Belief Updating with an Unknown DGP: Experimental Evidence from AI Recommendations*

Matthew Kovach Daniel Martin Gerelt Tserenjigmid

November 2, 2026

Abstract

To study how beliefs are updated after receiving qualitative information from an unknown data-generating process (DGP), we conduct a controlled experiment in which participants receive AI recommendations. Across 60,252 pairs of prior and posterior beliefs, we document three behavioral patterns: average updating is close to zero when recommendations confirm extreme priors, larger when recommendations contradict extreme priors, and smaller for intermediate priors. These three behavioral patterns suggest four properties of belief updating, and we assess their implications at the aggregate and participant levels. Finally, we examine how well updates are captured by three models of belief updating.

Keywords: Belief updating, AI recommendations, Grether rule, Inertial Updating, Contraction Rule, General Information.

*Kovach: Department of Economics, Purdue University, *mlkovach@purdue.edu*. Martin: Department of Economics, UC Santa Barbara, *danielmartin@ucsb.edu*. Tserenjigmid: Department of Economics, UC Santa Cruz, *gtserenj@ucsc.edu*. We thank Kenneth Chan, Yeon-Koo Che, Daniel Friedman, Shachar Kariv, Yoram Halevy, P. J. Healy, Collin Raymond, and seminar, conference, and workshop audiences for many helpful comments and suggestions. We thank Anastasiia Morozova and Leshan Xu for their excellent research assistance. Martin thanks the Sloan Foundation for supporting this research under the "Cognitive Economics at Work" grant. IRB approval was obtained at the University of California, Santa Barbara (11-22-0691) and Purdue University (2025-349). Preregistration can be found at https://aspredicted.org/fz8s-535m.pdf, and Appendix A details how our analyses relate to the preregistration.

1 Introduction

Understanding how a decision maker (DM) updates their beliefs after receiving information is a foundational issue in economics. The standard approach to modeling belief updating is to assume that the DM knows the information structure or data-generating process (DGP) of the received information and updates their beliefs accordingly. As a result, experiments that study belief updating often make the DGP explicit through transparent information structures, such as colored balls drawn with replacement from cups or urns with known compositions (Holt and Smith, 2009; Danz, Vesterlund, and Wilson, 2022), binary signals with known reliability (Coutts, 2019), biased computerized information sources with explicitly specified conditional signal distributions (Charness, Oprea, and Yuksel, 2021), or canonical base-rate updating problems in which base rates and signal reliabilities are provided (Esponda, Vespa, and Yuksel, 2024). In these treatments with a transparent DGP, the experimenter communicates the relevant mapping from states to signals, so the Bayesian posterior is well defined and deviations from the Bayesian benchmark are often interpreted as errors or biases (Benjamin, 2019). However, in many settings in the world, the process by which information is generated is neither transparent nor described in probabilistic terms, and there is less evidence on how beliefs are updated in such settings.[1]

To help fill this gap in the literature, we conduct an online experiment in which participants receive messages with qualitative information from an unknown DGP. We describe a DGP as *unknown* when the conditional probabilities for messages, given the state and the DM's existing information, are not fully specified to the DM. Importantly, this does not imply that the DM has no knowledge about the source or cannot form subjective beliefs about those probabilities. *Qualitative information* communicates the direction of the source's assessment without providing a numerical probability. We study this type of information because received information often takes this form in real-world settings, whether it be a suggested action, a most likely state, or an alert. For instance, a transaction may be flagged as fraudulent, a patient may be recommended for additional screening, or an applicant could be classified as high risk. These messages convey information about an underlying state, but do not provide a numerical probability or specify the mapping from states to messages.

In the world, unknown DGPs could come in many forms: human advice, expert

[1] See Section 6.1 for a discussion of existing evidence and related literature.

judgment, news, analyst forecasts, and institutional alerts. As a result, there are potentially many ways to generate an unknown DGP in an experiment. For instance, a qualitative message could be sourced from a human novice or expert, a crowd forecast, or a statistical model. Additionally, experimental participants could be told more or less about the source of those messages. In our experiment, the qualitative information comes in the form of artificial intelligence (AI) recommendations, which are particularly well suited for studying how beliefs are updated when the DGP is unknown. First, AI is often described as being a "black box" in news media, so there is reason to believe participants will treat the DGP for AI recommendations as unknown. Second, we can evaluate the AI's recommendations across many instances, which allows us to estimate recommendation probabilities and construct our Bayesian benchmark. Third, the DGP of the AI is stable, reducing unmodeled issues arising from dynamic changes in recommendations.[2]

Our online experiment adapts the "Bouncer" task introduced by Caplin, Deming, Li, Martin, Marx, Weidmann, and Ye (2026). In each of 160 rounds, participants viewed pictures of a face and reported their belief that the person was over 21 when the picture was taken. After the elicitation of their initial belief that the person was over 21 (henceforth, their *prior belief*), an "AI Assistant" on the screen announced one of two recommendations regarding the person's age: "More likely Over 21" or "More likely Under 21" (for simplicity, we will refer to these as the *Over 21* or *Under 21* recommendations, respectively). Participants could then revise their prior before reporting a final belief that the person was over 21 (henceforth, their *posterior belief*). Our main goal is to document and understand belief updating (i.e., how prior and posterior beliefs are related), so we incentivized truthful elicitation by having the prior and posterior beliefs be equally likely to determine payoffs and by determining the bonus payment based on a randomly selected belief using the incentive-compatible binarized quadratic scoring rule.

Participants in our experiment were randomized into one of two treatments. In the NOINFO treatment, participants received no information about the machine learning (ML) model used to generate the AI recommendation, the accuracy of the AI recommendation, or the mapping from ML outputs to the AI recommendation. In the INFO treatment, participants were provided information about the model and its per-

[2]Section 6.3 discusses how updating after AI recommendations might differ from updating after human recommendations and the empirical challenges in making this comparison.

formance, the AI's overall accuracy, and the AI's state-conditional recommendation rates. We varied the information provided to participants across these two treatments for several complementary reasons. First, we wanted to investigate whether providing some likelihood information would cause participants to update more like they do in experiments where the DGP is explicitly stated. Second, this variation allows us to examine how updating changes when participants have more or less uncertainty about the accuracy of an unknown DGP. Third, knowledge about an AI's accuracy can vary through experience or direct information, so this treatment variation lets us examine robustness along a dimension that may be important for generalizability.

Our experiment generated a panel of 60,252 pairs of prior and posterior beliefs from 377 U.S. residents that are roughly balanced across the two treatments. From these pairs, we document three behavioral patterns in participant updating. Average updating is close to zero when recommendations confirm extreme priors, larger when they contradict extreme priors, and smaller at intermediate priors.

These three behavioral patterns are illustrated in Figure 1. Panel (a) shows the average prior and posterior for each decile of prior belief by recommendation in the pooled sample. Black circles represent the Over 21 recommendation, and gray triangles represent the Under 21 recommendation. The 45-degree line represents no updating. After Over 21 advice, average absolute updating is less than 1 percentage point when the prior is at least 90%. Average movement toward the recommendation is 29.2 percentage points when the prior is at most 10%, compared with 8 percentage points when the prior is between 40% and 49%. The figure shows a similar asymmetry after Under 21 advice. Panels (b), (c), and (d) show these relationships separately for the first and second halves of rounds, the two information treatments, and correct and incorrect AI recommendations. We find similar patterns when we split the data by several other participant and task characteristics (Appendix Figure S1).

These three behavioral patterns suggest four candidate properties of belief updating. The first is *consistency* with the AI recommendation: average posterior beliefs move weakly toward the recommendation, rising after an Over 21 recommendation and falling after an Under 21 recommendation. The second is *monotonicity*: for each recommendation, average posterior beliefs are weakly increasing in prior beliefs. The third is *reactionary updating*: average movement toward the recommendation is weakly larger when the recommendation more strongly contradicts the prior, so updating after Over 21 is weakly larger at lower priors and updating after Under 21

is weakly larger at higher priors. The fourth is *threshold updating*: for each recommendation, average updating is zero once priors are sufficiently confirming.

