How We Think Others Update Beliefs: An Experiment*

Marina Agranov† and Polina Detkova‡

November 6, 2025

Abstract

We study how people think *others* update their beliefs upon encountering new evidence. We find support for 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 also find that when the two have different priors, people think that any information brings others' expected posteriors closer to their own prior, but this adjustment is less responsive to information quality than theory predicts. We identify the primary cause of this insensitivity and discuss the implications of our findings for strategic games with asymmetric information, information design, and, more broadly, for understanding societal polarization.

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

*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.

exploit the advantages of controlled laboratory experiments to provide some of the first empirical 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. We refer to the former scenario as *Bob's conditional posterior* and the latter as *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

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 why Anne's beliefs about Bob's expected posteriors are less responsive to the information quality than expected, we study three elements that jointly determine expected posteriors. The first element is how Anne updates her own beliefs. In line with previous literature, we find that Anne tends to underinfer both from her prior and from new evidence. This underinference results in less responsive (flatter) than Bayesian posteriors and it is stronger for politically charged statements.[5] Moreover, we present a novel finding indicating that corner beliefs (i.e., a prior of 0% or 100%) are not as rigid and degenerate, as previously thought; Anne is willing to revise these beliefs when confronted with contradictory evidence.

The second element concerns Anne's beliefs about Bob's conditional posteriors—how she ex-

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. (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]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).

pects Bob to update his prior when she observes the signal he receives. Our findings reveal that Anne tends to project her own belief-updating process onto Bob. She believes that, similar to her, Bob underinfers both from his prior and from new evidence. 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. Notably, Anne believes that Bob underweights his prior more than she does when updating her own beliefs; that is, she thinks Bob relies less on his prior than she relies on hers. For political statements, Anne expects that new information will have little effect on Bob's priors. Similarly to her own corner priors, Anne thinks that Bob's corner priors can shift when faced with contradictory evidence. Overall, these patterns lead to Bob's conditional posteriors being flatter than Bayesian ones, showing 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.

We also compare Anne's predictions of Bob's conditional posteriors to our behavioral benchmark—Bob's actual elicited posteriors. The results indicate that Anne is generally accurate in anticipating how information alters others' beliefs. Her largest errors stem from systematically overestimating the extent to which Bob underweights his prior when updating, but overall, her predictions exhibit a remarkable degree of accuracy.

The third element shaping Bob's expected posterior is the signal distribution, which determines how Bob's conditional posteriors are weighted in the expected value. Compared to the Bayesian benchmark, Anne expects the signal frequencies to be less responsive to both her own prior and the quality of the information Bob receives. In other words, she predicts signal frequencies that are closer to a uniform distribution than the Bayesian calculation implies. This constitutes the second "flattening" effect.

The two "flattening" effects operate in the same direction, collectively making Bob's expected posteriors quite rigid and less responsive to the quality of information he consumes relative to what Bayesian theory predicts. This finding is important as it underscores the limited impact of information in altering perceptions of others' beliefs, and consequently influencing their 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.[6] 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

[6]We explore several prominent models from the literature and find that the model proposed by Grether (1980) offers the best fit to our data, achieving this with the fewest parameters compared to alternatives. Other models we estimate include the social exchange model of Yuksel and Oprea (2022), Woodford (2020)'s model of cognitive imprecision used in the recent paper by Augenblick et al. (2024), and the base-rate neglect model (see chapter 6 in Benjamin (2019) survey for the evidence and theoretical underpinning of this phenomenon). Notably, the social exchange model of Yuksel and Oprea (2022) allows Anne to update her belief upon learning Bob's prior. This possibility, however, does not alter the main results of the paper. We discuss these models and the estimations in Section 5.3 and in Appendix in Section 7.1.

observed patterns and allows us to conduct counterfactual exercises, which often require extrapolation beyond the parameters directly observed in the experimental data.

In the first structural 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.

In the second structural exercise, we assess the magnitude of Anne's mistakes in predicting Bob's expected posteriors, comparing them to our behavioral benchmark, i.e., the *actual average Bob's posteriors*. By and large, the mistakes are quite small, indicating that Anne is remarkably good at predicting Bob's average posteriors. The largest mistakes she makes pertain to situations in which her own and Bob's priors are extreme and are on different sides of the spectrum. In these cases, Anne overestimates how much information will shift Bob's opinions toward her own. This result stems from Anne predicting that Bob underinfers from his prior to a larger extent than he actually does. This highlights the challenge of predicting how others respond to new information, particularly when strong and polarized opinions are involved.

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, this could help bridge the gap between polarized groups and reduce division, especially if 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, and people generally recognize its limited potential to reduce polarization.

1.1 Connection to the Literature

Information design. As discussed in the introduction, our experiment builds on the findings of Kartik et al. (2021), which contributes to the extensive theoretical literature on information design. While we do not aim to provide a comprehensive review of this literature, we want to highlight a few studies that explore strategic communication in environments with heterogeneous priors. For example, Hirsch (2016) examines a model in which a principal and an agent share common goals but hold heterogeneous prior beliefs about which policy is most effective. This disagreement complicates the agent's motivation but can be alleviated through policy ex-

perimentation and observing outcomes. Alonso and Camara (2016) considers a setting where the sender and receiver have differing prior beliefs, and the sender designs an experiment to persuade the receiver. The authors characterize the set of posterior belief distributions that can be induced by such experiments in setup with flexible information structure, i.e., no standard ordering assumptions, and identify necessary and sufficient conditions under which persuasion benefits the sender. Che and Kartik (2009) investigates a game in which a decision-maker consults an adviser before making a decision. The adviser can exert costly effort to obtain a signal about the state and communicate this information to the decision-maker. Although both parties care about the state, their differing prior beliefs create a tension: these differences incentivize information acquisition but simultaneously lead to information loss through strategic communication. In all these papers, agents must form beliefs about how others, who may hold different priors, update their beliefs when new evidence arrives—a process which we investigate experimentally in our study.

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.

Higher-order beliefs. 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.[7] 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.

[7] 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.

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. Thaler (2025) elicits beliefs of senders about the motivated reasoning of receivers and demonstrates that they adjust their message accordingly to these beliefs.

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.

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 playfield to document how people think others revise their beliefs in response to new information. It makes our approach similar to that of Evdokimov and Garfagnini (2022), who 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 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 the relationship between the extent to which one projects her information onto others and the extent to which one anticipates but underestimates the projection of others onto her as predicted by the behavioral model of Madarasz (2016). The results show strong

support for the projection equilibrium model. Among other things, the authors document adherence of the data to the Martingale property, albeit in a very different setting. Fedyk (2024) finds that individuals exhibit substantial sophistication regarding the present bias of others, while being mostly naive about their own. The psychology literature calls this phenomenon the "bias blind spot" (Wang and Jeon, 2020; Pronin et al., 2002, 2004). We show that individuals exhibit sophistication regarding base-rate neglect in others' belief updating while engaging in base-rate neglect themselves. However, they overestimate the extent to which others are susceptible to the base-rate neglect, connecting our findings to the broader literature on misperceptions about others (see Bursztyn and Yang, 2022, for a review). Trujano-Ochoa (2024) explores to what extent people consider the biases of others when updating their beliefs and on information acquisition patterns.[8]

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 reduced-form evidence on Martingale and IVP properties. Section 5 utilizes all conducted treatments to unpack the aggregate results presented in the previous section. Section 6 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 5.3, 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).[9]

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., $\theta = \Pr[s = \omega|\omega]$.

Anne knows both Bob's prior and signal accuracy, and she is tasked with predicting Bob's

[8] This work is still in progress. The preliminary draft we have access to suggests that, on average, people expect others to update in a similar manner to themselves. However, they show a significantly lower willingness to pay when others' strategies are implemented on their behalf. This latter finding is consistent with an expectation that others are more conservative in updating than they are themselves.

[9] 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.

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 $\mathbb{E}[b]$. If Anne is Bayesian and expects Bob to be Bayesian, then

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

where

$$b_{s=1} = \frac{b_0\theta}{b_0\theta + (1-b_0)(1-\theta)} \quad , \quad b_{s=0} = \frac{b_0(1-\theta)}{b_0(1-\theta) + (1-b_0)\theta} \, ,$$

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

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 frequencies. 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 belief would be, 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, i.e., $1 > \theta_1 > \theta_2 > \frac{1}{2}$. 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, i.e., 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[10]:

- 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 from T0 and observe the past participants' prior for the same statement and signal accuracy.

[10] Figures 13 and 14 in the Appendix present all statements used in the experiment and the visualization used alongside the statements.

Participants in T1 also observe the signal realization received by the past participants, while in T2, no such information is provided. In both treatments, after observing the information, participants are asked to guess the posterior reported by these past participants.[11]

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 (see an example in the Online Appendix).

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.[12]

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, $b_0 \in \{0.10, 0.20, 0.60, 0, 70, 0.90, 1.00\}$.[13] For each prior b_0, we selected two 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.[14]

Subject pool. The experiments were conducted on the Prolific platform in January 2024 with roughly 200 participants in each treatment, for a total of 603 participants. We recruited partici-

[11] We refer the reader to the Online Appendix for the screenshots detailing the language used to explain these tasks to participants.

[12] 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.

[13] Figures 13 and 14 in Appendix present the distribution of priors elicited from participants in T0 and indicate which prior was used as past participants' priors in T1 and T2.

[14] 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.

pants 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.[15] 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.[16] In addition, in 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).[17] The experimental software was programmed in Qualtrics. Instructions and screenshots of the interface are presented in the Online Appendix. Table 1 summarizes the details of all three treatments.

