obs	$n = 1048$	$n = 1048$	$n = 1048$
Nb participants	$i = 202$	$i = 202$	$i = 202$
adj R-squared	0.0082	0.2708	0.0168
root MSE	0.2430	0.2085	0.2419

Notes: The OLS regressions with bootstrapped standard errors clustered at the individual level. In all regressions, the dependent variable is the difference between (i) the absolute difference between Anne's belief about Bob's expected posterior and Anne's prior and (ii) the absolute difference between Bob's and Anne's priors. "acc 90%" denotes a high-accuracy signal with $\theta = 0.90$. Politically-sensitive statements are statements 3, 6, and 10 (Figures 7 abd 8). ** (*) indicates significance at the 5% (10%) level.

between (i) the absolute difference between Bob's expected posterior and Anne's prior and (ii) the absolute difference between Bob's and Anne's priors. Larger values of the dependent variable correspond to greater disagreement between Anne's prior and Bob's expected posterior, relative to their original disagreement.

Regression (1) provides support for both footprints of the IVP. First, Anne expects information to move Bob's average opinion closer to her prior, as indicated by the negative and significant constant, which captures the effect of the low-accuracy signal structure, as well as by the negative and significant sum of the constant and the high-accuracy indicator. Second, this effect is slightly stronger for more informative signals, as reflected in the negative and significant coefficient on the high-accuracy indicator, although the magnitude of the difference is small.

Regression (2) examines heterogeneity depicted in panel (B) of Figure 2. It shows that the difference between low- and high-accuracy signals is driven primarily by cases in which the initial disagreement between Anne's and Bob's priors is relatively large. Indeed, the shift in Bob's expected posterior toward Anne's prior is larger for accuracy 90% than for accuracy 65% if and only if $0.23 < 0.67 \cdot |b_0 - a_0|$.

Finally, Regression (3) shows that for politically-sensitive statements, the difference across information structures is essentially non-existent. Figure 10 in Appendix replicates Figure 2 for three politically charged statements used in our experiment.[24] It illustrates that Anne believes Bob's average posterior beliefs will still move closer to her own prior after receiving new information. Yet, unlike neutral statements, these shifts are identical regardless of the quality of the information Bob receives.

Observation 2: We find partial support for the IVP property. Consistent with the IVP, Anne thinks that any information brings Bob's average opinion closer to her own, and more precise information is (marginally) more effective at this job. However, strong precision effect occurs only when Anne has very different prior beliefs from Bob and when the statements are neutral. Otherwise, Anne expects Bob's beliefs to shift similarly regardless of information quality.

5 Reduced-form Diagnostics

In this section, we study what drives aggregate results presented in Section 4. In particular, we investigate the central puzzle in our data: why does signal precision play such a limited role in Anne's expectations of Bob's average beliefs, despite being fundamental in Bayesian theory?

Recall that in the Bayesian world (Section 2), Anne's expectation about Bob's average posterior depends on Bob's conditional posteriors $(b_{s=1}, b_{s=0})$ and the signal frequencies, which depend linearly on Anne's prior belief a_0 and signal precision θ:

$$\mathbb{E}^A[b] = \Pr[s=1] \cdot b_{s=1} + \Pr[s=0] \cdot b_{s=0} \qquad \text{where} \quad \Pr[s=1] = a_0\theta + (1-a_0)(1-\theta) \tag{2}$$

We examine these two components separately. We begin with Anne's beliefs about signal frequencies in Section 5.1, and then turn to Anne's beliefs about Bob's conditional posteriors in Section 5.2. Note that in our experiment we directly observe Anne's beliefs about Bob's conditional

[24] We have three politically sensitive statements: statement 3 about the United States foreign aid spending, statement 6 about the estimates of GDP growth under Democratic vs Republican presidents, and statement 10 about fraction of African-American residents in the United States.

posteriors (elicited in Part 2 of T1). By contrast, we do not elicit Anne's beliefs about signal frequencies, and instead have to estimate these given available data.[25]

5.1 Anne's beliefs about signal frequencies

Recall that before reporting her beliefs about Bob's expected posterior, Anne observes Bob's prior b_0. While in the Bayesian world, Anne is expected to use only her own prior a_0 to compute the likelihoods of signals, in reality, she might doubt whether her prior is reasonable after learning the prior of Bob and as a result might adjust it in some way.

In this section, we examine alternative ways Anne might form beliefs about signal frequencies and assess which specification best fits the data. Table 3 compares several benchmarks for how Anne might form beliefs about Bob's expected posterior. We construct all benchmarks based on $\mathbb{E}^A[b] = \Pr[s=1]\cdot b_{s=1} + \Pr[s=0]\cdot b_{s=0}$, where we vary $\Pr[s=1]$ between the benchmarks and use Anne's beliefs about Bob's conditional posteriors from T1 for $b_{s=1}$ and $b_{s=0}$.

The first benchmark computes signal frequencies using Anne's prior a_0, following equation (2). The next three use weighted averages of a_0 and b_0, with weights on a_0 equal to 0.8, 0.5, and 0.2, respectively. The fifth benchmark places full weight on Bob's prior b_0, and the final benchmark replaces model-implied frequencies with the empirical signal frequencies.

Table 3 reports OLS regressions of Anne's reported beliefs about Bob's expected posterior on the benchmark-predicted values (separately for each specification), along with model fit statistics.

Table 3: How much weight Anne puts on her prior when calculating signal frequencies?

	reg (1) a_0	reg (2) $0.8\cdot a_0 + 0.2\cdot b_0$	reg (3) $0.5\cdot a_0 + 0.5\cdot b_0$	reg (4) $0.2\cdot a_0 + 0.8\cdot b_0$	reg (5) b_0	reg (6) emp freq
Const	-21.5 (3.97)	-17.6 (3.94)	-7.7 (3.84)	3.8 (3.7)	11.2 (3.6)	4.1 (3.9)
Slope	1.36 (0.06)	1.30 (0.06)	1.14 (0.06)	0.95 (0.06)	0.83 (0.05)	0.97 (0.06)
Nb obs	1212	1212	1212	1212	1212	1212
Nb participants	202	202	202	202	202	202
R-squared	0.43	0.42	0.37	0.31	0.27	0.18
root MSE	21.5	21.6	22.2	23.1	23.9	25.3
MAE	17.5	17.6	17.9	18.4	18.7	21.0

Notes: The dependent variable in all regressions is Anne's beliefs about Bob's expected posterior (part 2 in T2). The independent variables include a constant and Anne's *predicted* belief about Bob's expected posterior. The latter is computed using Anne's observed beliefs about Bob's conditional posteriors (part 2 in T1) and signal frequencies defined as in equation 2, except that a_0 is replaced by the expression specified in the second row of the table. For example, in column (3), $\Pr[s=1] = \frac{a_0+b_0}{2}\theta + \left(1-\frac{a_0+b_0}{2}\right)(1-\theta)$, whereas column (6) uses the empirical signal frequencies observed in the experiment. The one to last row specifies root mean-squared errors and the last row lists the mean-absolute errors.

The results highlight different dimensions of fit. The slope and intercept capture co-movement

[25] In principle, one could design an experiment that elicits Anne's beliefs about signal frequencies in addition to her beliefs about Bob's average and conditional posteriors. We chose not to pursue this approach because such elicitation could be leading and might alter how participants naturally form expectations about others' posteriors. For example, a participant who does not spontaneously decompose Bob's average posterior into signal frequencies and conditional posteriors might begin to do so when prompted, thereby placing weight on signal frequencies that would not have been considered absent the elicitation.

and bias, while R^2, root MSE, and MAE measure overall predictive accuracy. Although no single metric is decisive, a clear pattern emerges: benchmarks that place greater weight on Anne's prior a_0 consistently outperform those emphasizing Bob's prior b_0 or empirical frequencies. From column (1) to column (5), R^2 declines from 0.43 to 0.27 and root MSE increases from 21.5 to 23.9, with the empirical-frequency benchmark performing worst on all metrics. This indicates that Anne's own prior is a central input in her perceived signal frequencies.