To provide a test of these properties, we group prior beliefs into four fixed ranges for each recommendation, with cutoffs at 25%, 50%, and 75%. For the consistency property, we test whether beliefs move away from the recommendation on average within each range. For monotonicity, we compare average posteriors across ranges to determine whether they decrease. For reactionary updating, we compare average movement toward the recommendation as contradiction increases. In aggregate tests for the pooled sample and reported splits, including by treatment, we detect no violations of these restrictions. For threshold updating, average updating does not differ significantly from zero in the confirming ranges of 75–100% after Over 21 and 0–24% after Under 21 in the pooled sample or either treatment.

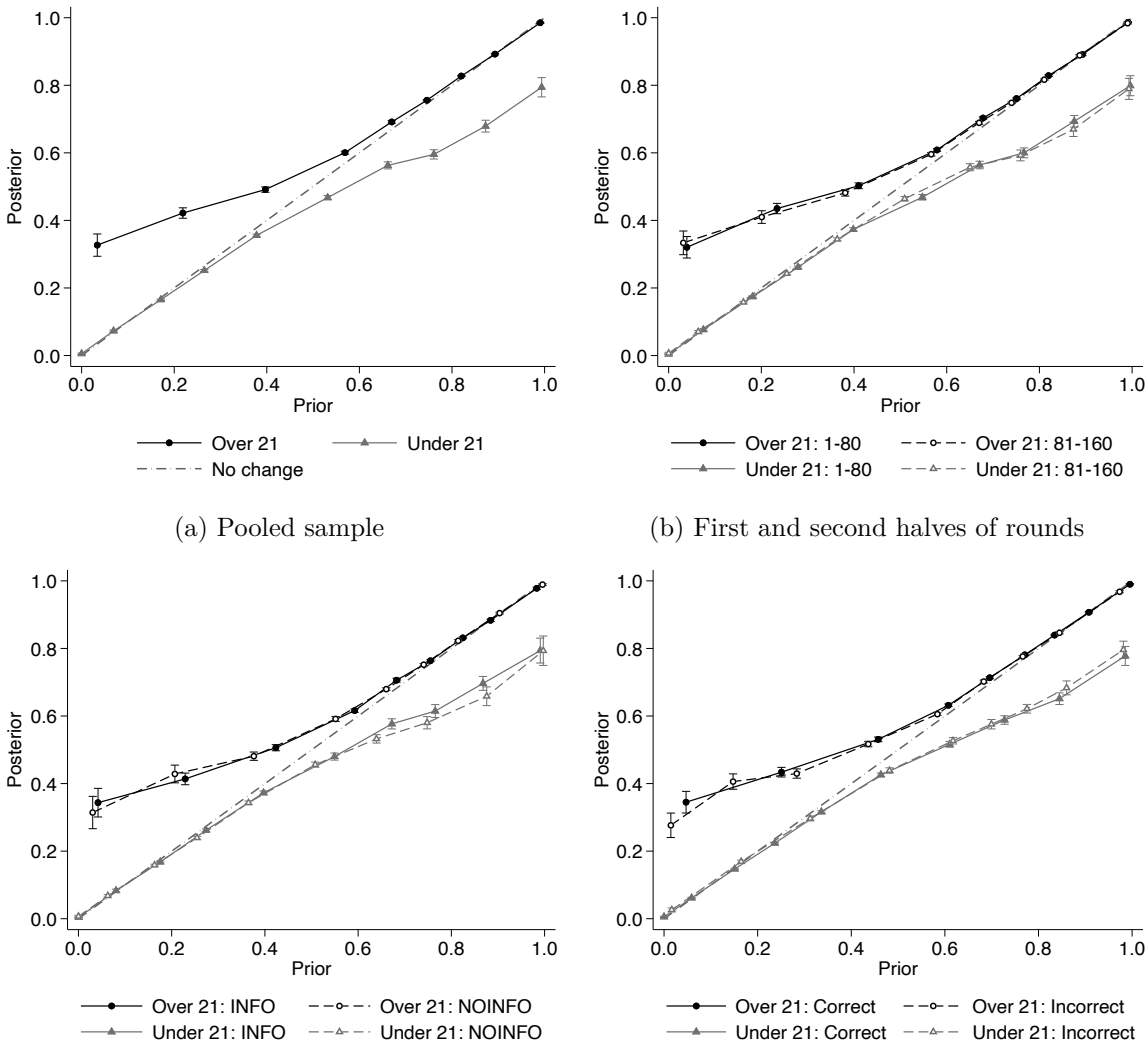
The evidence at the participant level is more mixed. We detect no consistency violations on observed support for 97.9% of participants, no monotonicity violations for 93.9%, and no reactionary-updating violations for 93.9%. For threshold updating, 85.4% satisfy our participant-level test. Overall, 77.5% meet all four empirical criteria. Another 15.6% meet exactly three of the four criteria.

Next, we examine how well the updates in our experiment are captured by several models of belief updating. As a Bayesian benchmark, we use the Bayesian update over binary states (Over 21 and Under 21) when recommendation probabilities depend on the reported prior and current state. Thus, our benchmark assumes that recommendation probabilities depend on private information only through the reported prior.[3] Also, we use a specification without learning about the DGP or oneself because we see similar aggregate updating across round halves of our experiment (Figure 1(b)), though a dynamic benchmark may be necessary in other settings.[4]

The first alternative that we consider relative to this Bayesian benchmark is a *quasi-Bayesian* (qB) specification that generalizes the Bayesian benchmark and nests Grether's rule (Grether, 1980). We also consider two non-Bayesian proposals for how DMs revise their beliefs after receiving information from a data-generating process

[3]This assumption makes the empirical comparison operational because a recommendation and prior do not determine a unique Bayesian posterior without additional restrictions on subjective likelihoods. Related theoretical discussions appear in Shmaya and Yariv (2016), Molavi (2026), and Dominiak et al. (2026).

[4]Appendix B.2 discusses the version imposing signal-dependence neglect, a related specification considered by Agarwal, Moehring, Rajpurkar, and Salz (2023).