Table 1: Design

Treatment		Part 1 own beliefs 6 rounds	Part 2 others' beliefs 6 rounds	Part 3 corner beliefs 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

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

[16] 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 the expected payoff in the experiment. This technique became standard in the literature as it helps participants to understand payment method and, as a result, helps the experimenter to elicit participants' true beliefs.

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

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).[18]

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.[19]

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.

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

[18] 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.

[19] 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.

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 (see Figures 13 and 14); and (b) fewer than 20% of participants report corner beliefs for more than a quarter of the statements (see Figure 1 in the Online Appendix). 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 Results

We start with presenting reduced-form evidence on Martingale and IVP properties (Sections 4.1 and 4.2), and illustrating the difference between them (Section 4.3). In this analysis, we use the data from T2 treatment, in which Anne predicts Bob's expected posterior. In Section 5 we explore what drives these aggregate results by studying how Anne updates her own beliefs, how Anne thinks Bob updates his beliefs when she knows his signal realization, and what this means for Anne's beliefs about signal distribution.

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, in which we cluster standard errors by individuals to account for the inter-dependency of observations that come from the same participant.[20]

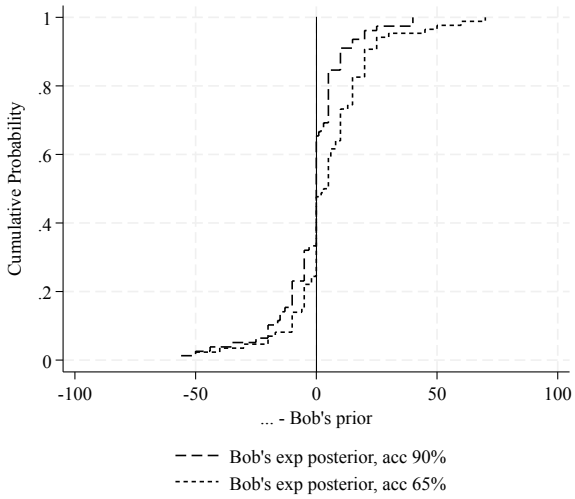
4.1 Martingale property

Does Anne expect information to systematically alter Bob's posterior when they share prior? Figure 1 displays the CDFs of the differences between Anne's prediction of Bob's expected posterior and Bob's original prior.

Figure 1 shows that the two CDFs are approximately symmetric around zero. The mean difference is not significantly different from zero when test accuracy is 65% ($p = 0.51$). It becomes statistically significant at 90% accuracy, but the magnitude remains small—about 4 percentage points ($p = 0.05$). Moreover, there is no significant difference between the two signal structures

[20] To be precise, we regress the variable of interest (for instance, 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 (for instance, the test accuracy), while clustering standard errors at the individual level. We say that the two treatments are significantly different when the estimated treatment indicator is different from zero at the standard 5% level and report the p-value associated with the estimated indicator.

Figure 1: Changes in Bob's beliefs when Anne and Bob share the same prior



Notes: The figure depicts the CDFs of the differences between Bob's expected posterior and his prior for each signal structures when Anne and Bob have same priors (priors differ by no more than 5 pp). The data come from Part 2 of treatment T2.

($p = 0.14$). 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

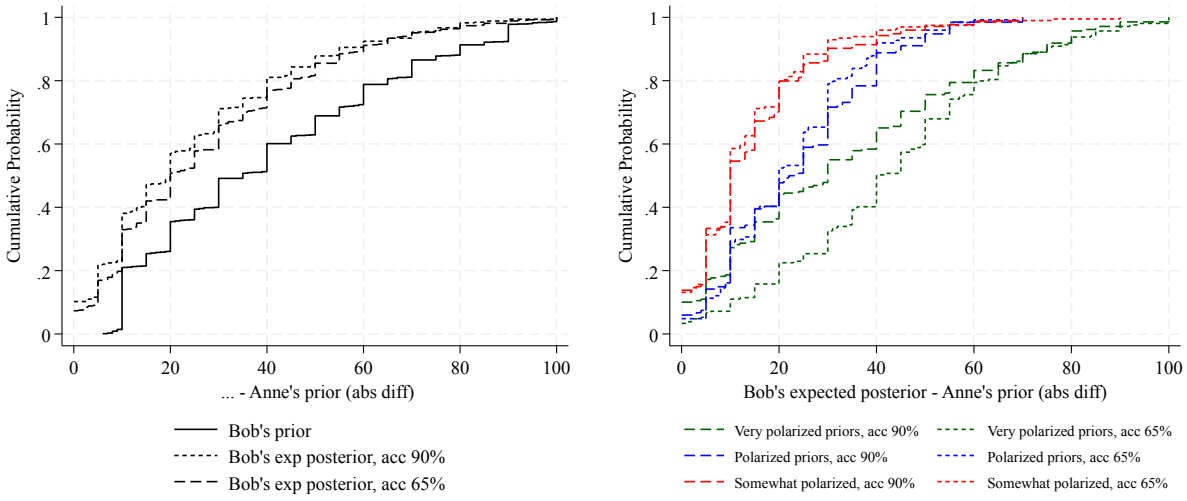
What does Anne think about Bob's expected posterior when they have different priors? The IVP property has two empirical footprints. First, Anne expects any information to be effective at bringing Bob's posterior closer to her prior relative to the original disagreement in their priors. Second, the more 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. Table 2 and Figure 2 depict the basic statistics and 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 (left panel of Figure 2). This shift is large in magnitude and statistically significant ($p < 0.01$).[21]

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%.

[21] 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: Changes in Bob's beliefs when Anne and Bob have different priors



Notes: The left panel depicts the CDFs of the absolute differences between Bob's prior and Bob's expected posterior and Anne's priors. The right panel displays the differences between Bob's expected posterior and Anne's prior, broken down by levels of priors' disagreement. We focus on cases where Anne and Bob have different priors. The data is from Part 2 of treatment T2.

However, the difference between the two CDFs is rather small and only marginally significant ($p = 0.08$).

The right panel in Figure 2 and the data in Table 2 show 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 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. At the same time, contrary to the IVP property, when Anne's and Bob's priors differ by less than 40 pp, we observe no difference between the two information structures in general (the two blue and the two red lines in the right panel of Figure 2 are very similar).[22] Overall, when Anne's and Bob's opinions are not too different, Anne believes that Bob's posteriors will shift similarly regardless of whether he learns from a more precise or a less precise source.

Anne's beliefs about changes in average Bob's beliefs for politically charged statements are similar to those for non-politically charged statements but with an even greater disregard for the quality of information compared to neutral statements. Figure 15 in Appendix replicates

[22]Table 1 in Online Appendix, presents similar statistics as the bottom part of Table 2 for the case in which Anne's and Bob's priors differ by at most 40 pp. The data shows that when Anne's beliefs are extreme, the more precise information structure shifts Bob's expected posterior closer to Anne's prior, consistent with the IVP property. However, the opposite is true when Anne's priors are intermediate or close to uniform; in these cases, Anne thinks that the less precise signals are more effective at reducing the polarization of opinions between Anne and Bob.

Table 2: Differences in Anne's and Bob's beliefs before and after Bob consumes new evidence, in absolute terms.

Bob's prior is different from Anne's prior (at least 5 pp difference)

	Anne's prior							
	all		extreme		intermediate		close to uniform	
	mean (se)	med	mean (se)	med	mean (se)	med	mean (se)	med
before info	39.8 (1.0)	35	44.4 (1.5)	40	35.6 (1.1)	30	30.6 (1.2)	35
info acc 90%	24.8 (1.1)	20	24.5 (1.6)	15	25.0 (1.5)	20	24.6 (1.6)	30
info acc 65%	27.3 (1.0)	20	31.1 (1.5)	25	24.0 (1.5)	20	20.7 (1.9)	20
p-values								
before vs 90%	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p = 0.018$	
before vs 65%	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
90% vs 65%	$p = 0.079$		$p = 0.002$		$p = 0.440$		$p = 0.098$	

Anne's and Bob's priors are very polarized (more than 40 pp difference)

	Anne's prior							
	all		extreme		intermediate		close to uniform	
	mean (se)	med	mean (se)	med	mean (se)	med	mean (se)	med
before info	67.9 (0.8)	65	75.0 (0.9)	70	55.9 (0.8)	55	49.6 (0.7)	50
info acc 90%	34.0 (2.0)	30	36.1 (2.7)	30	28.5 (2.8)	25	n/a	n/a
info acc 65%	42.4 (1.8)	40	47.2 (2.2)	49	37.0 (2.4)	40	29.5 (4.6)	33
p-value								
before vs 90%	$p < 0.001$		$p < 0.001$		$p < 0.001$		n/a	
before vs 65%	$p < 0.001$		$p < 0.001$		$p < 0.001$		$p < 0.001$	
90% vs 65%	$p = 0.001$		$p = 0.001$		$p = 0.024$		n/a	

Notes: Each table reports the difference between Anne's and Bob's prior beliefs in the first row and the differences between Anne's beliefs about Bob's expected posterior and her own prior in the second and third rows. The second and third rows differ by signal accuracy. We focus exclusively on cases where Anne and Bob have different priors. Entries marked n/a indicate instances with fewer than 10 observations. Anne's prior is categorized into three groups: extreme priors (below 20 or above 80), close-to-uniform priors (between 40 and 60), and intermediate priors (the remaining category, i.e., priors between 20 and 40 or between 60 and 80).