Overall, the equal-weight specification $0.5a_0 + 0.5b_0$ (column (3)) provides the best balance: its slope is closest to one (1.14), its intercept is small, and it performs favorably across the fit measures. These results suggest that Anne behaves as if she forms beliefs about signal frequencies using a combination of her own prior and Bob's prior. This pattern is reminiscent of the social exchange model of Yuksel and Oprea (2022), which describes how individuals adjust their beliefs after observing the beliefs of others. We examine this model in greater detail in Section 6.2.

Importantly, allowing Anne to revise her prior after observing Bob's prior does not eliminate sensitivity to precision: under all the benchmarks, signal frequencies are expected to remain responsive to changes in precision. We therefore conclude that variation in how Anne forms beliefs about signal frequencies alone cannot account for the weak precision effect observed in our data.

Observation 3: Anne places substantial weight on her own prior when forming beliefs about signal frequencies, while also incorporating Bob's prior. However, this mechanism cannot explain the limited sensitivity of Anne's beliefs about Bob's expected posterior to changes in signal precision.

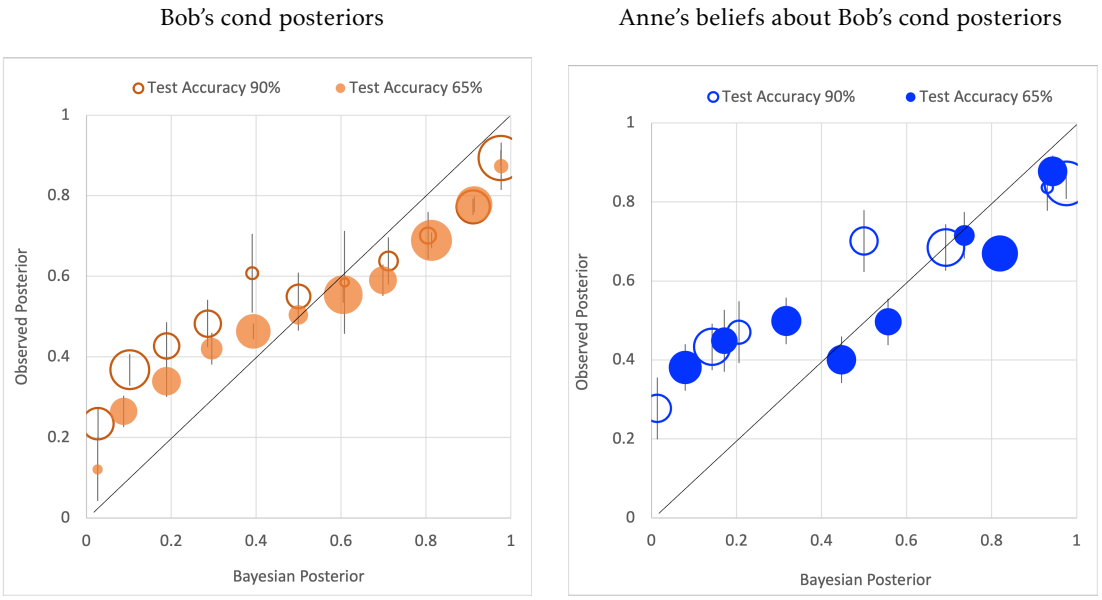
5.2 Bob's conditional posteriors

We now turn to Bob's conditional posteriors and evaluate Anne's beliefs relative to two benchmarks: the Bayesian-predicted conditional posteriors and Bob's actual conditional posteriors (elicited directly in the experiment). Figure 3 presents Bob's actual conditional posteriors in the left panel and Anne's predictions in the right panel.

Both Bob's actual conditional posteriors and Anne's beliefs about it display a familiar inverse S-shape: individuals overestimate small probabilities and underestimate large ones, compressing beliefs toward 0.5 (Benjamin, 2019; Enke and Graeber, 2023). This compression is more pronounced in Anne's beliefs about Bob's conditional posteriors than in Bob's actual conditional posteriors. An OLS regression of conditional posteriors on the Bayesian benchmark (with robust standard errors) estimates a substantially flatter relationship in the right panel than in the left: in the right panel, the constant is 0.36 ($se = 0.02$) and the slope is 0.39 ($se = 0.03$), whereas in the left panel, the constant is 0.25 ($se = 0.01$) and the slope is 0.59 ($se = 0.02$).

What about the corner beliefs? Does Anne think that a Bob with degenerate beliefs would revise them in response to new information, or—consistent with the Bayesian benchmark—does she view such corner beliefs as set in stone? Figure 4 plots the CDFs of Anne's beliefs about Bob's posteriors after a confirming signal (red solid) and a contradicting signal (red dashed). It also shows Bob's actual posteriors in the same cases (black solid for confirming, black dotted for contradicting). For Bob's beliefs, we pool both corners and normalize all corner priors to 100. A

Figure 3: Bob's conditional posteriors and Anne's beliefs about it



Notes: The left panel plots Bob's actual conditional posteriors, which are also Anne's own beliefs from T0 and part 1 from T1 and T2. The right panel plots Anne's beliefs about Bob's conditional posteriors (data from part 2 of T1). Both panels depict conditional posteriors as a function of Bayesian posteriors. The whiskers are 95% confidence intervals, where standard errors are clustered at the individual level. In both panels, we exclude degenerate corner priors.

confirming signal is the more likely signal conditional on the state being one, while a contradicting signal is the less likely one.[26] For Anne's beliefs, we observe only the case where Bob's prior is 100 (see Section 3).

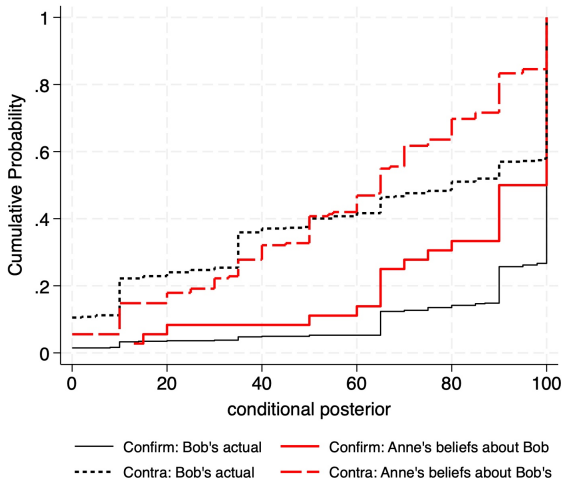
Focusing first on Bob's own corner beliefs, Figure 4 shows that when Bob receives a confirming signal, his posterior typically remains at the corner: the median posterior is 100 and the average is 92. While a fraction of subjects report posteriors below 100, this might reflect measurement noise or reporting constraints near the boundary (since beliefs cannot exceed 100). Importantly, posteriors in this case remain significantly higher than those observed for near-corner priors (90–99%) under confirming signals ($p < 0.001$). In contrast, contradicting signals lead to substantial revisions. The median posterior falls to 80 and the average to 63, and these beliefs are not statistically different from those observed for priors between 90% and 99% under contradicting signals ($p = 0.152$). Thus, even when individuals report degenerate priors, they meaningfully update their beliefs when faced with conflicting evidence.