(a) Pooled sample

(b) First and second halves of rounds

(c) INFO and NOINFO treatments

(d) Correct and incorrect AI recommendations

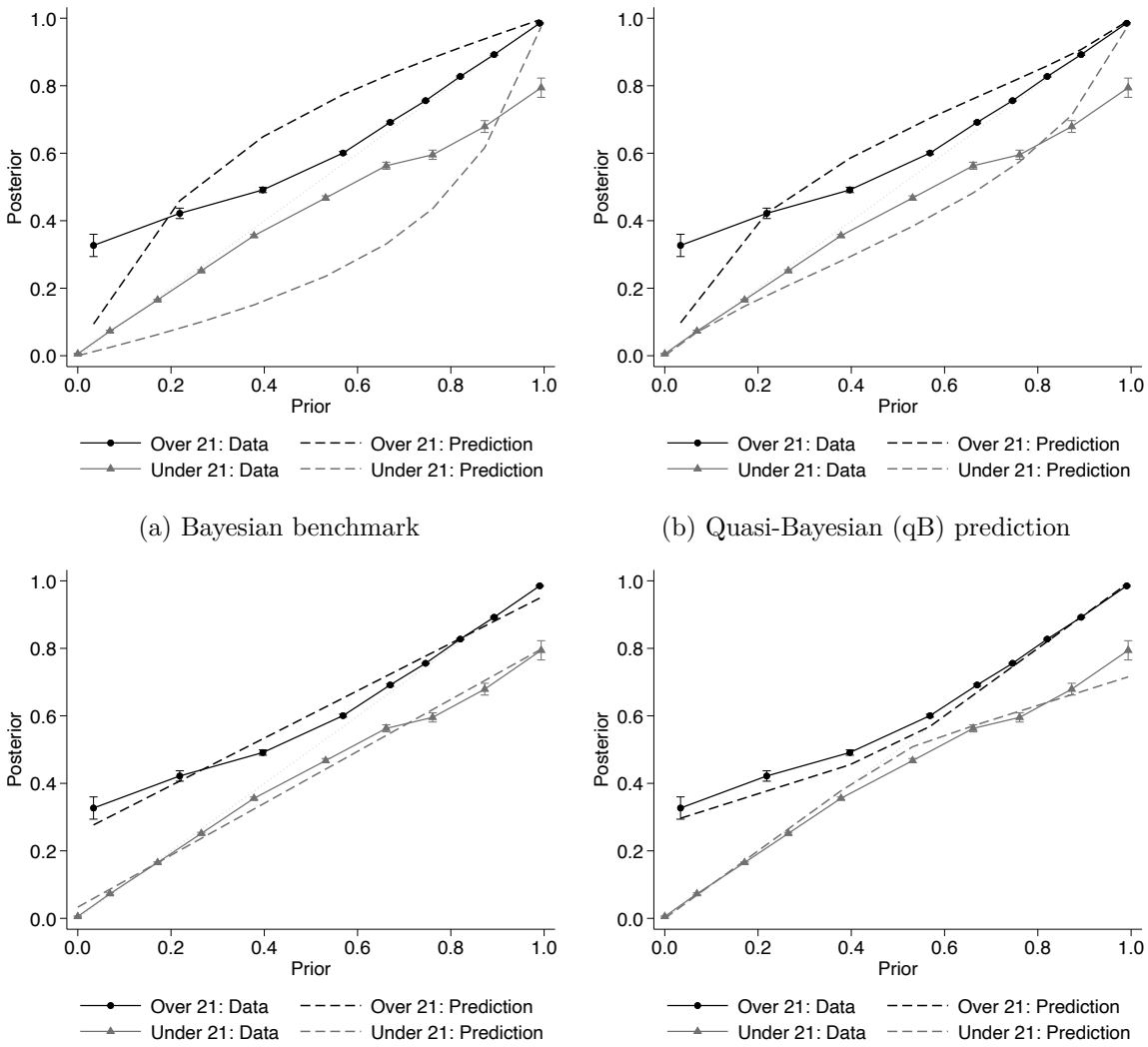
Figure 1: Average prior and posterior beliefs by AI recommendation. Panel (a) pools all observations. Panel (b) separates rounds 1–80 and 81–160. Panel (c) separates the INFO and NOINFO treatments. Panel (d) separates correct and incorrect AI recommendations. Black circles represent Over 21 and gray triangles Under 21. Points show mean posterior beliefs against mean prior beliefs within prior deciles formed separately by recommendation and, where applicable, comparison group. The 45-degree line represents no change in mean beliefs. Bars show ±1 standard error clustered by participant and image, conditional on the observed bins.

that is not well understood, but is nevertheless informative. The first of these is the *Contraction Rule* (CR) of Ke, Wu, and Zhao (2024), which captures a fixed shrinkage toward a representative belief that is consistent with the recommendation. The other is *weighted Inertial Updating* (wIU), which we operationalize within the general-information framework of Dominiak, Kovach, and Tserenjigmid (2026). This model posits that DMs choose the belief closest to their prior that satisfies an information constraint, then partially mix that belief with the prior to form their posterior. While Ke, Wu, and Zhao (2024) and Dominiak, Kovach, and Tserenjigmid (2026) both discuss AI recommendations as a setting where these updating rules might apply, we are not aware of existing empirical applications of either CR or wIU.

Figure 2 plots average beliefs in the pooled sample alongside predictions from our Bayesian benchmark, qB, CR, and wIU. From this figure, CR and wIU appear closer to aggregate belief updating than the Bayesian benchmark and qB. Because these models differ in the number of free parameters, we assess fit more rigorously in Section 5 by comparing the three models using out-of-sample performance as well as the completeness and restrictiveness measures of Fudenberg, Kleinberg, Liang, and Mullainathan (2022) and Fudenberg, Gao, and Liang (2026), respectively.[5] In the pooled sample, we find that CR and wIU have lower out-of-sample MSE than qB across all 100 random partitions and higher in-sample completeness. wIU is also more restrictive than both qB and CR under both of the restrictiveness benchmarks we consider, whether the eligible class imposes monotonicity or only recommendation consistency.

The three behavioral patterns of belief updating can help to explain the relative fit of these models. Figure 1(a) documents large updating for extreme priors and contradicting recommendations, but almost no updating for extreme priors and confirming recommendations. In contrast, Figure 2(a)–(b) shows that the Bayesian benchmark and qB predict minimal updating for extreme priors, while maximum updating occurs at an interior prior for both recommendations. As shown in Figure 2(c)–(d), CR and wIU naturally accommodate larger updates at extreme contradicting priors and smaller updates at intermediate priors. In addition, wIU captures average updating that is close to zero when recommendations confirm extreme priors because it predicts unchanged beliefs after confirming recommendations.

[5] qB has four free parameters: three bias parameters and one prior-dependence coefficient. In our setting, CR has four free parameters and wIU has one.



(a) Bayesian benchmark

(b) Quasi-Bayesian (qB) prediction

(c) Contraction Rule (CR) prediction

(d) Weighted Inertial Updating (wIU) prediction

Figure 2: Average beliefs and fitted predictions by AI recommendation. Black circles represent Over 21 and gray triangles Under 21. Solid lines connect observed means, and dashed lines show predictions averaged within the same prior deciles. Panel (a) shows the Bayesian benchmark. Panels (b)–(d) show qB, CR, and wIU fitted in-sample. Bars show ±1 standard error of the posterior mean, clustered by participant and image.

Across participants, CR and wIU have lower mean out-of-sample MSE than qB. Comparing models for each participant, we find that roughly a third of participants have the lowest MSE under each of the models. In head-to-head comparisons with qB, CR and wIU have a lower MSE for around half of participants (58.1% for CR and 48.5% for wIU).

The remainder of the paper is organized as follows. Section 2 details the experimental design. Section 3 presents the high-level results of the experiment, including the property tests. Section 4 describes the three models and presents their estimates and predictions. Section 5 assesses the three models using our experimental data. We treat the three updating rules as functional forms (i.e., reduced-form) to be fit to data, not as tests that identify a unique behavioral microfoundation, but in Section 6.2 we discuss possible interpretations. Section 6.1 briefly discusses related literature. Section 7 concludes.

2 Experimental Design

2.1 Task

In each round of our experiment, participants were shown an image of a human face and asked to make two decisions based on it. In the instructions, participants were told that all images were taken between 2010 and 2014. In half of the images, the person was over 21 years old at the time the picture was taken, and in the other half of the images, the person was under 21 years old. Participants were told that the people in the images were either over 21 or under 21 when the picture was taken, but not the fraction of people that were over 21 or under 21.[6]

The first decision participants were asked to make was to report the probability that the person in the image was over 21 at the time that the picture was taken, which we refer to as their *prior belief*. Participants were provided a slider bar to indicate their belief and, to minimize anchoring effects, the slider bar had no preset value.

After submitting their prior belief, participants were shown an AI recommendation as a pop-up message. This recommendation was framed as the prediction of an "AI Assistant". As previously stated, there were two possible recommendations:

[6]Instructions and screenshots for both treatments, including the payment screens, are provided in the experimental materials accompanying this paper.

"More likely Over 21" and "More likely Under 21". After closing the pop-up message, the recommendation was presented in a gray bar above the submit button.

Once participants closed the message box, the second decision task began. In this decision, they were allowed to revise their reported probability before submitting their final report, which we refer to as their *posterior belief*. The posterior slider retained the participant's submitted prior, which could be resubmitted without moving the slider. Participants were not required to change their beliefs and were reminded in the pop-up message that "you can change your reported probability, but you are not required to do so".

After submitting their posterior, participants proceeded to the next round, and they could not go back and change any of their previous choices. Altogether, participants completed 160 rounds of this task, broken into blocks of 20 rounds. To minimize fatigue effects, participants were given the opportunity to pause the experiment after each block of rounds. For each participant, images were randomly allocated to rounds.