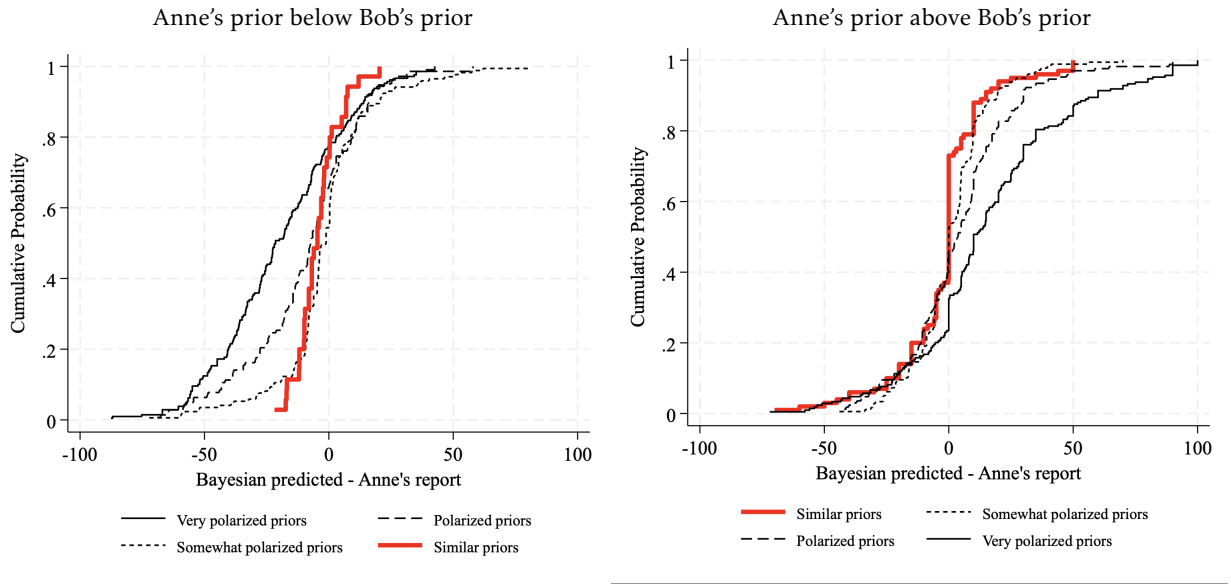
Figure 2 for two politically charged statements used in our experiment.[23] It illustrates that Anne believes Bob's average posterior beliefs will still move closer to her own prior after receiving new information. Yet, unlike neutral statements, these shifts are identical regardless of the quality of the information Bob receives.

How do Anne's estimates about shifts in Bob's posteriors compare to those predicted by Bayesian theory? Figure 3 combines the data from both information structures and depicts the difference between predicted and observed posteriors depending on whether Anne's prior is above or below that of Bob's and how different the two priors are (black solid and dashed lines).[24] As a reference, we also plot the case in which Anne and Bob share the same prior (red lines).

[23] We have two politically charged statements: statement 6 about the estimates of GDP growth under Democratic vs Republican presidents and statement 3 about the United States foreign aid spendings.

[24] Figure 2 in Online Appendix presents Figure 3 separately for each information structure and shows very similar patterns.

Figure 3: Anne's estimates of Bob's expected posteriors vs Bayesian predictions



Notes: We plot the CDFs of the differences between Bob's Bayesian-predicted posterior expectations and Anne's estimates of these values. The plots are separated into cases where Anne's prior is lower than Bob's (left panel) and cases where Anne's prior is higher than Bob's (right panel). The data is sourced from Part 2 of treatment T2.

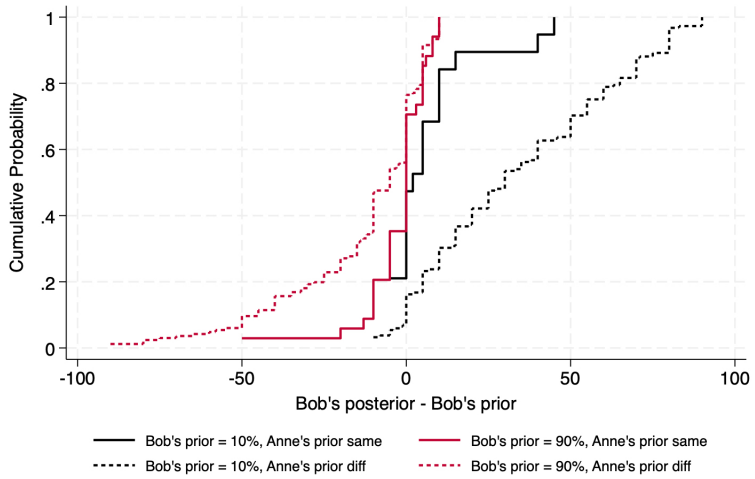
As we've discussed above, when Anne shares the same prior as Bob, she expects Bob's posterior to be on average the same as Bob's or her own prior. In Figure 3 this is depicted by red lines being very close to zero and actually being exactly equal to zero for a large portion of the data. However, when Anne and Bob do not share the same priors, Anne tends to underestimate how Bob's posteriors move relative to those predicted by Bayesian theory. Indeed, when Anne's prior is above that of Bob, the majority of Anne's estimates are below predicted ones, while the opposite happens when Anne's prior is below that of Bob. Figure 12 in Appendix replicates Figure 3 for Anne's extreme beliefs, for which we find the strongest qualitative support of the IVP, and document similar patterns. Overall, Anne expects that Bob's reaction to information is much more inert than what the Bayesian theory predicts.

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, the significant difference in the effectiveness of more precise information structures is observed only when Anne has very different beliefs from Bob, when Anne holds relatively extreme prior beliefs herself, and when the statements are neutral. Otherwise, Anne expects Bob's beliefs to shift similarly regardless of information quality. In other words, Anne thinks that Bob's average posteriors are not responsive to information quality, contrary to what the Bayesian theory predicts.

4.3 Difference between Martingale and IVP properties

We conclude this section by illustrating the distinction between the Martingale property and the IVP property. Figure 4 plots the CDFs of the difference between Bob's posterior and his prior when Anne and Bob share similar priors versus when their priors differ. The illustration focuses on two priors of Bob—10% and 90%.[25]

Figure 4: Anne's expectations regarding changes in Bob's beliefs as a function of the difference in their original opinions



Notes: The figure depicts the CDFs of the differences between Anne's prediction of Bob's expected posterior and Bob's prior, separately for cases in which Anne and Bob share similar priors (solid lines) and cases in which their priors differ (dashed lines). Results are shown for Bob's prior of 10% (black) and 90% (red). Data from two information structures is pooled together and is based on Part 2 of treatment T2.

Figure 4 rules out the possibility that, when Anne and Bob begin with different beliefs, Anne expects Bob's posterior to remain close to his prior. Instead, it reinforces both the Martingale property and the core component of the IVP property. When Anne and Bob hold the same prior, information has little effect on Bob's posterior (solid lines), consistent with the Martingale prediction. By contrast, when their priors differ, information moves Bob's posterior away from his original prior (dashed lines) and toward Anne's prior, consistent with first part of the IVP prediction.[26]

[25] Other priors exhibit similar patterns; these graphs are available from the authors upon request.

[26] Figure 11 in the Appendix presents analogous CDFs of the *absolute* differences between Bob's expected posterior and his prior for the same-prior and different-prior cases. The CDFs for the same-prior case shows that the difference between Bob's expected posterior and his prior is small in most cases: less than 15 pp in nearly 80% of observations, regardless of test accuracy. The CDFs for the different-prior case first-order stochastically dominates that for the same-prior case, with the difference statistically significant at the 1% level for each signal structure.

5 Unpacking Aggregate Results

In this section, we study what drives aggregate results presented in Section 4. We start by documenting how Anne revises her own beliefs in response to new information using the data collected in treatment T0 and in Part 1 of treatments T1 and T2 (Section 5.1). We then investigate how Anne thinks Bob forms his conditional posteriors when new information arrives using the data collected in Part 2 of treatment T1 (Section 5.2). After that, in Section 5.3, we study how Anne predicts the signal frequencies based on what we learned about Anne's updating process in Section 5.1. Section 5.4 closes the loop by bringing all these elements together and explaining why Anne's beliefs about Bob's expected posteriors are much more rigid than Bayesian theory predicts.

The analysis is structured as follows. We first present model-free raw data patterns. Then, we analyze the data through the lens of the Grether (1980) model, which is one of the most popular behavioral models used in the literature to account for deviations in belief-updating tasks (Benjamin, 2019). In our case, it turns out that the Grether model outperforms alternative behavioral models proposed in the literature. The alternative models we discuss include the social exchange model of Yuksel and Oprea (2022), the cognitive imprecision model of Woodford (2020) used in a recent paper by Augenblick et al. (2024), and the base rate neglect model. We discuss these alternatives and run a horse race between the models in Appendix 7.1.

5.1 How Anne Updates Her Beliefs

The left panel in Figure 5 depicts Anne's reported posteriors as a function of Bayesian posteriors. For both signal accuracies, we observe a familiar inverse S-shape (Benjamin, 2019; Enke and Graeber, 2023). People tend to overestimate the probabilities of unlikely events and underestimate the probabilities of very likely events.