A similar pattern holds for Anne's beliefs about Bob: she expects substantial revisions in response to contradicting signals when Bob initially holds a corner belief. Taken together, these findings indicate that corner beliefs are not as rigid as might be previously thought.[27]

[26] For example, if a participant assigns probability 100% to the statement being correct and receives a 90%-accurate signal, a positive signal is confirming, whereas a negative signal is contradicting.

[27] Additional evidence comes from Part 3 of the experiment, where participants chose between a safe \$10 payment and a risky bet paying \$11 if their reported corner belief was correct. Roughly three-quarters of participants who

Figure 4: Do corner beliefs respond to information?



Notes: We plot the CDFs of normalized Bob's conditional posteriors and Anne's beliefs about Bob's conditional posteriors. Normalization is defined as follows: beliefs equal to 0 are transformed to 100− belief, while beliefs equal to 100 remain unchanged. A confirming signal is the signal that is more likely given the reported prior that the statement is true, while a contradicting signal is the less likely signal given that prior.

Observation 4: Anne correctly predicts the direction in which Bob updates his beliefs in response to signals, but she underestimates the magnitude of this updating: her predicted conditional posteriors are flatter than the Bayesian benchmark and flatter than Bob's actual conditional posteriors. In addition, Anne does not view Bob's corner beliefs as fixed: she expects Bob to revise them in response to contradictory signals.

6 Structural Analysis

In this section, we analyze the data through the lens of two behavioral models: the model of Grether (1980), which accounts for deviations of conditional posteriors from Bayesian predictions (Benjamin, 2019) and the model of social exchange by Yuksel and Oprea (2022), which accounts for revising beliefs based on observation of beliefs of others.

reported a corner belief chose the risky option, consistent with genuinely holding degenerate priors. Taking the bet is optimal only with high confidence in the belief. Among participants whose final surprise round involved a statement for which they had reported a corner belief and then received a signal, those receiving a confirming signal chose the risky bet more than 80% of the time, compared to about 65% among those receiving a contradicting signal.

6.1 Conditional Posteriors, Structural Estimations

Grether (1980) model proposes a parsimonious way to modify Bayes's rule which allows accommodation of over- and under-inferences from either or both the prior and the signals.[28] This model is parameterized by parameters (c, d) which captures the degree to which updating deviates from the Bayesian one. Specifically, Anne's posterior given signal $s = 1$ can be written as

$$b_{s=1} = \frac{b_0^c \theta^d}{b_0^c \theta^d + (1 - b_0)^c (1 - \theta)^d}$$

The model collapses to the Bayes's rule when $c = d = 1$. Otherwise, the parameter c controls the weight on the prior, and the parameter d controls the weight on the new information. Both parameters matter in determining how sensitive Anne's posterior is to her initial beliefs and newly received signals.

Table 4: Bob's Conditional Posteriors and Anne's beliefs about it, estimates of Grether model

	Dependent Variable = ln [Posterior odds]				
	Bob's cond posteriors		Anne's beliefs about Bob's conditional posteriors		All together
	reg (1)	reg (2)	reg (3)	reg (4)	reg (5)
ln [Prior odds]	0.55** (0.02)	0.56** (0.02)	0.31** (0.04)	0.29** (0.04)	0.29** (0.04)
ln [Likelihood ratio]	0.46** (0.02)	0.48** (0.02)	0.43*** (0.04)	0.49** (0.04)	0.42** (0.04)
ln [Prior odds] *x* Political		-0.04 (0.04)		0.45** (0.12)	
ln [Likelihood ratio] *x* Political		-0.09** (0.03)		-0.19** (0.06)	
ln [Prior odds] *x* Bob					0.26** (0.04)
ln [Likelihood ratio]*x* Bob					0.04 (0.04)
Nb obs	$n = 3534$	$n = 3534$	$n = 865$	$n = 865$	$n = 4261$
Nb participants	$i = 581$	$i = 581$	$i = 195$	$i = 195$	$i = 582$
R-squared	0.44	0.44	0.30	0.32	0.42
Data	both parts in T0 Part 1 in T1 and T2		Part 2 in T1		both parts in T0 both parts in T1 Part 1 in T2

Notes: We express Grether's formula in the log form, i.e., $\ln \frac{b_{s=1}}{1-b_{s=1}} = c \cdot \ln \frac{b_0}{1-b_0} + d \cdot \ln \frac{\theta}{1-\theta}$. This implies a linear relationship between the posterior odds, the prior odds, and the likelihood ratio. We estimate this relationship using linear regression with the standard errors clustered at the individual level. We exclude Bob's degenerate priors. Political is an indicator of three politically charged statements (statements 3, 6, and 10). Bob is an indicator of Bob's actual conditional posteriors. ** indicates significance at the 5% level.

Regression (1) in Table 4 reports estimates of parameters (c, d) for Bob's actual beliefs when he receives new information. Our results are consistent with the canonical findings in the literature for so-called balls-and-urns experiments with induced priors: both parameters c and d are significantly smaller than the Bayesian benchmark (Benjamin, 2019).[29] This means that people

[28] In Appendix 8.2, we discuss alternative structural models proposed in the literature and run the horse race between these models. We show that Grether (1980) model emerges as a clear winner among the considered alternative.

[29] Our estimates for Bob are also close to those that Benjamin (2019) obtains in his meta-analysis of incentivized experiments. His estimate for the coefficient on ln[Prior odds] is 0.43 ($se = 0.09$), and his estimate for the coefficient on ln[Likelihood ratio] is 0.38 ($se = 0.03$).

tend to under-infer from both the new information they receive and their own homegrown genuine priors. Regression (2) distinguishes between neutral and politically charged statements and shows that people put similar weight on their priors in both cases but update less in light of new evidence related to political statements.

Moving from Bob's actual beliefs to Anne's beliefs about Bob's conditional posteriors, we find that Anne believes Bob also underinfers both from new evidence and from his prior (both coefficients are less than one in regression (3)). Moreover, she expects Bob's underinference from the prior to be stronger than what it is in actuality (positive and significant interaction term in regression (5)). Finally, Anne thinks that Bob puts a significantly higher weight on his prior and a significantly lower weight on the new evidence for politically charged statements relative to the neutral ones (regression (4)). In other words, Anne believes that relative to the neutral statements, Bob's posteriors regarding political statements will be close to his priors and new information will not have much effect on these priors.