To incentivize truthful reporting of beliefs, a single round was selected (all equally likely) and either the prior or posterior was randomly selected from that round (both equally likely).[7] To determine payment based on the randomly selected belief and true age of the person in the image, we used the incentive-compatible binarized quadratic scoring rule to calculate the probability of getting a $6 bonus payment.[8] Danz, Vesterlund, and Wilson (2022) show that detailed information about binarized scoring rule (BSR) incentives can increase deviations from truthful reporting. We therefore initially provided no information about the payment rule aside from telling participants that they would maximize the likelihood of receiving the bonus payment by truthfully reporting their belief in each round. However, participants were given the opportunity to learn more about the payment rule after completing the experiment.

Participants were given 60 seconds to complete each round, and they were informed that if they did not complete a round within 60 seconds and that round was randomly selected for payment, then they would have no chance of receiving the bonus

[7] In the pop-up message, participants were reminded that "your reports before and after seeing the AI Assistant's guess are equally likely to be selected for payment".

[8] The idea of paying in probabilities predates the term *binarized scoring rule*. Smith (1961) describes probability currency, and Savage (1971, p. 785) discusses its use in elicitation. McKelvey and Page (1990) provide an early experimental application of the quadratic scoring rule. Hossain and Okui (2013) develop a general construction and analyze its incentive properties.

payment. Approximately 99.9% of rounds were completed within the time limit, and the rounds in which the time limit was hit were excluded from our analysis sample.

2.2 AI Recommendations

To generate AI recommendations, we used the outputs of the model for age estimation developed by Rothe, Timofte, and Van Gool (2018), implemented in the Caffe framework. This model was trained to predict the ages of humans based on images of faces, and it outputs a probability score between 0 and 1 for each possible age, where the scores sum to 1 across all possible ages. To generate a recommendation for an image, we summed the probability scores for all ages over 21. The AI Assistant gave the "More likely Over 21" recommendation whenever this sum was above 50% and otherwise gave the opposite recommendation. The AI model probability of Over 21 was never exactly 50%.

For our primary treatment, which we refer to as the NOINFO treatment, we provided participants no information about the underlying model used to generate AI recommendations, nor any information about how its predictions were mapped into AI recommendations. For instance, we did not tell participants anything about the accuracy of the AI or whether the AI recommendations were balanced across recommendations. The reason for this was twofold. First, we wanted to maximize the extent to which the DGP was viewed as a "black box" by the participants, which is the focus of our research question. Second, in many situations in which individuals receive AI recommendations outside of our experiment, such as the results of an Internet search, a music recommendation, or a navigation route, they are not told about how the results are generated.

As a complement to this treatment, we also conducted an INFO treatment, in which participants were provided with several pieces of information about the AI recommendations. First, they were told the overall accuracy of the AI ("the AI's prediction matches whether the person was actually under or over 21 approximately 76% of the time") and were given qualitative information about the AI's relative performance ("based on past studies, this performance is better than most humans, but is worse than the very best humans"). Following Agarwal, Moehring, Rajpurkar, and Salz (2023), we also told participants that the AI did not have extra information beyond the picture ("it uses only the image to predict the person's age") and who

programmed the AI ("the tool is based on state-of-the-art machine learning algorithms developed by a leading team of researchers at the Computer Vision Lab at the Federal Institute of Technology in Zurich, Switzerland"). In this treatment, participants were also informed of the AI's state-conditional rates of correct recommendations alongside the AI advice in the pop-up message in each round: "when the person is actually over 21, the AI Assistant predicts 'More likely Over 21' about 83% of the time" and "when the person is actually under 21, the AI Assistant predicts 'More likely Under 21' about 70% of the time".

2.3 Implementation

Our experiment was conducted on Prolific. Participants are restricted by Prolific to be at least 18 years old, and we added a requirement that they be residents of the U.S. as well. In the posting for the experiment, participants were told that the expected time to complete the study was 40–45 minutes, that the payment for completing the experiment was $6, and that they could possibly receive a bonus payment of $6 based on their decisions in the experiment.

Participants who chose to complete our experiment were shown a consent form that provided additional details about our experiment. Participants who chose to participate were then asked two questions as a standard attention check. Failure to correctly answer either question on the first try resulted in participants being prevented from completing the experiment.

After completing all 160 rounds and before payment results were disclosed, participants completed an unincentivized survey measuring perceived AI accuracy and confidence in their prior and posterior beliefs. The separate *Supporting Experimental Documentation* describes the survey page sequence, slider wording, and payment-screen sequence.

2.4 Design Decisions

Task Choice: We chose the Bouncer task for several reasons. First, it has a documented ground truth, so performance can be assessed. Second, it requires no specialized training, which allowed us to obtain a large sample of participants. Third, when implemented previously by Caplin et al. (2026), it generated a wide range of beliefs at the participant and image levels without AI recommendations.

Image Selection: We used the same 160 images used in Caplin et al. (2026, endnote 7). These images were originally drawn from the IMDB-WIKI dataset, which is a large publicly available face image dataset with age labels. The age labels are 16–19 years for Under 21 images and 23–26 years for Over 21 images. Given the recorded birth dates and capture years, none of these images could depict someone aged exactly 21.[9]

Belief Elicitation: The belief elicitation literature contains several incentive-compatible methods, including the binarized scoring rule (BSR), the multiple-price list (MPL), and the Becker-DeGroot-Marschak (BDM) method. We chose the BSR because MPL and BDM are less practical in our setting, given our desire to elicit two beliefs for 160 images.

Number of Rounds and Time Limit: To determine the number of rounds and to test the time limit, we ran a pilot session on Amazon Mechanical Turk (MTurk) with a similar design to our final design. We did not see evidence of dynamic effects or time constraints in that pilot, so we kept those parameters the same.

Information about the AI: Both treatments leave aspects of the recommendation process unspecified. INFO additionally supplies recommendation probabilities conditional on the true state. These probabilities suffice for Bayesian updating under conditional independence of recommendations and private visual information given the state, but INFO does not specify that dependence. In our pilot session, we did not see treatment differences even when we provided additional information about the AI, such as the accuracy by recommendation. However, we decided to keep this treatment variation in our final experiment to verify the robustness of our results along this dimension.

3 Experimental Results

In total, 377 participants completed our experiment, of whom 189 were in the NOINFO treatment and 188 were in the INFO treatment. Participants were required to be U.S. residents and at least 18 years old. The median self-reported age was 35 years old, 59% self-reported being women, and 62% self-reported being White. We find no significant differences in average age ($p = 0.072$), fraction women ($p = 0.490$),

[9] Appendix Table S9 reports summary statistics at the image level for this selection of images.

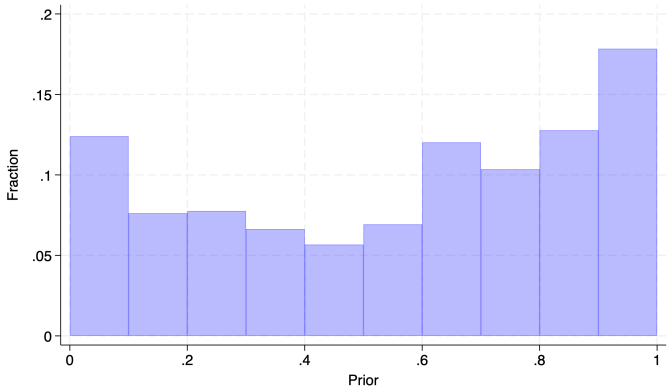


Figure 3: Distribution of prior beliefs in the pooled sample. Bins contain integer reports of 0–9%, 10–19%, ..., 90–100%.

and fraction White ($p = 0.485$) across treatments.[10]

This sample yielded a total of 60,320 rounds (377 participants $\times$ 160 rounds each). We excluded 68 rounds ($\approx 0.11\%$) because the 60-second time limit was hit. The exclusion rate did not differ significantly across treatments ($p = 0.983$). This exclusion left 60,252 pairs of prior and posterior beliefs for analysis.[11] From the start of the round, the median and mean response times for prior beliefs were 4 seconds and 4.8 seconds, respectively, and for the posterior beliefs were 7 seconds and 8.4 seconds, respectively.[12] The median time to complete the experiment was 40 minutes. Among the 377 participants in our experiment, the average payment was \$9.26, including a \$6 participation payment and a possible \$6 bonus.