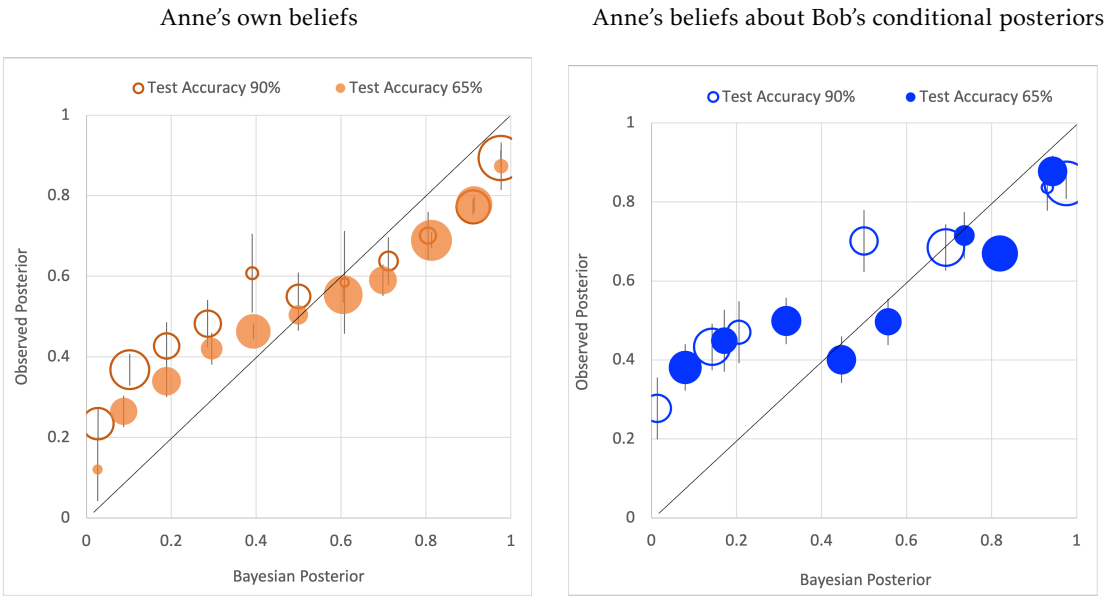
Grether (1980) model proposes a parsimonious way to modify Bayes's rule which allows accommodation of over- and under-inferences from either or both the prior and the signals. This model is parameterized by parameters (c, d) which captures the degree to which updating deviates from the Bayesian one.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

$$a_{s=1} = \frac{a_0^c \theta^d}{a_0^c \theta^d + (1 - a_0)^c (1 - \theta)^d}$$

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.

Regression (1) in Table 3 reports estimates of parameters (c, d) for Anne's own beliefs when she receives new information. Our results are consistent with the canonical findings in the literature established for the belief-updating tasks with so-called balls-and-urns experiments and

Figure 5: Observed versus Bayesian posteriors



Notes: The left panel uses data from treatment T0 as well as Part 1 from treatments T1 and T2, while the right panel uses the data from Part 2 of treatment T1. The whiskers are 95% confidence intervals, where standard errors are clustered at the individual level. In both panels, we exclude degenerate corner priors.

induced beliefs: both parameters c and d are significantly smaller than the Bayesian benchmark (Benjamin, 2019). This means that people 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.

Anne's Corner Beliefs. Corner beliefs are degenerate, and, by definition, unchangeable. If you are absolutely sure that a statement is either true or false, then no new evidence should alter your conviction, and your opinion of the statement should remain unchanged. Our experiment provides one of the first empirical evidence evaluating this prediction.[27]

How often do people report corner beliefs? That depends on the statement. The fraction of corner beliefs ranges from 11% to 49% per statement, with an average of 22%. Some participants are more likely to report the corner beliefs than others. However, as Figure 1 in Online Appendix shows, participants rarely report corner beliefs for more than 4 statements out of 12 in total.[28]

Do people take corner beliefs seriously? The data in part 3 of the experiment provides some insights into this question. Recall that in this part, we offer participants a choice between a safe

[27] Note that, by design, both signals are conceivable even when one holds a corner prior. This is true because the signals are only partially informative: conditional on the state, there is a positive chance of receiving either a signal that coincides with the state or contradicts it. Thus, a participant cannot learn from a signal that their prior is wrong.

[28] Recall, that Figures 13 and 14 present the histograms of prior beliefs for each statement observed in T0. Participants' priors in the other two treatments T1 and T2 are very similar to those in T0 and are omitted for brevity.

Table 3: Anne's Own Posteriors and Anne's Beliefs about Bob's Conditional Posteriors, estimates of Grether model

	Dependent Variable = ln [Posterior odds]					
	Anne's own beliefs		Anne's beliefs about Bob's conditional beliefs			All together
	reg (1)	reg (2)	reg (3)	reg (4)	reg (5)	reg(6)
ln [Prior odds]	0.55** (0.02)	0.55** (0.02)	0.31** (0.04)	0.27*** (0.04)	0.31** (0.04)	0.29** (0.04)
ln [Likelihood ratio]	0.46** (0.02)	0.47** (0.02)	0.43** (0.04)	0.43*** (0.04)	0.47** (0.04)	0.42** (0.04)
ln [Prior odds] *x* Political		-0.02 (0.05)			0.70** (0.32)	
ln [Likelihood ratio] *x* Political		-0.09** (0.04)			-0.17** (0.09)	
ln [Prior odds] *x* Anne						0.26** (0.04)
ln [Likelihood ratio] *x* Anne						0.04 (0.04)
ln [Prior odds] *x* Same Priors				0.38*** (0.08)		
ln [Likelihood ratio] *x* Same Priors				0.04 (0.08)		
Nb obs	$n = 3534$	$n = 3534$	$n = 865$	$n = 865$	$n = 865$	$n = 4261$
Nb participants	$i = 581$	$i = 581$	$i = 195$	$i = 195$	$i = 195$	$i = 582$
R-squared	0.4433	0.4443	0.2956	0.3116	0.3068	0.4193
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., $\ln \frac{a_{s=1}}{1-a_{s=1}} = c \cdot \ln \frac{a_0}{1-a_0} + d \cdot \ln \frac{\theta}{1-\theta}$. 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 Anne's degenerate priors in reg (1), (2), and (6) and Bob's degenerate priors in reg (3) - (6). Political is an indicator of two politically charged statements (the GPD growth and the foreign aid spending). Anne is an indicator of Anne's own posteriors. Same is an indicator that Anne's and Bob's priors are within 5 p.p. from each other. ** indicates significance at the 5% level.

payment of \$10 and a risky bet which pays \$11 if one's reported corner belief is correct and nothing otherwise. About three-quarters of participants who reported a corner belief chose the risky bet in the last part of the experiment. We take this evidence as supportive of the fact that people do originally believe in their corner priors. Taking such a risky bet makes sense only if one has little doubt about correctly assessing the truthfulness of the statement.[29]

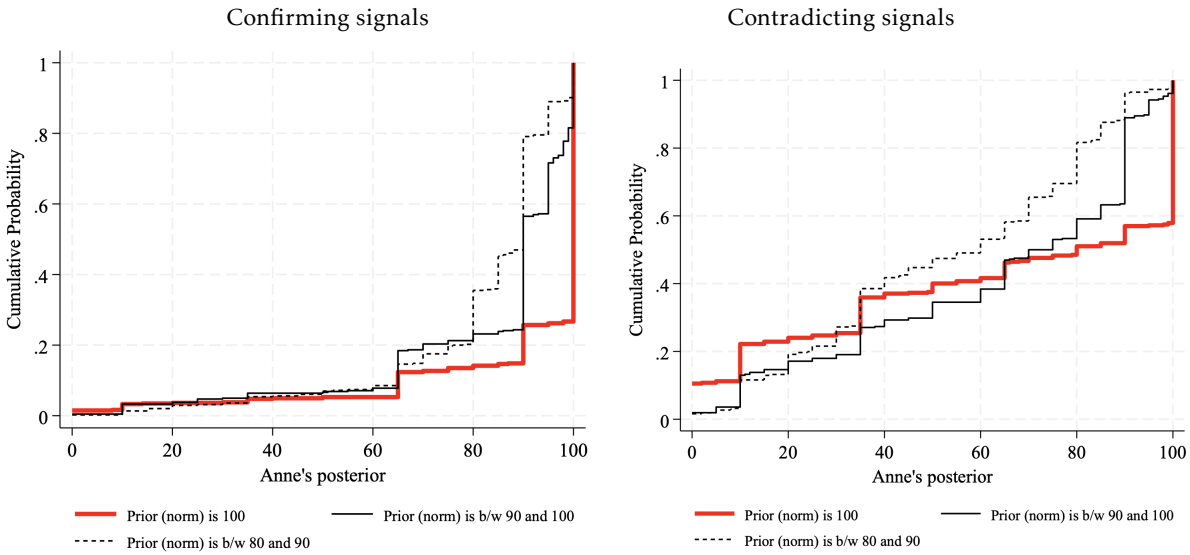
Do people update corner beliefs? Figure 6 depicts the CDFs of Anne's posteriors after receiving either a confirming original prior signal (left panel) or a contradicting original prior signal (right panel). We pool the data from both corners and redefine all corner beliefs to be one. The confirming signal is, then, a more likely signal conditional on the state being one, while the contradicting signal is the less likely signal.[30] The red thick lines depict updating for corner beliefs,

[29] Focusing on participants whose last surprise round involved the statement where they reported a corner belief and received a signal about that statement—i.e., the 'own beliefs' portion of the experiment—we find that they are more likely to choose the risky bet when the signal confirms their prior belief than when it contradicts it. Specifically, participants with confirming signals chose the risky bet over 80% of the time, compared to approximately 65% for those with contradicting signals.

[30] Say, a participant believes that the statement is correct with probability 100% and receives a signal which is 90%

while black solid and dashed lines provide a benchmark of how Anne updates her beliefs when her prior is close to the corner but still interior.

Figure 6: How Anne updates her own corner beliefs



Notes: We plot the CDFs of Anne's normalized posteriors of Anne from Part 1 in all treatments. The normalized prior equals to 100 - elicited prior for priors below 50 and equals itself for the priors above 50. The confirming signal is a more likely signal and the contradicting signal is the less likely signal conditional on the statement being true (100% correct),

When Anne receives a confirming signal, she rarely updates (left panel of Figure 6). The median posterior, in this case, is 100, the average is 92, and it is significantly different from Anne's posterior when she has a slightly lower prior between 90% and 99% and similarly receives a confirming signal ($p < 0.001$). However, when Anne receives a contradicting signal (right panel of Figure 6), she updates her beliefs substantially. The median belief in this case is 80, the average is 63, and it is not significantly different from the posterior beliefs of Anne whose prior is between 90% and 99% and who similarly receives a contradicting signal ($p = 0.152$). This evidence suggests that corner beliefs are not really corners: people are willing to change their minds in light of new evidence that goes against their prior beliefs, even when they were initially certain in their opinion.