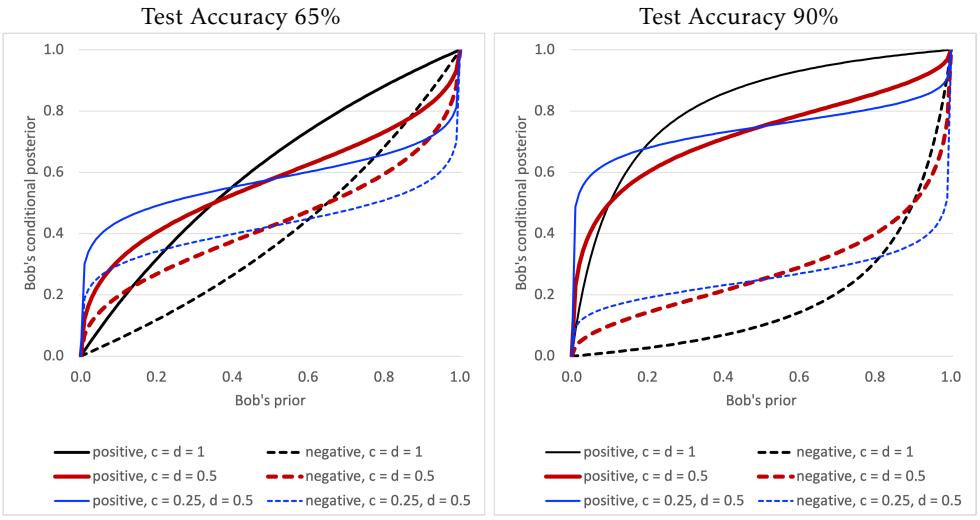
Overall, the structural estimation reported in Table 4 are consistent with the reduced-form patterns discussed in Section 5.2. Notably, because Bob's conditional posteriors coincide with Anne's own conditional posteriors, the results are consistent with Anne projecting her own updating behavior onto others, while expecting larger deviations from Bayesian predictions for them.[30]

The estimations reported in Table 4 allow us to extrapolate Anne's beliefs about Bob's conditional posteriors beyond the specific parameter values observed in the experiment and to derive general patterns in her beliefs. In particular, the two forms of underinference documented above—from the prior and from the signals—compress Anne's beliefs about Bob's conditional posteriors toward a belief of 0.5. In general, underinference from the prior alone does not necessarily lead to flatter beliefs. For example, under full base-rate neglect combined with full inference from a highly precise signal, posteriors are close to either 0 or 1. However, as we illustrate below, even moderate underinference from signals can generate flatter posteriors, even when test accuracy is 90%.

Figure 5 illustrates this point by plotting Bayesian posteriors together with posteriors predicted by the Grether model for $(c, d) = (0.5, 0.5)$ (roughly the best-fitting parameters for Bob's actual beliefs) and $(c, d) = (0.25, 0.5)$ (roughly the best-fitting parameters for Anne's beliefs about Bob's conditional posteriors). Both panels illustrate compression and flattening effects. First, the general tendency to underinfer from both new information and priors flattens conditional posteriors relative to Bayesian predictions. This substantially reduces the predicted difference in posteriors across signal realizations (compare the black lines to the red lines). Second, stronger underinference from the prior—captured by a lower parameter c—produces conditional posteriors that are even less responsive to the prior (compare the red and blue lines, holding d fixed). Together, these effects constitute the first flattening effect, which will help us understand why the quality of information has a limited impact on Anne's beliefs about Bob's expected posteriors.

[30] Loewenstein et al. (2002) show that people project their current tastes onto their future selves, while Danz et al. (2024) document projection of one's own biases onto others. Our results provide some of the first evidence that individuals also project their updating behavior onto others when interpreting new evidence.

Figure 5: Bob's conditional posteriors after receiving a signal as a function of his prior



Notes: Each panel depicts Bayesian posteriors and Grether's posteriors for two signal realizations conditional on the parameters (c, d). The x-axis on both pictures gives Bob's prior, and the y-axis gives Bob's posterior after receiving a signal. The left picture is for weak signals (accuracy 65%), while the right is for strong ones (accuracy 90%).

Observation 5: Anne's beliefs about Bob's conditional posteriors mirror the way she updates her own beliefs, except that she expects Bob to underinfer from his prior more strongly than she does herself—and more strongly than he actually does. For political statements, Anne believes that new information has very limited effect on Bob's prior beliefs. Underinference from both the prior and new signals leads Anne to believe that Bob's conditional posteriors are only weakly responsive to his prior, a phenomenon we refer to as the first flattening effect.

6.2 Signal Frequencies, Structural Estimations

We next turn to the signal frequencies and examine two behavioral models. The first one is Grether (1980)'s model which does a good job at tracking Bob's conditional posteriors and Anne's beliefs about it as we have shown in Section 6.1. Applying this model to signal frequencies requires some normalization to guarantee that both frequencies are bounded between zero and one and sum up to one. Incorporating these restrictions, we arrive at

$$\Pr[s=1] = \frac{a_0^c \theta^d + (1-a_0)^c (1-\theta)^d}{a_0^c \theta^d + (1-a_0)^c (1-\theta)^d + a_0^c (1-\theta)^d + (1-a_0)^c \theta^d} \tag{3}$$

A crucial feature of both the Grether and the Bayesian models is that Bob's prior does not play a role in determining signal frequencies; the latter is solely based on Anne's prior and signal accuracy.

The second model is social exchange model by Yuksel and Oprea (2022). In this model, Anne, upon observing that Bob holds a different prior from her own, takes this into account and revises

her prior to $\tilde{a}$. Specifically, Anne takes Bob's prior b_0 at 'face value' and treats it as an additional signal about the state: Anne considers b_0 to be generated with probability b_0 if the statement is true, and with probability $1 - b_0$ if the statement is false. Thus, we can express the revised prior odds ratio as

$$\log \frac{\tilde{a}}{1-\tilde{a}} = \alpha \cdot \log \frac{a_0}{1-a_o} + \gamma \cdot \log \frac{b_0}{1-b_0} \tag{4}$$

Parameters (α, γ) govern the weight that Anne puts on her prior relative to Bob's prior, and can be estimated from the collected data. If Anne was fully Bayesian, then $\alpha = 1$ and $\gamma = 0$ indicating that Anne is fully confident in her prior and learns nothing from Bob's prior. If, on the contrary, $\gamma > 0$ then Anne adjusts her prior after observing Bob's prior. After forming new prior $\tilde{a_0}$, Anne formulates signal frequencies as suggested by the Bayesian model using the revised prior $\tilde{a}$ instead of her original prior a_0.

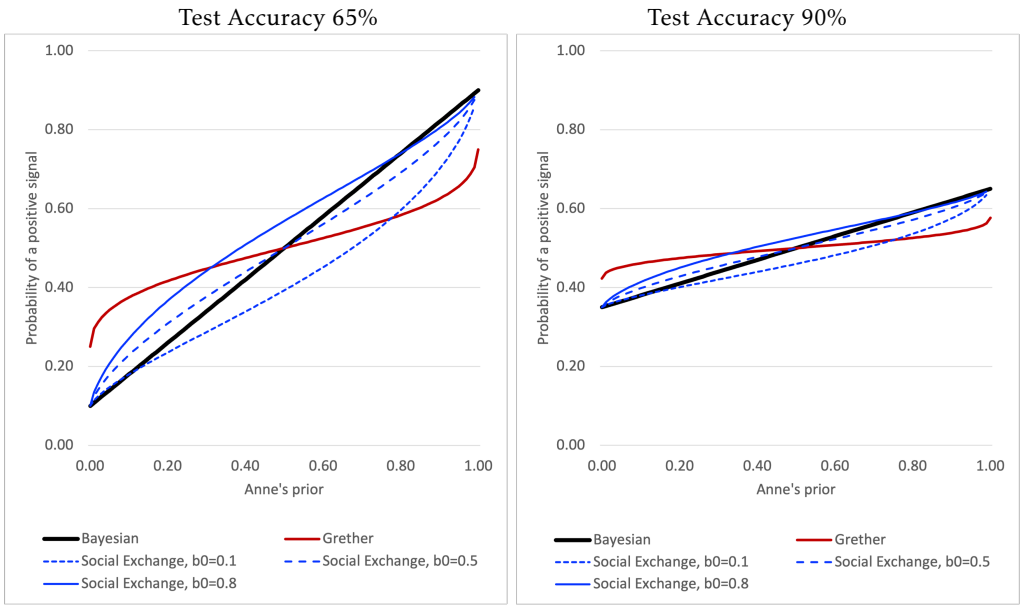
While the Grether and the Social Exchange models are obviously different, they share two properties. First, both flatten signal frequencies with respect to Anne's own prior relative to the steepness embedded in the Bayesian benchmark. Second, for a fixed Anne's prior, both reduce the difference in signal frequencies across more and less precise signals. Figure 6 demonstrates this point by plotting Anne's estimation of the probability that Bob will receive a positive signal as a function of Anne's prior. The left panel focuses on signals with low precision, $\theta = 0.65$, and the right one on signals with high precision, $\theta = 0.9$. In both panels, the black lines depict the Bayesian benchmark, the red lines are for the Grether model, and the blue lines are for the Social Exchange model.[31]