3.1 Prior Beliefs

We document substantial heterogeneity in prior beliefs across participants and images. Figure 3 shows that prior beliefs are widely dispersed in the pooled sample. The most populated 10-percentage-point bin contains only 19.0% of observations. We also see considerable variation within participants across images (mean largest bin fraction=35.6%) and within images across participants (mean largest bin fraction=23.2%). While prior beliefs are broadly dispersed, 19.0% of reports are at the

[10] Appendix Table S9 reports balance across participants on these and other dimensions (e.g., survey measures, accuracy, response times, non-updating rates).

[11] Appendix B.9 compares the binary accuracy of prior beliefs, posterior beliefs, and AI recommendations.

[12] See Appendix Figure S4 for the incremental response time by prior and AI recommendation.

endpoints and less than 1% are at the midpoint (exactly 50%). Reports of exactly 0%, 25%, 50%, 75%, and 100% account for 7.5%, 0.9%, 0.6%, 1.3%, and 11.5% of observations, respectively. Reports near 50% are also less common than reports near the endpoints. With 25 bins of width four percentage points $[0, 4), [4, 8), \ldots, [96, 100]$, the highest share is 14.1% in $[96, 100]$ and the lowest is 1.7% in $[48, 52)$.

3.2 Three Patterns and Four Properties

We document the three behavioral patterns described in the introduction using Figure 1(a). First, the figure shows near-zero average changes at extreme confirming priors. This reflects little movement in the underlying reports, rather than simply offsetting increases and decreases. In the pooled sample, unchanged reports are particularly common when recommendations confirm extreme priors. After Over 21 advice, 95.2% of reports are unchanged when the prior is at least 90%. After Under 21 advice, 94.8% are unchanged when the prior is at most 10%. Average absolute updating in these two groups is 0.79 and 0.82 percentage points, respectively.

In contrast, average movement toward the recommendation is 29.2 percentage points after Over 21 advice when the prior is at most 10%, and 20.0 percentage points after Under 21 advice when the prior is at least 90%. At moderate contradicting priors, the corresponding averages are 8.0 percentage points for priors of 40–49% after Over 21 advice and 7.8 percentage points for priors of 51–60% after Under 21 advice.

These three updating patterns motivate a more systematic examination of the effects of recommendations on prior beliefs. We focus on four properties, which we assess using conditional mean beliefs. *Consistency* requires that mean movement is not away from the recommendation, *monotonicity* requires that higher priors do not produce lower mean posteriors, and *reactionary updating* requires that mean movement toward the recommendation does not shrink as the prior becomes more contradictory. *Threshold updating* requires conditional mean updating to equal zero once priors are sufficiently confirming.[13]

To test implications of these properties, we group priors into four fixed ranges for each recommendation, 0–24%, 25–49%, 50–74%, and 75–100%. Figure 4 illustrates the four checks using these range means. For consistency, we ask whether

[13] Given the full consistency and reactionary updating restrictions, zero conditional mean updating at any prior other than certainty in favor of the recommendation implies zero conditional mean updating at all more confirming priors. Monotonicity is not needed.

mean movement is away from the recommendation in any range. There are eight checks, one for each range and recommendation. Panel (a) shows that movement is toward the recommendation or close to zero, and none of the checks detects a violation. For monotonicity, we ask whether mean posteriors fall between neighboring ranges as priors rise. The six comparisons detect no such decline. Panel (b) shows that mean posteriors rise across the four ranges after both recommendations. For reactionary updating, we ask whether mean movement toward the recommendation becomes smaller as contradiction increases. These six comparisons detect no such reversal. In panel (c), greater contradiction lies farther to the right, and mean movement rises for both recommendations.

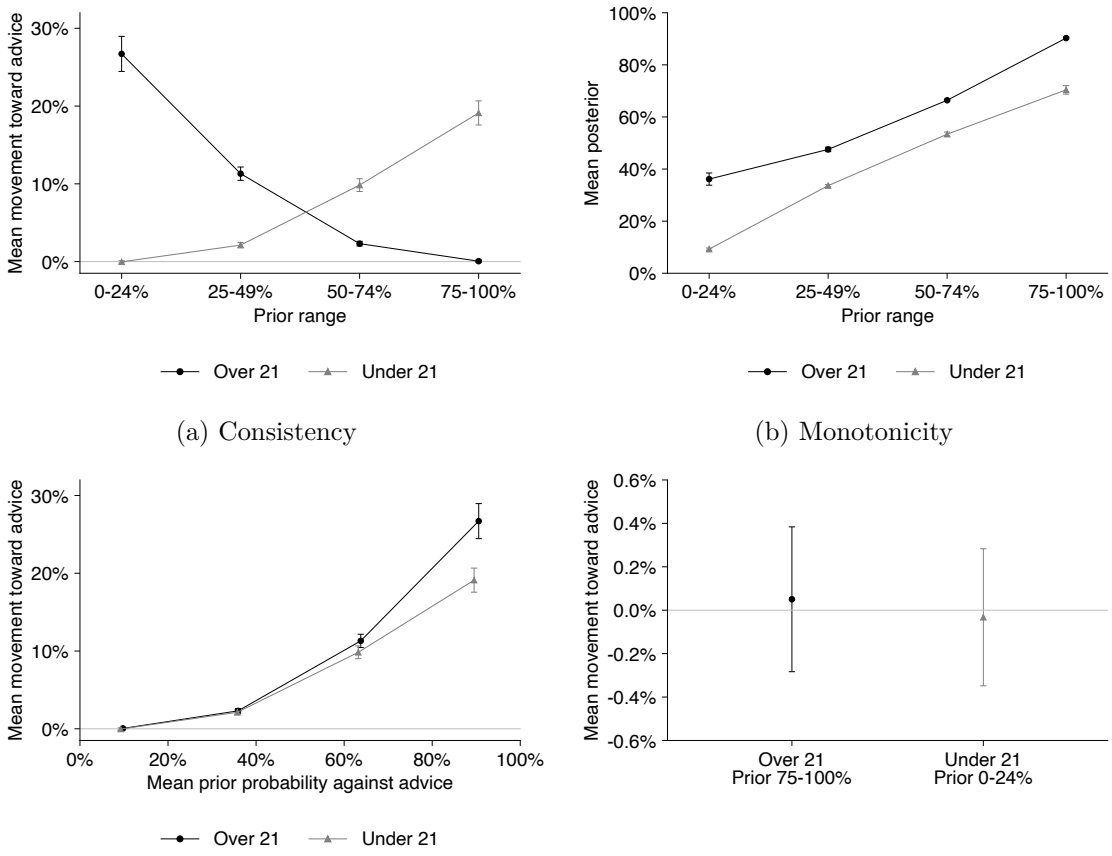
To apply the threshold criterion, we measure movement toward the recommendation, which we call *signed updating*, with positive values indicating movement toward the recommendation. Specifically, we examine whether mean signed updating differs from zero in the most confirming range, 75–100% after Over 21 and 0–24% after Under 21. In the pooled sample, mean signed updating is 0.05 percentage points after Over 21 and −0.03 points after Under 21, with standard errors of 0.15 and 0.14 points.[14] Neither differs significantly from zero. Panel (d) shows these aggregate estimates. Appendix Table S6 reports absolute changes and the frequency of unchanged reports in these confirming ranges.