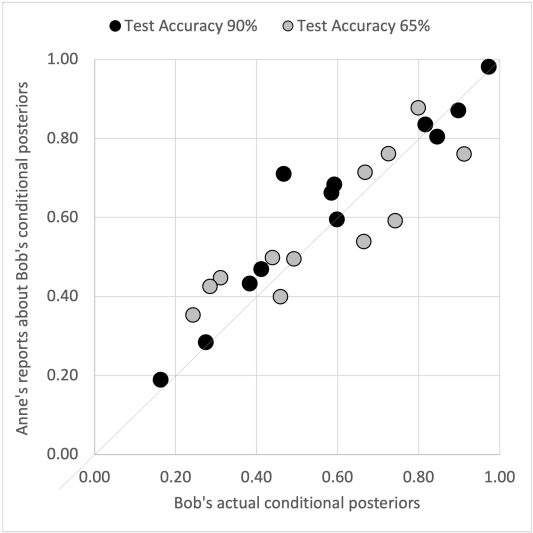
Observation 3: When updating her own beliefs, Anne underinfers both from her prior and from new information. The underinference from new information is stronger for politically charged statements. Corner beliefs are not really degenerate, they are malleable to some degree and can be updated in light of contradictory evidence.

accurate. The positive signal is a confirming signal, as it confirms the original belief of a participant. The negative signal, however, is a contradicting one, as it goes against one's prior.

5.2 How Anne Thinks Bob Updates His Beliefs

The right panel in Figure 5 depicts Anne's guesses about Bob's conditional posteriors as a function of Bayesian posteriors. For both signal accuracies, we observe a familiar inverse S-shape similar in form to Anne's own posteriors (left panel). The shape similarity between the left and the right panels in Figure 5 is consistent with Anne projecting her way of updating onto how she thinks others update, albeit larger deviations from the Bayesian predictions for Anne's own posteriors compared to Bob's posteriors.[31] Regression (3) presented in Table 3 confirms what we see in Figure 5: Anne thinks that Bob, like her, underinfers both from new evidence and from his prior, i.e., $c^{\text{Bob}} < 1$ and $d^{\text{Bob}} < 1$. Regression (6) shows that Bob's underinference from the prior is stronger than her own, i.e., $c^{\text{Bob}} < c^{\text{Anne}}$.

Figure 7: How good Anne is at predicting Bob's conditional posteriors?



Notes: We plot Anne's reports about Bob's conditional posteriors in T1 against Bob's actual conditional posteriors, which are Anne's own conditional posteriors in T0 holding fixed the priors, the signal accuracy, and signal realizations.

Figure 7 compares Anne's predictions of Bob's conditional posteriors with Bob's actual posteriors, which correspond to Anne's own posteriors holding fixed their priors, signal accuracies, and signal realizations.[32] Most of the dots lie close to the 45-degree line, indicating accurate predictions. In other words, Anne is generally quite good at predicting Bob's conditional posteriors. The biggest mistakes she makes reflect that pattern documented above: Anne thinks that Bob underinfers from his prior more than he actually does.

Four more patterns regarding Bob's conditional posteriors are worth noting. First, regression

[31] Loewenstein et al. (2002) show that people project their current tastes on their future selves. Danz et al. (2024) show evidence that people project their own biases on others. Our results are one of the first documenting that people project the way they update on how others do so when encountering new evidence.

[32] Recall that Anne predicts the conditional posteriors of past participants from treatment T0. This is why the sample sizes for the actual and the predicted Bob conditional posteriors differ for a given signal, prior, and signal precision, and this is the reason we plot a simple scatter plot rather than a bubble graph.

(5) in Table 3 distinguishes between the types of statements that Bob encounters. Interestingly, 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. 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.

Second, another manifestation of projection bias is evident in Anne's susceptibility to base-rate neglect and her predictions about Bob's likelihood of exhibiting the same bias. The base-rate neglect in its pure form suggests that a decision-maker completely ignores their prior and, for instance, after receiving a positive signal with 90% accuracy updates their posterior to 90%.[33] Using individual-level data, we find a strong and significant correlation between the number of questions where Anne exhibits perfect base-rate neglect and the number of questions where she believes Bob will do the same: corr = 0.59 ($p < 0.01$).

Third, Bayesian theory posits that Anne's beliefs about Bob's conditional posteriors are independent of her prior. Regression (4) shows that this prediction does not match our data as Anne thinks that Bob will put more faith in his prior when he shares the same prior as she does.

Fourth, Anne also projects her own way of updating corner beliefs on Bob. Anne thinks that Bob's corner beliefs are not set in stone, especially when he receives a signal that contradicts his initial prior. Figure 16 in the Appendix shows trends similar to those in Figure 6, illustrating the similarity between how Anne updates her corner beliefs and what Anne thinks about Bob's updating his corner beliefs. There is minimal updating when the signal aligns with the initial prior, but significant adjustment when the signal contradicts it.

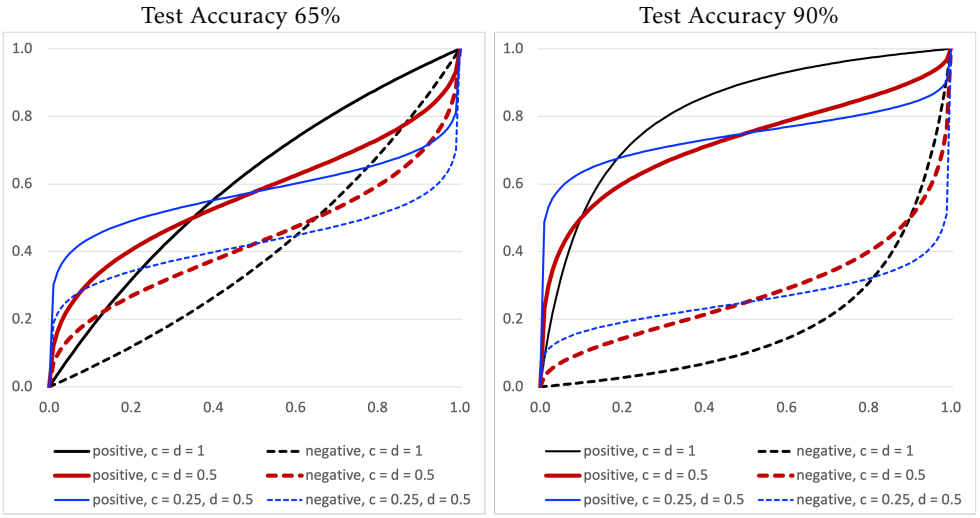
Observation 4: Anne projects the way she updates her beliefs on the way she thinks others do. Anne thinks that Bob underinfers both from his prior and from new evidence and that Bob's corner beliefs may shift, just like hers. Compared to her own updating process, Anne thinks Bob underinfers from his prior to a larger degree than she does herself. For political statements, Anne believes that new information is quite ineffective at moving Bob's beliefs, and his posteriors will not differ much from his priors. Overall, Anne is quite good at predicting Bob's conditional posteriors and she expects these posteriors to be less responsive to both his prior and information quality compared to what the Bayesian theory predicts.

The role of documented underinferences in determining the shape of Bob's conditional posteriors. The two underinferences (from the prior and from the signals) are important as they compress Bob's conditional posteriors towards a 50/50 belief and result in Anne believing that Bob's posteriors are not as responsive to Bob's priors as Bayesian theory predicts or as Anne's own beliefs respond. Figure 8 illustrates this point and plots Bayesian posteriors as well as posteriors predicted by the Grether model for $(c, d) = (0.5, 0.5)$ and $(c, d) = (0.25, 0.5)$. The latter parameters correspond roughly to those estimated in regression 6 of Table 3, which are based on Anne's

[33] Base-rate neglect is one of the most prevalent biases in decision-making, garnering considerable attention in the literature due to its persistence and widespread occurrence (Benjamin, 2019; Esponda et al., 2023; Gneezy et al., 2023).

observed beliefs about Bob's conditional posteriors. Both figures illustrate the compression and flattening effects. First, the general tendency of individuals to underinfer from new information and priors flattens conditional posteriors relative to Bayesian predictions. This substantially reduces the predicted difference in posteriors for two signal realizations (compare the black lines to the red lines). Second, stronger underinference from one's prior—represented by a lower parameter c leads to conditional posteriors that are even less responsive to the prior (compare the red and the blue lines, holding d fixed). Both effects are crucial for understanding why the quality of information has a limited impact on Anne's beliefs about Bob's expected posteriors.

Figure 8: 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%).

5.3 What Anne Thinks about Signal Distribution

The last element in the equation determining Bob's expected posteriors is signal distribution. In the Bayesian world (Section 2), the likelihood of receiving a positive signal depends linearly on Anne's prior belief and the signal accuracy:

$$\Pr[s = 1] = a_0\theta + (1 - a_0)(1 - \theta) \tag{2}$$

However, the evidence presented so far documents systematic deviations from the Bayesian model. How do these deviations affect Anne's beliefs about signal frequencies? This is what we study in this section.