The two effects described above and depicted on Figure 6 are important for the following reason. In the Bayesian world, the substantial difference in Bob's expected posteriors for signals of different quality is driven by the significant difference in the *likelihood of receiving positive and negative signals* in two information structures. This is captured by a large difference in the slopes of the two black lines across panels in Figure 6. Contrary to that, in Grether model, the difference between the two red lines is significantly smaller and not very responsive to Anne's prior. This means that Anne expects Bob to receive signals with similar likelihoods regardless of whether he is exposed to a high- or a low-accuracy information structure and regardless of her own prior. The same conclusion follows from examining the Social Exchange model (the blue lines). This constitutes the *the second flattening effect*.

Observation 6: Anne expects signal frequencies to be less responsive to her own prior and the quality of information Bob consumes relative to what Bayesian model predicts. This conclusion holds regardless of the behavioral model Anne uses to formulate signal frequencies.

[31] A simplified version of the social–exchange model in which Anne updates her prior by taking a weighted average of her own prior and Bob's prior yields qualitatively similar predictions.

Figure 6: Signal Frequencies



Notes: For each behavioral model, we plot the probability that Anne assigns to Bob receiving a positive signal (y-axis) as a function of Anne's own prior (x-axis). For Grether's model, we use $c = d = 0.5$. For the social exchange model, we use weight $\alpha = 0.75$ on Anne's own prior and weight $\gamma = 0.25$ on Bob's prior. We compute Anne's beliefs about the likelihood of Bob receiving a positive signal given these weights and plot three lines: the solid line is for Bob's high prior $b_0 = 0.8$, the dashed line is for Bob's intermediate prior $b_0 = 0.5$, and the dotted line is for Bob's low prior $b_0 = 0.1$.

6.3 Bringing All Pieces Together

In Section 4.2, we have documented partial support for the IVP property. Consistent with this property, Anne predicts that, in general, information will bring Bob's expected posterior closer to her own prior. However, in contrast with this property, Anne predicts that these posterior moves are similar for information sources with different accuracy. Both the reduced-form and structural analysis of Anne's own updating process (which is how Bob actually updates his beliefs) and Anne's beliefs about Bob's updating process presented in the previous sections helps us understand why this is the case. We identified two "flattening effects" that jointly reduce the disparity in average posteriors predicted for signals of varying strength. The first flattening effect reflects Bob's conditional posteriors being only weakly responsive to his prior, while the second captures the reduced responsiveness of signal frequencies to both signal accuracy and Anne's prior beliefs. Together, these effects result in Bob's expected posteriors remaining relatively stagnant, showing limited responsiveness to the quality of information he encounters.

We finish this section by estimating the magnitudes of the two flattening effects to measure their relative importance in determining weak precision effect. We do that through the prism of Grether's model.

Table 5: Decomposition of the Combined Flattening Effect

	Conditional Posteriors	Signal frequency	Model Fit β_0	β_1	root MSE
(1)	Bayesian	Bayesian	0.28** (0.01)	0.55** (0.01)	0.2522
(2)	Bayesian	Grether	0.19** (0.01)	0.73** (0.02)	0.2553
(3)	Grether	Bayesian	0.07** (0.01)	0.96** (0.02)	0.2440
(4)	Grether	Grether	0.08** (0.01)	0.94** (0.02)	0.2461

Notes: We use all the data from all treatments and modify corner priors from 100% and 0% to 99% and 1%, respectively. The results are similar when we exclude corner priors (see Table 2 in Online Appendix).

Table 5 performs a decomposition exercise and turns on/off the two flattening effects one at a time. The first row is the Bayesian benchmark, where both the signal frequencies and Bob's conditional posteriors are assumed to be Bayesian. This model is a good benchmark but does not fully account for behavioral patterns observed in our experiments. Allowing either signal frequencies or conditional posteriors to follow Grether's model improves the fit significantly, with the latter modification outperforming the former one. The last row is Grether's model, where both elements follow Grether's logic. The message from this table is clear: the lack of sensitivity in Bob's average posteriors to information quality is predominantly driven by the lack of sensitivity in Bob's conditional posteriors. In fact, the model in which Anne uses Bayesian signal frequencies performs just as well as the one in which she augments signal frequencies through the lens of Grether's model.

Observation 7: Grether's model offers a parsimonious framework to account for the lack of responsiveness in Anne's beliefs about Bob's average posteriors to information quality. This non-responsiveness

is primarily driven by the lack of sensitivity in Bob's conditional posteriors to the information quality he is exposed to.

7 Conclusions

This paper provides empirical evidence on how people think others revise their beliefs in response to new information. Our findings show that individuals generally believe others' beliefs follow the Martingale property—i.e., from an ex-ante perspective, new information cannot systematically shift beliefs in one direction. However, we find only partial support for the Information Validates the Prior (IVP) property. Specifically, while people do expect new information to bring others' beliefs closer to their own effectively reducing polarization of opinions, the degree of this adjustment is less sensitive to information quality than predicted by the Bayesian model. This reduced sensitivity stems from flatter-than-expected conditional posteriors and signal frequencies. Moreover, we observe that even extreme or "corner" beliefs are not entirely degenerate, as individuals are open to revising them, and they believe others will do the same when confronted with contradictory evidence.

Our findings carry important implications for various strategic environments. The rigidity of others' beliefs and their limited responsiveness to information quality can be both advantageous and disadvantageous, depending on the setting. From a policy standpoint, a lack of responsiveness to high-quality information is often problematic, as information campaigns are designed to shift public beliefs, influence subsequent actions, and regulate markets. However, in certain scenarios, this reduced sensitivity may prove beneficial. To illustrate, consider the voluntary testing game, in which an agent with private knowledge about their ability or product quality can choose to undergo a costly test that generates an independent public signal of quality. The agent's payoff is based on the market's posterior belief of their quality minus the cost of testing. Kartik et al. (2021) theoretically demonstrate that, under standard informational assumptions, more informative tests lead to lower participation rates. However, our results suggest that participation will be less responsive to test quality, which, in this case, might be a welfare-improving outcome.

Our findings have also implications for information design literature. When people anticipate others to be relatively unresponsive to the quality of information, it may be more effective to expose them to a sequence of weak signals than a single strong signal, even if the collection of weak signals in theory conveys the same amount of information as a strong signal alone. We are hoping future research will provide empirical evidence on response to these types of information framings.

References

Agranov, M., Dasgupta, U., and Shotter, A. (2024). Trust me: Competition and communication in a psychological game. *Journal of the European Economic Association*.

Agranov, M. and Reshidi, P. (2024). Disentangling suboptimal updating: Task difficulty, structure, and sequencing. *working paper*.

Aina, C., Amelio, A., and BrÃ¼tt, K. (2023). Contingent belief updating. ECONtribute Discussion Papers Series 263, University of Bonn and University of Cologne, Germany.