3.3 Properties by Splits

No violations of consistency, monotonicity, or reactionary updating are detected in any of the 22 subgroups reported in Table 1. The threshold criterion is not met in two subgroups. When Under 21 recommendations are incorrect and priors are 0–24%, mean signed updating is −0.69 percentage points ($p = 0.034$ after adjustment), indicating movement away from the recommendation. High-ability participants instead update toward confirming recommendations by about 0.45 percentage points in each confirming range (both adjusted $p < 0.003$).

The subgroups distinguish recommendation characteristics, rounds, and participant characteristics. Correct AI recommendations match the true age category. High AI confidence means that the image-level mean of $\max\{p, 1-p\}$, where p is the

[14] Standard errors cluster by participant and image. Bonferroni adjustments use eight, six, six, and two comparisons for the four properties. Appendix B.6 describes the procedures. Appendix Tables S4 and S7 report component estimates and support in the confirming ranges.



(a) Consistency

(b) Monotonicity

(c) Reactionary updating

(d) Threshold criterion

Figure 4: Range means used in the four property tests. Black circles represent Over 21 and gray triangles Under 21. Consistency checks whether movement is away from the recommendation in any range. Monotonicity compares posteriors as priors rise. Reactionary updating compares movement as contradiction increases. The threshold criterion compares movement with zero in the most confirming range for each recommendation. Bars in panels (a)–(c) show ±1 standard error. Panel (d) shows 97.5% intervals, matching the two Bonferroni-adjusted tests at a family level of 5%. All standard errors cluster by participant and image.

AI model's probability of Over 21, exceeds the median image value of 0.772. Hard images have prior MSE above the median image value of 0.267. The round split separates rounds 1–80 from rounds 81–160. Fast participants have a median prior response time below the participant median of 3.5 seconds. Other participant splits distinguish women, participants older than the median age of 35, and high-ability participants. High ability means prior MSE at or below the median among the 377 participants, calculated after timeout exclusions. The survey belief splits are defined in the next subsection.

Table 1: Aggregate updating property tests using fixed ranges in the pooled sample and by split.

Sample	Participants	Obs.	Cons.	Mono.	React.	Thresh.
All	377	60,252	NR	NR	NR	NR
AI correct	377	45,946	NR	NR	NR	NR
Not AI correct	377	14,306	NR	NR	NR	R
AI high confidence	377	30,122	NR	NR	NR	NR
Not AI high confidence	377	30,130	NR	NR	NR	NR
Hard image	377	30,129	NR	NR	NR	NR
Not Hard image	377	30,123	NR	NR	NR	NR
Second half	377	30,133	NR	NR	NR	NR
Not Second half	377	30,119	NR	NR	NR	NR
Fast prior report	188	30,044	NR	NR	NR	NR
Not Fast prior report	189	30,208	NR	NR	NR	NR
Woman participant	224	35,800	NR	NR	NR	NR
Not Woman participant	153	24,452	NR	NR	NR	NR
Older participant	185	29,575	NR	NR	NR	NR
Not Older participant	192	30,677	NR	NR	NR	NR
High ability	189	30,218	NR	NR	NR	R
Not High ability	188	30,034	NR	NR	NR	NR
NOINFO	189	30,206	NR	NR	NR	NR
INFO	188	30,046	NR	NR	NR	NR
High belief about AI	186	29,735	NR	NR	NR	NR
Not High belief about AI	191	30,517	NR	NR	NR	NR
High confidence in prior	185	29,564	NR	NR	NR	NR
Not High confidence in prior	192	30,688	NR	NR	NR	NR

Notes: Each *Not* row is the complement of its preceding row. NR denotes non-rejection and R rejection of at least one component at a family level of 5%, using participant and image clustered standard errors. Fixed Bonferroni family sizes are eight, six, six, and two. All planned components are assessable. Appendix B.6 provides the methods.

Appendix Table S8 reports checks that exclude never-updaters, participants with no updating in at least 95% of rounds, fast prior reports, large INFO belief discrepancies, or concentrated priors. In each restricted sample, the first three tests detect

no violations, and zero mean updating is not rejected under the threshold criterion.

3.4 Beliefs, Confidence, and INFO

Using survey responses collected after the experiment, we split participants at the median belief about overall AI accuracy of 71%. Following contradicting advice, average movement is 20.4 percentage points among the 186 participants above this median and 14.4 points among the 191 at or below it. We also split participants at the median confidence in their prior beliefs of 78 out of 100. Average movement following contradicting advice is 15.8 points among the 185 participants above this median and 18.7 points among the 192 at or below it. Appendix B.12 reports the survey distributions.[15]

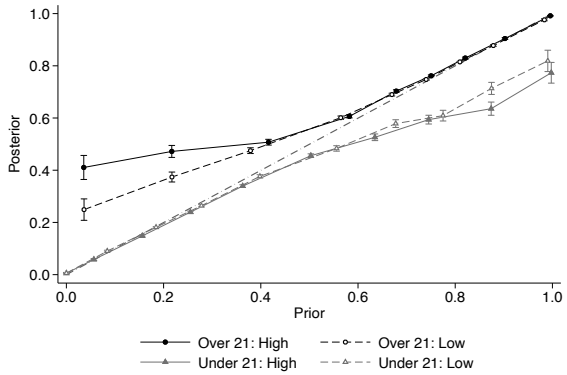
Despite these differences in the amount of updating, the three patterns remain visible in both the belief and confidence splits (Figure 5, panels (a) and (b)). The corresponding tests detect no violations of consistency, monotonicity, or reactionary updating in any of these four groups, and zero mean updating is not rejected under the threshold criterion (Table 1).

The treatments provide different information about the AI, but their aggregate updating curves are similar (see Figure 1(c)). After the experiment, mean beliefs about overall AI accuracy were 68.2% in INFO and 67.0% in NOINFO ($p = 0.544$), compared with empirical accuracy of 76.3%. INFO participants' beliefs were closer to this rate, with mean absolute errors of 13.8 rather than 18.4 percentage points for NOINFO. Panel C of Appendix Table S9 reports these comparisons.

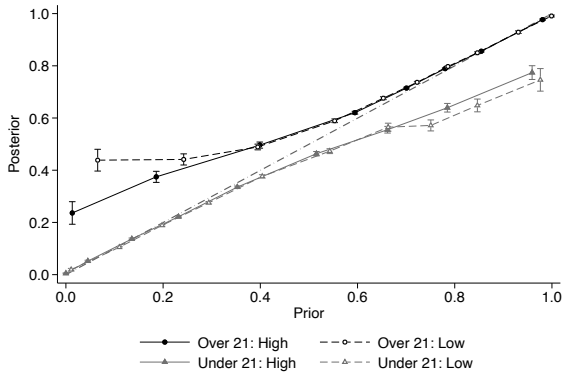
3.5 Never-Updaters and Partial Non-Updaters

Participants were explicitly told that they could leave their report unchanged. A natural question is whether this non-updating represents a deliberate decision by participants not to update their beliefs. The evidence on response times in Appendix B.13 is consistent with participants spending more time on decisions when the AI recommendation is contradictory, but also with spending some time on all decisions.

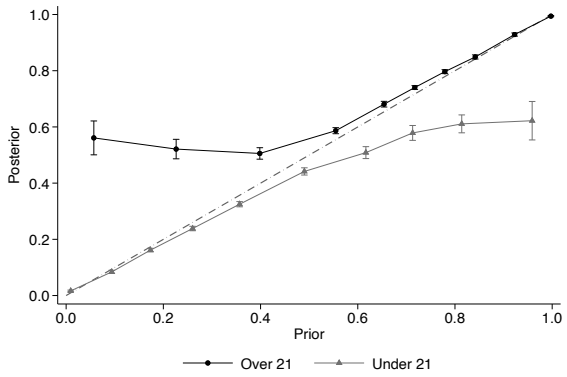
[15] Dreyfuss and Raux (2026) show that people form beliefs about AI performance by projecting human task difficulty onto AI. Their analysis concerns the formation of reliability beliefs, while ours examines how reported beliefs about AI accuracy are associated with updating.