Note that understanding what Anne thinks about signal frequencies is an inference exercise, since we do not directly observe Anne's beliefs about signal distribution.[34] We therefore proceed

[34]One could envision an experiment in which Anne's beliefs about signal frequencies are elicited, in addition to her

as follows. We start by formulating several alternative behavioral models that describe how Anne may form beliefs about signal frequencies. The models we consider are either based on the behavioral patterns documented above or are popular models in the literature relevant to our setting. We examine the general properties of these models and compare their predictions with those of the Bayesian model. In Section 5.4, we estimate these models using data from treatment T2 and evaluate their ability to fit the data.

The first model is that of Grether (1980). As documented in Section 5.1, Anne tends to underinfer both from her prior a_0 and from signals, the accuracy of which is depicted by parameter θ. Grether's model is summarized by two parameters (c, d) and it does a good job at tracking the deviations of Anne's beliefs from the Bayesian ones. 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 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 we consider differs from the Grether and Bayesian models in that it allows for *social exchange*. Anne, upon observing that Bob holds a different prior from her own, takes this into account and revises her prior to $\tilde{a}$. She then formulates signal frequencies as suggested by the Bayesian model in equation 2, using the revised prior $\tilde{a}$ instead of her original prior a_0.

We follow the model of social exchange by Yuksel and Oprea (2022) to define how Anne revises her prior after observing Bob's prior. This model suggests that Anne takes Bob's prior b_0 at 'face value': 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. That is, Anne believes that Bob's prior b_0 is an additional signal about the state of the world, i.e., the truthfulness of the statement, with a likelihood ratio being $\frac{b_0}{1-b_0}$. 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.

While the Grether and the Social Exchange models are obviously different, they share two

beliefs about Bob's average posteriors. However, we chose not to pursue this approach out of concern that it might be leading and could alter how people naturally think about others' average posteriors. For instance, consider someone who does not instinctively break down Bob's average posterior into signal frequencies and Bob's conditional posteriors. If asked a question that prompts this decomposition, they might learn to focus on signal frequencies as a crucial element, even though they might not have done so on their own without such a suggestion.

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 9 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.[35]

The two effects described above and depicted on Figure 9 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 9. 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).

Observation 5: Compared to the Bayesian benchmark, Anne expects signal frequencies to be less responsive to her own prior and the quality of information Bob consumes. This conclusion holds regardless of the behavioral model Anne uses to formulate signal frequencies.

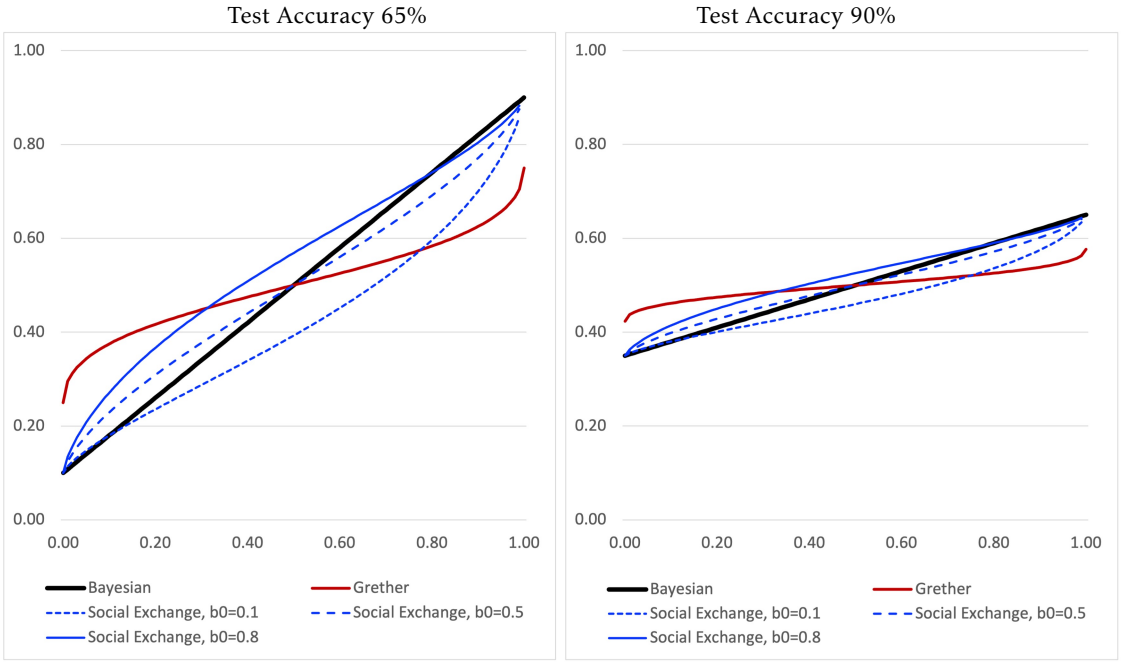
5.4 Bringing All Pieces Together

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. The analysis of Anne's own updating process and her beliefs about Bob's updating process presented in the previous sections helps us understand why this might be 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 diminished sensitivity to conditional posteriors, 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've discussed two behavioral models that can account for the effect of 'average posteriors being less responsive to information quality than predicted by the Bayesian model'. Both models use Grether's framework to calculate Anne's and Bob's posteriors conditional on signal realizations. The two models differ in how signal frequencies are computed; the fully Grether's model

[35] 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 9: 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 $b_0 = 0.1$.

uses Grether's logic to compute signal frequencies, while the Social Exchange model allows Anne to revise her prior after observing Bob's prior before formulating signal frequencies.

Table 4 estimates both models and compares their fit to the Bayesian benchmark. We perform this exercise twice: once using all observations and modifying corner priors to close but non-corner values (the top part) and once excluding the corner priors (the bottom part). The modification of corner beliefs is warranted given our analysis of how people update their corner priors, which shows that corner priors are not degenerate and can change as new evidence arrives.[36]

To compare the fit we run a simple linear regression of observed posteriors on the predicted ones, clustering standard errors at the individual level, i.e.,

$$\text{Observed Posterior} = \beta_0 + \beta_1 \cdot \text{ Predicted Posterior} + \epsilon. \tag{5}$$

The best fit is achieved when $\beta_0 = 0$ and $\beta_1 = 1$.

Table 4: Parameter Estimates and Model Fit for Behavioral Models

	Anne's parameters $(c^{\text{Anne}}, d^{\text{Anne}})$	Bob's parameters $(c^{\text{Bob}}, d^{\text{Bob}})$	Revised prior (α, γ)	Model fit β_0	β_1	root MSE
All data						
Bayesian				0.28** (0.01)	0.55** (0.01)	0.2522
Grether	(0.36,0.43)	(0.30,0.47)		0.08** (0.01)	0.94** (0.02)	0.2461
Grether + Social Exchange	(0.39,0.40)	(0.26,0.45)	(1,1)	0.14** (0.01)	0.82** (0.02)	0.2517
Without corners						
Bayesian				0.27** (0.01)	0.57** (0.01)	0.2303
Grether	(0.49,0.41)	(0.33,0.43)		0.07** (0.01)	0.96** (0.02)	0.2288
Grether + Social Exchange	(0.48,0.40)	(0.32,0.46)	(1,0.58)	0.06** (0.01)	0.97** (0.02)	0.2278

Notes: The estimates of each model are presented alongside the model fit (see equation 5). We use data from all parts of all treatments. The top part of the table uses all data including the corner priors, where corner priors of 100% and 0% are replaced by 99% and 1%, respectively. The bottom part of the table excludes these corner priors.

Two patterns emerge from Table 4. First, the estimated parameters (c, d) for Anne's own updating and Anne's beliefs about Bob's updating are similar to those presented in Sections 5.1 and 5.2, and display the same qualitative patterns. In particular, Anne thinks that Bob's under-inference from new evidence is similar to her own, i.e., $d^{\text{Anne}} = d^{\text{Bob}}$. At the same time, Anne thinks that Bob under-infers from his prior more than she does herself, i.e., $c^{\text{Anne}} > c^{\text{Bob}}$.[37]

Second, both models perform very well at explaining deviations from the Bayesian benchmark. Among the two of them, we favor Grether's model since it uses fewer parameters than the model that combines elements of Grether's model and the Social Exchange and has a similar or better fit.

[36] Similar modification is done in the analysis of Enke and Graeber (2023).

[37] We cannot reject the hypothesis that $d^{\text{Anne}} = d^{\text{Bob}}$. We obtain $p = 0.24$ ($p = 0.55$) for Grether model using all data (data without corners). We obtain $p = 0.10$ ($p = 0.11$) for Grether + Social Exchange model using all data (data without corners). At the same time, we reject the hypothesis that $c^{\text{Anne}} = c^{\text{Bob}}$ in all specifications of all models ($p < 0.01$).

Magnitudes of Two Flattening Effects. Here, we measure the relative importance of the two flattening effects. We do that through the prism of Grether's model which, as we've argued above, is an elegant and parsimonious way of organizing our data.

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 6: Grether's model offers a parsimonious explanation for the lack of responsiveness in 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.