Andreoni, J. and Mylovanov, T. (2012). Diverging opinions. *American Economic Journal: Microeconomics*, page 209–232.

Augenblick, N., Lazarus, E., and Thaler, M. (2024). Overinference from weak signals and underinference from strong signals. *Quarterly Journal of Economics*, forthcoming.

Azzimonti, M. and Fernandes, M. (2023). Social media networks, fake news, and polarization. *European Journal of Political Economy*, 76.

Ba, C., Bohren, A., and Imas, A. (2023). Over- and underreaction to information. *working paper*.

Becker, G., DeGroot, M., and Marschak, J. (1964). Measuring utility by a single response sequential method. *Behavioral Science*, 9:226–232.

Benjamin, D. J. (2019). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics: Applications and Foundations*, 1:69–186.

Bikhchandani, S., Hirshleifer, D., Tamuz, O., and Welch, I. (2024). Information cascades and social learning. *Journal of Economic Literature*.

Bursztyn, L. and Yang, D. Y. (2022). Misperceptions About Others. *Annual Review of Economics*, 14(1):425–452.

Calford, E. and Chakraborty, A. (2023). Higher-order beliefs in a sequential social dilemma. *Working Paper*.

Carlsson, H. and van Damme, E. (1993). Global games and equilibrium selection. *Econometrica*, 61(5):989–1018.

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

Charness, G., Gneezy, U., and Rasocha, V. (2014). Experimental methods: Eliciting beliefs. *Journal of Economic Behavior and Organization*, 189:234–256.

Danz, D., Madarasz, K., and Wang, S. (2024). The biases of others: Projection equilibrium in an agency setting. *working paper*.

Danz, D., Vesterlund, L., and Wilson, A. (2021). Belief elicitation and behavioral incentive compatibility. *American Economic Review*, 112(9):2851–2883.

DellaVigna, S. and Kaplan, E. (2007). The fox news effect: media bias and voting. *Quarterly Journal of Economics*, 122(3):1187–1234.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior*, 30:163–182.

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *Quarterly Journal of Economics*.

Esponda, I., Vespa, E., and Yuksel, S. (2023). Mental models and learning: The case of base-rate neglect. *American Economic Review*.

Evdokimov, P. and Garfagnini, U. (2022). Higher-order learning. *Experimental Economics*, pages 1234–1266.

Fedyk, A. (2024). Asymmetric naivete: Beliefs about self-control. *Management Science*, forthcoming.

Francetich, A. and Kreps, D. (2014). Bayesian inference does not lead you astray... on average. *Economics Letters*, 125:444–446.

Friedenberg, A. and Kneeland, T. (2024). Beyond reasoning about rationality: Evidence of strategic reasoning. *working paper*.

Garrett, R. (2009). Echo chambers online? politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication*, 14(2):265–285.

Gneezy, U., Enke, B., Hall, B., Martin, D., Nelidov, V., Offerman, T., and van de Ven, J. (2023). Cognitive biases: Mistakes or missing stakes? *Review of Economics Studies*.

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

Healy, P. and Leo, G. (2024). Belief elicitation: a user's guide. *Handbook of Experimental Economics Methodology*.

Healy, P. J. (2024). Epistemic experiments: Utilities, beliefs, and irrational play. *working paper*.

Kartik, N., Lee, F. X., and Suen, W. (2021). Information validates the prior: A theorem on bayesian updating and applications. *American Economic Review: Insights*, 3(2):165–182.

Kneeland, T. (2015). Identifying higher-order rationality. *Econometrica*, 83:2065–2079.

Loewenstein, G., O'Donoghue, T., and Rabin, M. (2002). Projection bias in predicting future utility. *Quarterly Journal of Economics*.

Madarasz, K. (2016). Projection equilibrium: Definition and applications to social investment, communication and trade. *working paper*.

Manski, C. and Neri, C. (2013). First- and second-order subjective expectations in strategic decision-making: Experimental evidence. *Games and Economic Behavior*, 81:232–254.

Martin, G. and Yurukoglu, A. (2017). Bias in cable news: persuasion and polarization. *American Economic Review*, 107:2565–2599.

McCarthy, N. (2019). Polarization: what everyone needs to know. *Oxford University Press*.

McCarthy, N., Poole, K., and Rosenthal, H. (2006). Polarized america: the dance of ideology and unequal riches. *MIT Press*.

McGranaghan, C., O'Donoghue, T., Nielsen, K., Somerville, J., and Sprenger, C. (2024). Distinguishing common ratio preferences from common ratio effects using paired valuation tasks. *American Economic Review*.

Morris, S. and Shin, S. (2002). Social value of public information. *American Economic Review*, 92(5):1521–1534.

Möbius, M. M., Niederle, M., Niehaus, P., and Rosenblat, T. S. (2022). Managing Self-Confidence: Theory and Experimental Evidence. *Management Science*, 68(11):7793–7817.

Pronin, E., Gilovich, T., and Ross, L. (2004). Objectivity in the Eye of the Beholder: Divergent Perceptions of Bias in Self Versus Others. *Psychological Review*, 111(3):781–799.

Pronin, E., Lin, D. Y., and Ross, L. (2002). The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality and Social Psychology Bulletin*, 28(3):369–381.

Schlag, K., Tremewan, J., and van der Weele, J. (2015). A penny for your thoughts: a survey of methods for eliciting beliefs. *Experimental Economics*, 18:457–490.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87(3):355–374.

Stroud, N. (2010). Polarization and partisan selective exposure. *Journal of Communication*, 60(3):556–576.

Szkup, M. and Trevino, I. (2020). Sentiments, strategic uncertainty, and information structures in coordination games. *Games and Economic Behavior*, 124:534–553.

Thaler, M. (2024). The fake news effect: Experimentally identifying motivated reasoning using trust in news. *American Economic Journal: Microeconomics*, pages 1–38.

Thaler, M. (2025). The Supply of Motivated Beliefs. *SSRN Electronic Journal*.

Trevino, I. and Schotter, A. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, 6:103–128.

Trujano-Ochoa, D. (2024). Do others learn like me? higher order willingness to pay for information. *working paper*.

Wang, Q. and Jeon, H. J. (2020). Bias in bias recognition: People view others but not themselves as biased by preexisting beliefs and social stigmas. *PLOS ONE*, 15(10):e0240232.

Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, 12:579–601.