(a) Belief about overall AI accuracy



(b) Confidence in own prior



(c) High AI accuracy belief and low prior confidence

Figure 5: Updating by beliefs and confidence. Panels (a) and (b) split participants by their survey beliefs. Panel (c) shows only participants with higher beliefs about overall AI accuracy and lower confidence in their own prior. High AI accuracy beliefs are above the participant median, and low prior confidence is at or below its median. Plotting conventions and clustered standard errors are the same as in Figure 1.

We also find there is heterogeneity at the participant level in the amount of non-updating. 8.5% of participants (32 of 377) never update their priors. All 32 faced recommendations contradicting their prior (Over 21 with a prior below one half, or Under 21 with a prior above one half), with 41–94 such cases per participant. Their non-updating therefore cannot be attributed solely to priors that always satisfy the recommendation. We retain these participants in the aggregate analyses and updating-property tests but exclude them from participant-level model comparisons because all three models can reproduce their unchanged reports. One possible interpretation of zero or low AI weight is that participants place greater confidence in their own assessment than in the AI recommendation. Never-updaters report higher average confidence in their own prior (78.5 vs. 73.5 for others) and lower average beliefs about AI accuracy (62.3% vs. 68.1% for others). Neither difference is statistically significant ($p = 0.199$ and $p = 0.127$, respectively). They are also not significantly different from the rest of the participants in prior accuracy (61.1% vs. 61.8% for others, $p = 0.594$).

Also, 80.1% of participants (302 of 377) are *partial non-updaters*: they update sometimes, but in fewer than half of observed rounds. In Section 4, we show that both the behavior of partial non-updaters and the non-updating rate in the pooled sample (78.6%) are consistent with wIU.

3.6 Participant-Level Property Tests

The four properties of belief updating also receive substantial empirical support at the participant level, although some heterogeneity exists across participants. We test the updating properties separately for each participant using the same four prior ranges and adjustments for multiple comparisons. Some comparisons cannot be assessed because the relevant ranges contain too few observations, so we test whether participants meet the criteria on ranges with sufficient observations. Appendix B.6 gives the procedures and coverage requirements.

We detect no consistency violation on observed support for 97.9% of participants and no monotonicity or reactionary-updating violation for 93.9%. Monotonicity and reactionary updating each detect violations for 11 participants, but the two groups do not overlap. We use the same threshold criterion as in the aggregate analysis. The criterion is met by 85.4% of participants (322 of 377). Overall, 77.5% of participants

(292 of 377) meet all four empirical criteria. Another 15.6% (59 of 377) meet exactly three. Appendix Table S5 reports classifications and assessability for each criterion.

4 Three Models: Estimates and Predictions

We compare three updating rules that impose different restrictions on the relationship between prior beliefs, recommendations, and posterior beliefs. Our likelihood-based model is a generalized Grether specification, which we call *quasi-Bayesian* (qB). This model captures a DM who uses Bayesian machinery but is subject to several belief distortions. The *Contraction Rule* (CR) models adjustment toward a representative belief associated with each recommendation. *Weighted Inertial Updating* (wIU) instead allows participants to retain their prior when it already satisfies the recommendation's qualitative information and otherwise adjust toward the closest belief that does. We compare these models through their implications for behavior and their predictive fit.

We first introduce the Bayesian benchmark and qB specification, then CR, and finally wIU. For each model, we define the updating rule, report aggregate parameter estimates, and explain which of the four updating properties it satisfies or violates. We then revisit the frequency of unchanged reports and discuss the model-implied weight placed on AI recommendations.

4.1 Model Notation

To formalize these models, we require a bit of notation. Let S be a finite set of states. In our experiment, the images that participants consider have many features, but we only consider the two payoff-relevant states: whether the individual is over or under 21. Thus, we model updating in our experiment using a binary state space $S = \{o, u\}$ where $o =$ Over 21 and $u =$ Under 21.

Let $\mu \in \Delta(S)$ be a prior belief before getting the AI recommendation. We do not take a stand on how participants form their prior beliefs from the images. A participant may form an approximately Bayesian perceptual belief from visual information and then use a non-Bayesian rule to incorporate a black box recommendation. As emphasized by Ortoleva (2024), different updating environments may call for different behavioral models. There is therefore no contradiction between non-Bayesian

updating over AI recommendations and approximately Bayesian visual perception, as in Bayesian brain theories or economic models of perceptual inference such as Caplin and Martin (2015).

The DM receives information in the form of a recommendation R from a set of possible recommendations $\mathcal{R}$. For $i \in \{o, u\}$, let R_i denote the recommendation that state i is more likely. Let $I(R) \subseteq \Delta(S)$ denote a convex set of probability distributions that are consistent with the qualitative information of recommendation R.

A qualitative recommendation can be represented either as a signal or through its qualitative information. For example, qB interprets an Over 21 recommendation using its perceived probability of occurring in each state. A DM can posit subjective conditional recommendation probabilities and apply Bayes' rule, provided the observed recommendation has positive subjective probability. In CR and wIU, the same recommendation is associated with the set of beliefs assigning probability at least one half to Over 21, $I(R_o) = \{\pi \in \Delta(S) \mid \pi(o) \geq 1/2\}$, while $I(R_u) = \{\pi \in \Delta(S) \mid \pi(o) \leq 1/2\}$, consistent with the objective construction of recommendations. The qualitative set constrains the representative belief in CR or the closest belief in wIU, not necessarily the final posterior after mixing with the prior.

Finally, let $\mu_R \in \Delta(S)$ be the posterior after receiving a recommendation R. We want to model how the DM revises their prior μ and obtains a new belief (posterior) μ_R after receiving a signal R or the qualitative information generated by R.

4.2 Bayesian and Quasi-Bayesian Updating

Bayesian conditioning uses the recommendation process relevant after the DM observes private visual information X. If the true joint distribution of states, visual information that the participant perceives, and recommendations were known, the Bayesian posterior would be

$$\Pr(o \mid X, R) = \frac{\Pr(R \mid o, X)\ \Pr(o \mid X)}{\Pr(R \mid o, X)\ \Pr(o \mid X) + \Pr(R \mid u, X)\ \Pr(u \mid X)}.$$

In its most general form, where the conditional recommendation probability $\Pr(R|o, X)$ depends on the picture-participant-specific information X and is unknown, it has essentially no testable restrictions on interior prior and posterior pairs. See Shmaya and Yariv (2016), Molavi (2026), and Dominiak et al. (2026)

for related theoretical "anything goes" results. Our empirical Bayesian benchmark is a Markovian specification where the reported prior summarizes the private information relevant to recommendation probabilities conditional on the true state: $\Pr(R \mid s, X) = \Pr(R \mid s, \mu)$. As mentioned previously, the relative stability of aggregate updating across rounds and treatments motivates our specification without additional learning components.

We first estimate how recommendation probabilities vary with the true state and reported prior. We then use those probabilities either to calculate the Bayesian benchmark or to estimate how participants weight recommendations and priors under qB. The estimated recommendation process is

$$(1)\quad \widehat{\delta}(R_o \mid s, \mu) = \Lambda\big(\widehat{\gamma}_0 + \widehat{\gamma}_1 \, \mathbf{1}\{s = o\} + \widehat{\gamma}_2 \, \mu(o)\big), \qquad \widehat{\delta}(R_u \mid s, \mu) = 1 - \widehat{\delta}(R_o \mid s, \mu),$$

where Λ is the logistic function. Our aggregate estimate of $\gamma_2 = 0.992$ indicates a positive association between the prior and receiving an Over 21 recommendation conditional on the true state. Setting $\gamma_2 = 0$ removes dependence between priors and recommendations conditional on the true state, as in the signal-dependence neglect specification presented in the Appendix. Priors and recommendations can remain correlated unconditionally. Given the prior-dependence coefficient γ_2 and the observed distribution of priors within each state, γ_0 and γ_1 can be recovered by matching average fitted recommendation probabilities to the state-conditional recommendation frequencies.