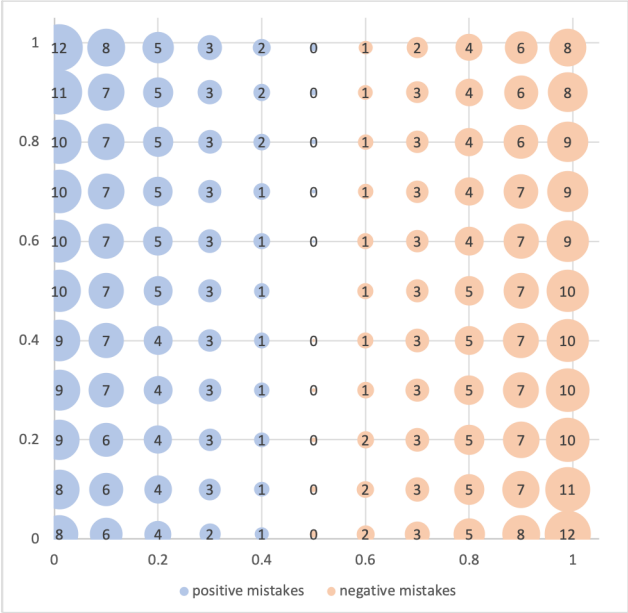
Magnitudes of Anne's Mistakes. How accurately does Anne predict Bob's expected posteriors? Figure 10 illustrates the differences between Bob's actual and Anne's predicted posteriors for signals with 90% accuracy using the estimated parameters of Grether's model reported in Table 4.[38,39] In the figure, positive mistakes are represented by blue circles, while negative mistakes are shown as orange circles. The size of each circle, along with the number inside it, indicates the magnitude of the mistake for a specific pair of Anne's and Bob's priors, with Anne's priors depicted on the horizontal axis and Bob's priors on the vertical axis. A positive mistake means Anne

[38] We use parameters reported in the second-to-last row of Table 4. The figure depicts mistakes for the following values of priors: 0.01, 0.1, 0.2, ..., 0.9, and 0.99. A similar analysis for weak signals with 65% accuracy is provided in Figure 3 in Online Appendix, showing similar results.

[39] To assess Anne's accuracy, we rely on the estimated parameters rather than the raw data, because Anne's actual prediction is influenced by her prior. Using the raw data alone would make it impossible to determine whether the inaccuracy in her prediction arises from her misunderstanding of Bob's belief updating or from an incorrect assumption about the distribution of signals Bob receives.

predicts Bob's expected posterior to be lower than it actually is. Conversely, a negative mistake indicates the opposite, i.e., Anne's prediction about Bob's expected posterior is higher than it is.

Figure 10: The difference between Bob's actual average posteriors and Anne's prediction of Bob's average posterior based on estimates of Grether's model, signal accuracy 90%



Notes: For each pair of Anne (x-axis) and Bob's (y-axis) priors, the size of the mistake is represented by the bubble size, with the exact value displayed inside the bubble. Anne's priors are shown on the horizontal axis, while Bob's priors are depicted on the vertical axis. Both priors take values of 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 0.99. The mistake estimates are derived using the parameters of Grether's model reported in Table 4 for the dataset excluding corner beliefs. Blue bubbles represent positive mistakes, where Anne overestimates Bob's expected posterior (i.e., she believes Bob's posterior is higher than it actually is). In contrast, orange bubbles indicate negative mistakes, where Anne underestimates Bob's expected posterior (i.e., she believes it is lower than it actually is).

Figure 10 demonstrates that, apart from extreme priors, Anne is fairly accurate in predicting Bob's expected posteriors. For all priors except the most extreme cases (0.01 and 0.99), the prediction errors are at most 7 percentage points, which is remarkably good. The largest mistakes occur for highly polarized priors. For example, when Anne holds a prior of 0.01 and attempts to predict Bob's expected posterior given his prior of 0.99, she overestimates the extent to which Bob's belief shifts towards hers predicting that Bob's expected posterior will be 12 percentage points lower than it actually is. Similarly, when Anne's prior is 0.99 and she predicts Bob's posterior given his prior of 0.01, Anne again overestimates the magnitude of Bob's shift toward her high prior by 12 percentage points, resulting in a negative mistake.

These errors are consistent with the estimates of Grether's model reported in Table 4. As we've shown, Anne believes that Bob places less weight on his own prior than he actually does, i.e., $c^{\text{Bob}} < c^{\text{Anne}}$. Consequently, when the two have very extreme and polarized priors, Anne thinks that Bob will on average shift further away from his prior relative to what he actually does. This

highlights the limitations of Anne's predictions in such scenarios.

Observation 7: Anne is quite accurate in predicting movements in Bob's expected posteriors. However, the largest errors arise when Anne and Bob hold highly polarized and extreme priors. In these cases, Anne overestimates the extent to which Bob's beliefs shift toward her own due to information.

6 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*.

Alonso, R. and Camara, O. (2016). Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165:672–706.

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.

Che, Y. and Kartik, N. (2009). Opinions in incentives. *Journal of Political Economy*, 117:815–860.

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*.

Hirsch, A. (2016). Experimentation and persuasion in political organizations. *American Political Science Review*, 110:68–84.

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.

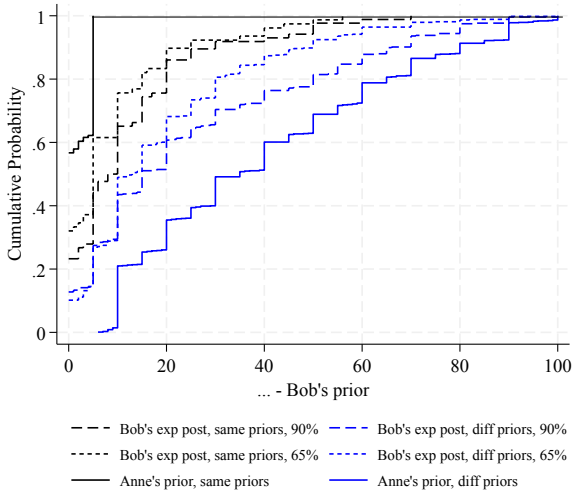
Woodford, M. (2020). Modeling imprecision in perception, valuation, and choice. *Annual Review of Economics*, 12:579–601.

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

7 Appendix

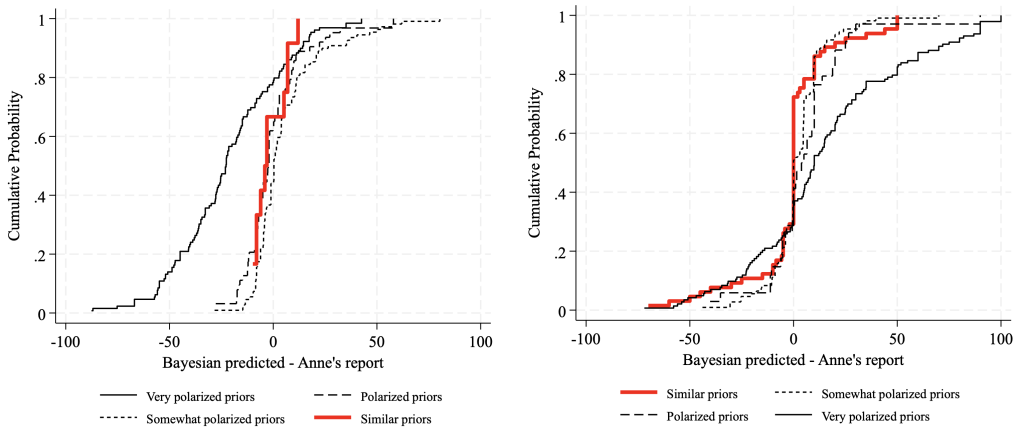
In this section, we present additional data analysis, which is referenced in the paper.

Figure 11: The CDFs of the difference in Bob's expected posterior and his prior, depending on the difference in Anne's and Bob's original priors, in absolute terms.



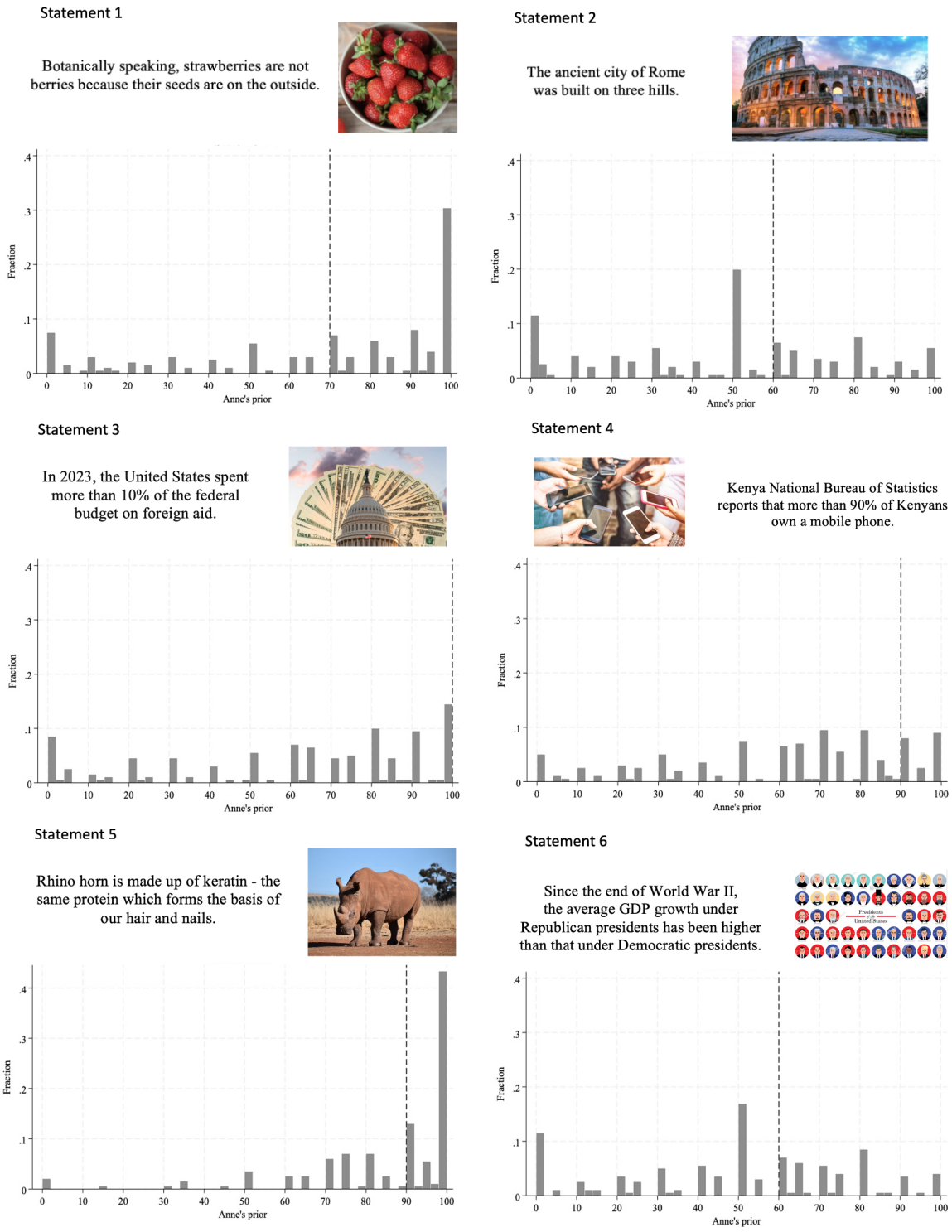
Notes: We plot the CDFs of the absolute difference between Bob's expected posterior his prior separately for the case in which Anne and Bob share the same prior and the case in which they have different priors. We also plot the CDFs of the differences between Anne's and Bob's priors in the two cases. The data is sourced from Part 2 of treatment T2.