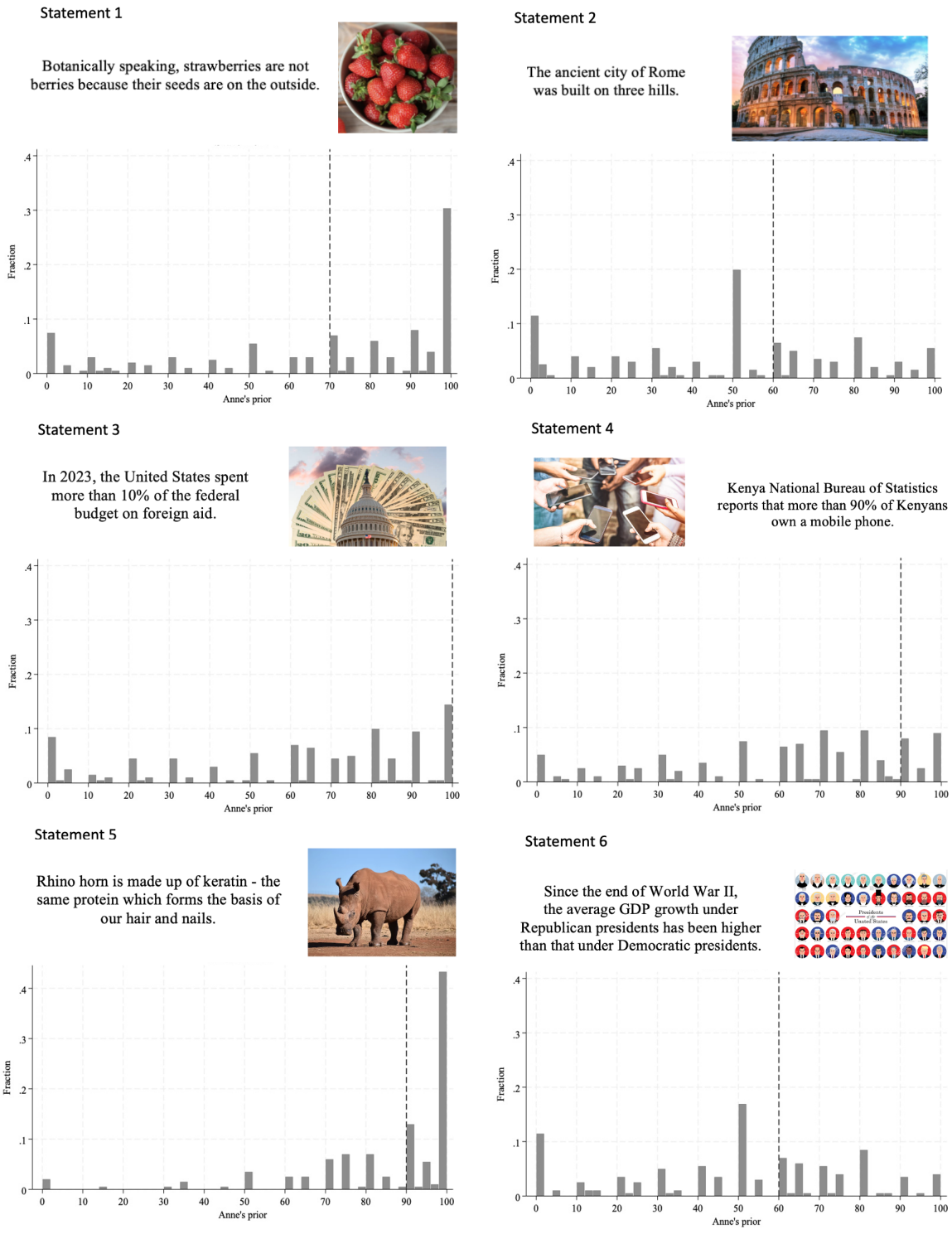
Yuksel, S. and Oprea, R. (2022). Social exchange of motivated beliefs. *Journal of the European Economic Association*, pages 667–699.

8 Appendix

In this section, we present additional data analysis, which is referenced in the paper and discuss alternative structural models one can use to analyze the data.

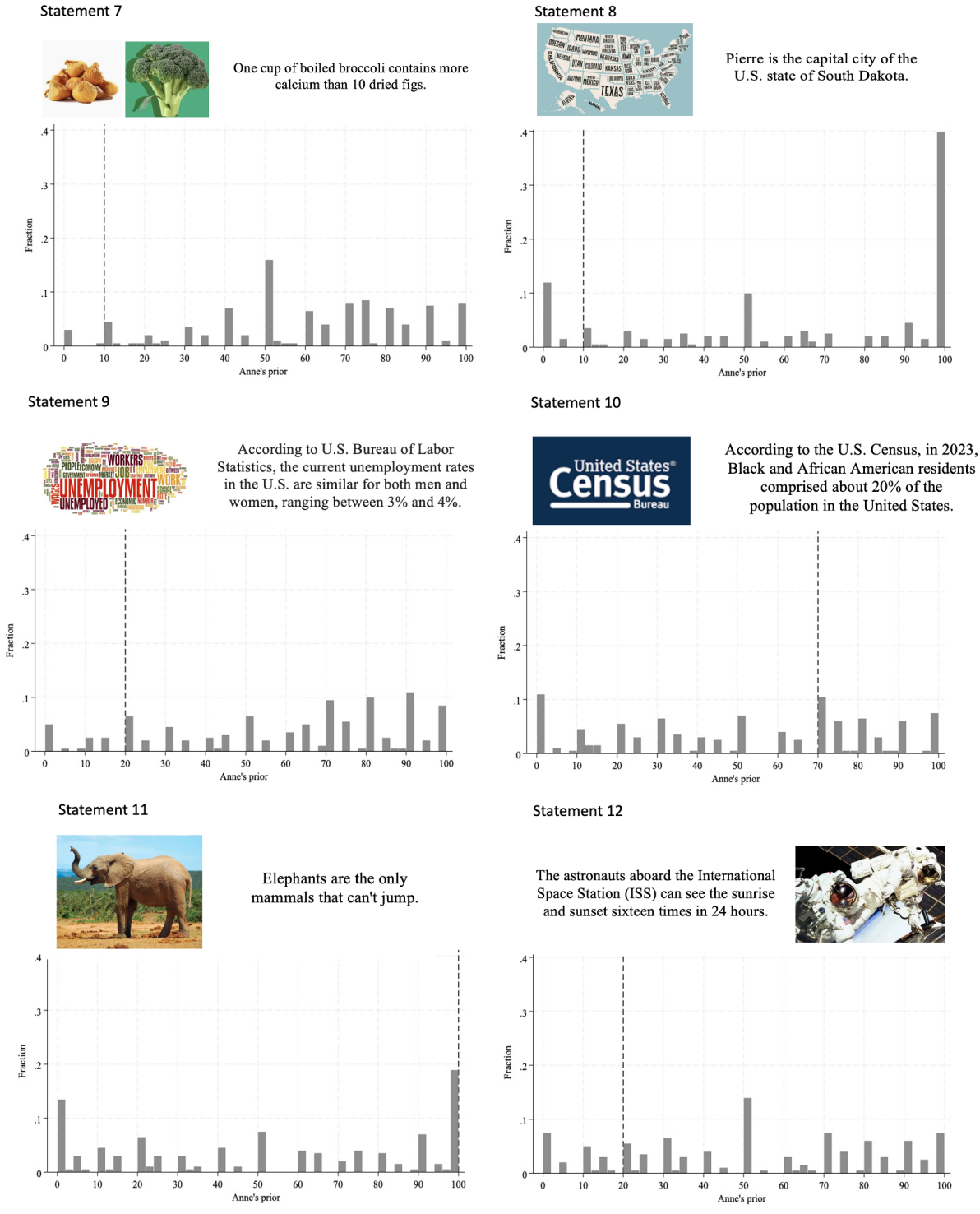
8.1 Additional Analysis

Figure 7: Statements and Anne's Prior Beliefs (treatment T0, part 1)



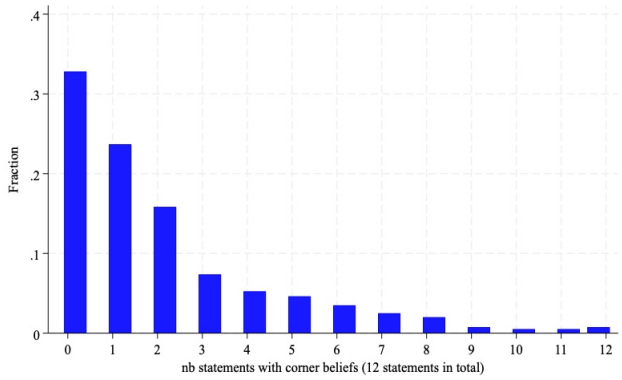
Notes: We present all statements used in the experiment and Anne's prior beliefs for each statement. The dashed line indicates the value of the prior used in Part 2 of T1 and T2, i.e., Bob's prior.