Figure 12: Anne's estimates of Bob's expected posteriors vs Bayesian predictions when Anne's priors are extreme



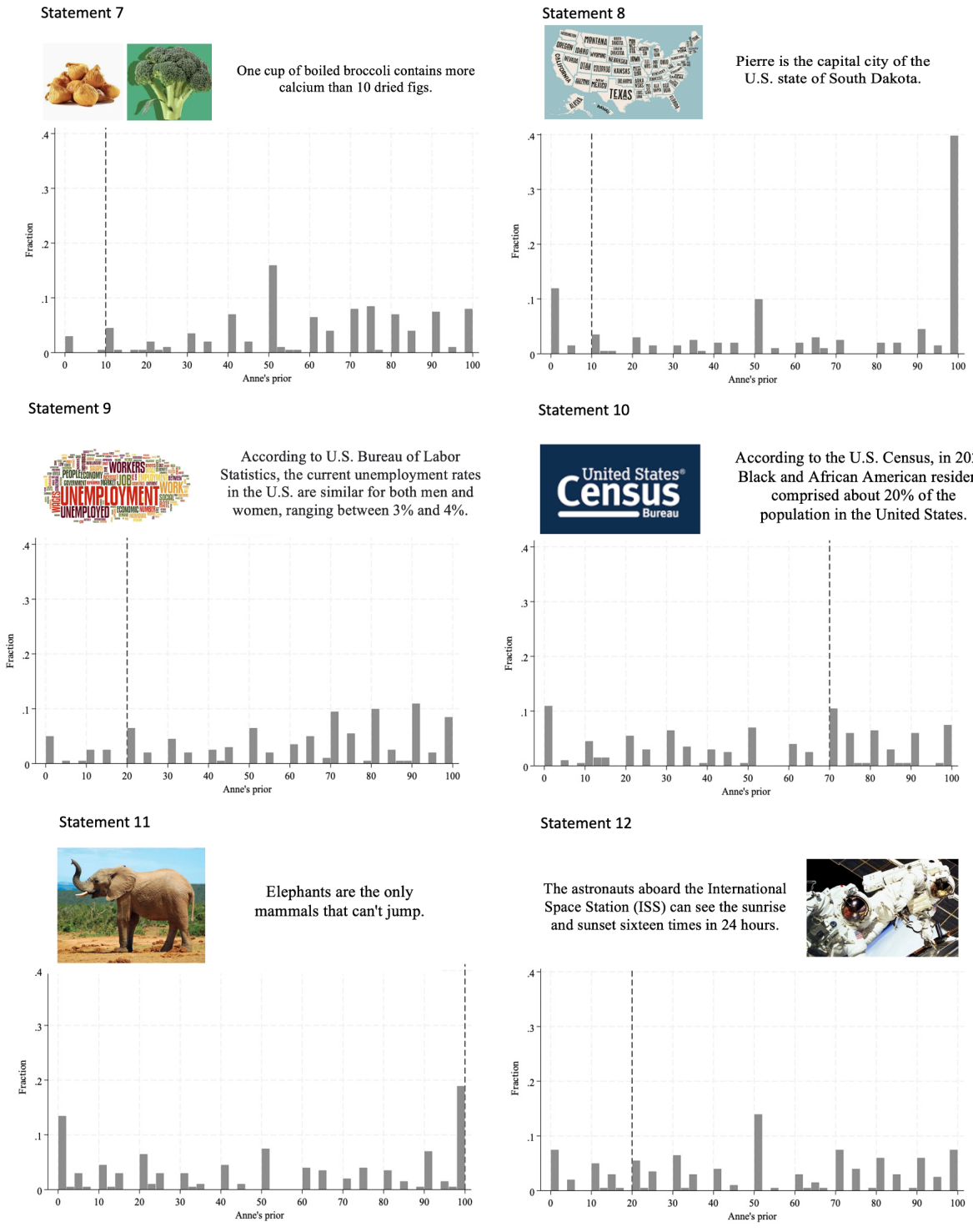
Notes: We plot the CDF of the differences between Bob's Bayesian-predicted posterior expectations and Anne's estimates of these values. We focus on cases in which Anne's prior is extreme, i.e., below 20 or above 80. The plots are separated into cases where Anne's prior is lower than Bob's (left panel) and cases where Anne's prior is higher than Bob's (right panel). The data is sourced from Part 2 of treatment T2.

Figure 13: Statements and Anne's Prior Beliefs (treatment T0, part 1)



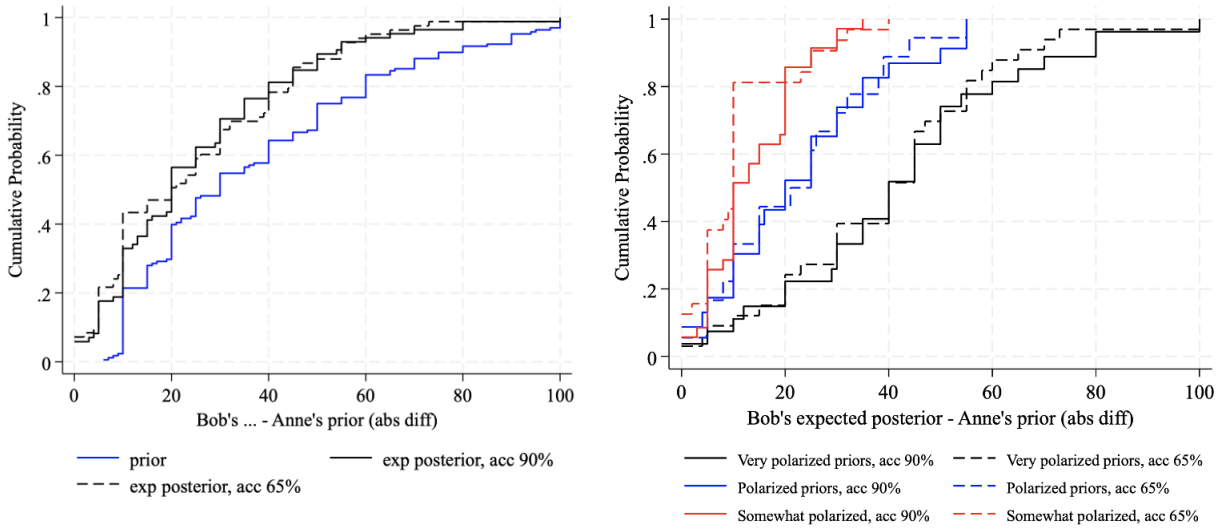
Notes: We present all statements used in the experiment and Anne's prior beliefs for each statement. The dashed line indicates the value of the prior used in Part 2 of T1 and T2, i.e., Bob's prior.

Figure 14: Statements and Anne's Prior Beliefs (treatment T0, part 2)



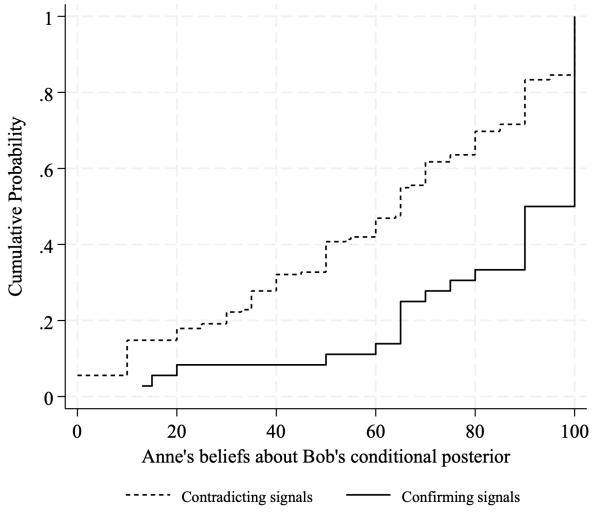
Notes: We present all statements used in the experiment and Anne's prior beliefs for each statement. The dashed line indicates the value of the prior used in Part 2 of T1 and T2, i.e., Bob's prior.

Figure 15: Changes in Bob's beliefs when Anne and Bob have different priors for politically-charged statements (statements 3 and 6)



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.

Figure 16: How Anne Thinks Bob Updates his Corner Beliefs



Notes: The figure depicts two CDFs, one for what Anne thinks Bob's posterior will be after observing a confirming signal and one for a contradicting signal. In both cases, Bob starts from the degenerate prior equal to 100. The data is from Part 2 of treatment T1.

7.1 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 3 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 5 clearly shows.

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