Figure 8: Statements and Anne's Prior Beliefs (treatment T0, part 2)



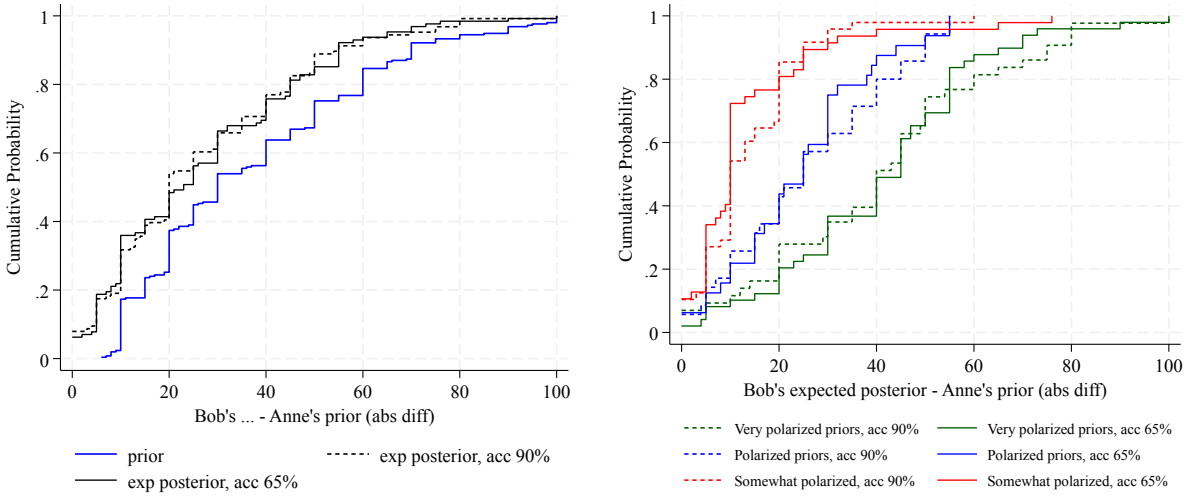
Notes: We present all statements used in the experiment and Anne's prior beliefs for each statement. The dashed line indicates the value of the prior used in Part 2 of T1 and T2, i.e., Bob's prior.

Figure 9: How Often Anne Reports Corner Beliefs in All Rounds?



Notes: We present the histogram of the number of statements in which corner belief is reported at the individual level. Data is from both parts in treatment T0.

Figure 10: Changes in Bob's beliefs when Anne and Bob have different priors for politically-sensitive statements (statements 3, 6, and 10)



Notes: The left panel depicts the CDFs of the absolute differences between Bob's and Anne's priors, as well as the absolute differences between Bob's expected posteriors and Anne's priors. The right panel displays the differences between Bob's expected posteriors and Anne's priors, broken down by each level of prior disagreement. The analysis in both panels focuses on cases where Anne and Bob have different priors.

8.2 Alternative Structural Models

In this section, we estimate beliefs' revisions through alternative structural models proposed in the literature and run the horse race between these models. We do this separately for Anne's own beliefs and Anne's beliefs about Bob's conditional posteriors. We consider four alternative models:

1. BAYESIAN model, according to which the posterior-odds ratio depends on signal precision θ and Anne's prior a_0, i.e.,
$$\frac{a_{s=1}}{1-a_{s=1}} = \frac{a_0}{1-a_0} \cdot \frac{\theta}{1-\theta}$$

2. BASE-RATE NEGLECT (BRN) model, according to which Anne completely ignores her prior and, as a result, the posterior-odds ratio depends only on the signal-odds ratio, i.e.,
$$\frac{a_{s=1}}{1-a_{s=1}} = \frac{\theta}{1-\theta}$$

3. COGNITIVE IMPRECISION model of Woodford (2020), according to which Anne misperceives signal strength but otherwise uses the Bayes' rule. In particular, we follow Augenblick et al. (2024) paper and define the true signal-odds ratio as $\mathbb{S} = \log\left(\frac{\theta}{1-\theta}\right)$ and perceived signal-odds ratio as $\mathbb{E}\left(\hat{\mathbb{S}}\right) = k \cdot \mathbb{S}^{\beta}$. Then, the difference between posterior-odds and prior-odds ratios in log terms can be written as
$$\log\left(\frac{a_{s=1}}{1-a_{s=1}}\right) - \log\left(\frac{a_0}{1-a_0}\right) = \log(k) + \beta \cdot \log\left(\frac{\theta}{1-\theta}\right).$$
Using this formulation, we can estimate the two parameters of this model (k, β).

4. GRETHER model used in the paper, according to which the posterior-odds ratio in log terms can be written as
$$\log\left(\frac{a_{s=1}}{1-a_{s=1}}\right) = d \cdot \log\left(\frac{\theta}{1-\theta}\right) + c \cdot \log\left(\frac{a_0}{1-a_0}\right)$$
and we estimate the two parameters of this model, (c, d), which represent how Anne under-/over- infers from her own prior and from the new signal she receives.

To judge which behavioral model fits our data best, we run a simple linear regression of observed posteriors on the predicted ones, clustering observations at the individual level:

$$\text{Observed Posterior} = \text{const} + \text{intercept} \cdot \text{Predicted Posterior} + \epsilon.$$

The best fit is achieved when the estimated constant is close to zero, the estimated intercept is close to one, and the value of the mean-squared errors is small.

Table 6 presents the results and shows that Grether's model emerges as a clear winner among the considered alternatives. This model captures most variation in Anne's own posteriors as well

as Anne's beliefs about Bob's conditional posteriors and significantly improves the fit relative to the Bayesian model, the BRN model, and the cognitive imprecision model.

Table 6: Comparing the fit of different behavioral models

	BRN	BAYESIAN	COGNITIVE IMPRECISION	GRETHER
Anne's own posteriors				
const	0.33** (0.01)	0.27** (0.01)	0.21** (0.01)	0.12** (0.01)
intercept	0.48** (0.02)	0.55** (0.02)	0.63** (0.02)	0.84** (0.02)
root MSE	0.2514	0.2239	0.2312	0.2217
Bob's conditional posteriors				
const	0.36** (0.02)	0.35** (0.02)	0.34** (0.02)	0.16** (0.03)
intercept	0.40** (0.04)	0.44** (0.03)	0.42** (0.03)	0.80** (0.05)
root MSE	0.2470	0.2341	0.2481	0.2320

Notes: For Anne's own posteriors, we use data from both parts in T0 and part 1 in T1 and T2. For Anne's beliefs about Bob's conditional posteriors, we use the data from Part 2 in T1. In all estimations, we exclude corner priors and corner posteriors. This is done to maintain with results reported in Table 4 and general comparability across models. This is because Grether's and Woodford's models involve logs of prior-odds and posterior-odds ratios and, as a result, are not defined for corner priors and posteriors.

As a final exercise, we take the cognitive imprecision model and the estimated parameters (k, β) obtained by Augenblick et al. (2024) and ask what would these estimates predict in our experiment. Augenblick et al. (2024) obtains $k = 0.88$ and $\beta = 0.76$ which imply very close to Bayesian posteriors for low-precision signals (65% accuracy) and significant underinference relative to Bayesian posteriors for high-precision signals (90% accuracy). These predictions do not fit our data as Figure 3 clearly shows.

Our discussion above supports the use of the Grether model for analyzing revisions of beliefs in a structural manner.