The resulting Bayesian prediction using raw priors is

$$(2)\qquad \mu_R^B(o) = \frac{\widehat{\delta}(R \mid o, \mu)\mu(o)}{\widehat{\delta}(R \mid o, \mu)\mu(o) + \widehat{\delta}(R \mid u, \mu)\mu(u)}.$$

Interpreting this prediction as the conditional probability of Over 21 requires calibrated priors and a correctly specified recommendation process.[16]

The qB specification generalizes our Bayesian benchmark by applying power distortions to prior probabilities and conditional recommendation likelihoods, while al-

[16] We calibrate the recommendation process to the data while taking participants' reported priors as given, rather than adjusting them for miscalibration.

lowing a bias toward one state:

$$\mu_R(o) = \frac{\Pr(R \mid o, \mu)^{\beta_1}\, \mu(o)^{\beta_2}\, \exp(\alpha)}{\Pr(R \mid o, \mu)^{\beta_1}\, \mu(o)^{\beta_2}\, \exp(\alpha) + \Pr(R \mid u, \mu)^{\beta_1}\, \mu(u)^{\beta_2}}, \tag{3}$$

where $\alpha \in \mathbb{R}$ and $\beta_1, \beta_2 \geq 0$. Here, β_1 weights the conditional recommendation likelihood ratio, β_2 weights the direct prior-odds term, and α is a state-direction bias. In our empirical implementation, we replace $\Pr(R \mid s, \mu)$ with the estimated recommendation probabilities $\widehat{\delta}(R \mid s, \mu)$ from Equation (1). Using these same estimates, setting $\alpha = 0$ and $\beta_1 = \beta_2 = 1$ recovers our empirical Bayesian benchmark in Equation (2).

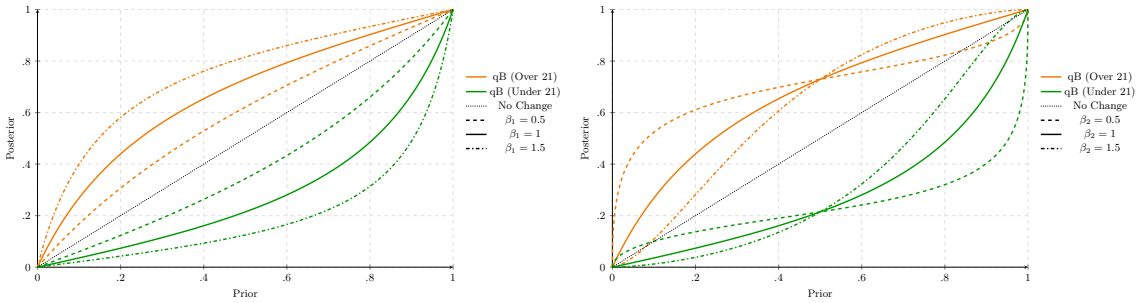
When $\alpha = 0$, qB reduces to the model of Grether (1980), which is one of the most widely applied non-Bayesian updating rules in the behavioral literature. The bias terms, β_1 and β_2, can have two interpretations. In the first interpretation, the DM is in fact Bayesian, but they only have noisy estimates of recommendation informativeness and prior probabilities, and the bias terms capture directions of noise. Under the second interpretation, which is more behavioral, the DM knows the DGP, but fails to apply Bayes' rule appropriately. Typically, $\beta_1 < 1$ captures underweighting of recommendation information or underreaction to news, while $\beta_2 < 1$ captures base-rate neglect.[17]

Figure 6 varies the weight on the recommendation in panel (a) and the direct prior weight in panel (b), holding the other weight at one. Both panels set $\alpha = 0$ and use the same estimated recommendation process, so $\beta_1 = \beta_2 = 1$ gives our Bayesian benchmark.

The estimation of this model is relatively standard, and details can be found in Appendix B.3. We estimate the three updating parameters, α, β_1, β_2, by least squares in capped log odds with nonnegative prior and recommendation coefficients. For this second stage, capping replaces reported priors and posteriors below 0.005 with 0.005 and those above 0.995 with 0.995 before taking log odds. The full-sample estimation gives $\widehat{\alpha} = 0.127$, $\widehat{\beta}_1 = 0.539$, and $\widehat{\beta}_2 = 0.811$. Figure 2(a) shows the Bayesian prediction and Figure 2(b) shows qB, averaged within the same bins as the observed data.

For the estimated parameter values, qB satisfies monotonicity but violates consis-

[17] See Benjamin (2019) and Ortoleva (2024) for discussions of the Grether rule. Appendix B.2 presents the version imposing signal-dependence neglect, its parameter interpretations, and its empirical comparisons.



(a) Varying recommendation weight ($\beta_2 = 1$) (b) Varying prior weight ($\beta_1 = 1$)

Figure 6: Predictions of qB.

tency at sufficiently confirming priors, reactionary updating, and threshold updating.

4.3 Contraction Rule

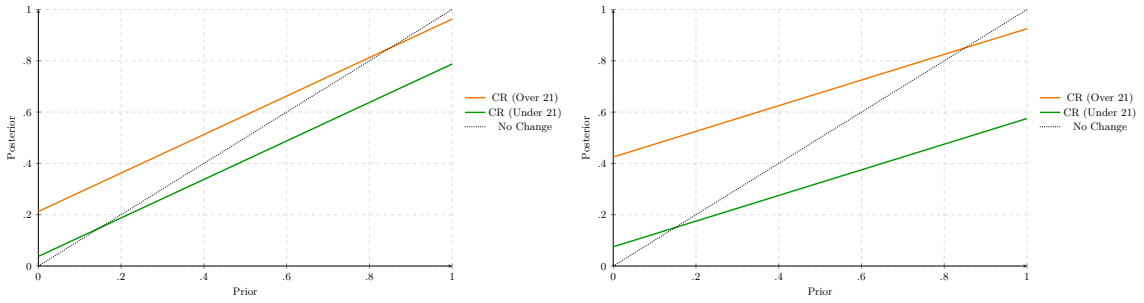
In the Contraction Rule of Ke, Wu, and Zhao (2024), the DM first selects a *representative* belief ρ_R for each recommendation R from $I(R)$, the set of probability distributions that are consistent with the recommendation. This is then mixed with their prior with the weight $\epsilon_R \in [0, 1]$. Hence, the posterior is given by

$$\mu_R = \epsilon_R \, \mu + (1 - \epsilon_R) \, \rho_R. \tag{4}$$

The parameter ϵ_R is the weight on the DM's prior. The corresponding recommendation or AI weight is $1 - \epsilon_R$. We assume that $I(R_o) = \{\pi \in \Delta(S) \mid \pi(o) \in [\frac{1}{2}, 1]\}$ and $I(R_u) = \{\pi \in \Delta(S) \mid \pi(o) \in [0, \frac{1}{2}]\}$. For example, the recommendation of Over 21 is consistent with all possible beliefs of Over 21 from fully uncertain ($\frac{1}{2}$) to fully certain of Over 21 (1).

CR has four free parameters. For each recommendation R_i, these are the prior weight ϵ_{R_i} and the probability of Over 21 assigned by the representative belief ρ_{R_i}. Figure 7 illustrates CR with a common prior weight across recommendations, set to 0.75 in panel (a) and 0.5 in panel (b). Estimating CR in the pooled sample gives $\hat{\epsilon}_{R_o} = 0.70$, $\hat{\epsilon}_{R_u} = 0.77$, $\hat{\rho}_{R_o}(o) = 0.85$, and $\hat{\rho}_{R_u}(o) = 0.14$. Figure 2(c) illustrates these estimates overlaid with the observed average beliefs in the pooled sample.

CR satisfies weak monotonicity for every $\epsilon_R \in [0, 1]$, since the posterior is linear in the prior with nonnegative slope ϵ_R. The posterior is constant when $\epsilon_R = 0$ and strictly increasing when $\epsilon_R > 0$. It can also satisfy reactionary updating on



(a) $\epsilon_{R_i} = 0.75$ for both recommendations, $\rho_{R_o}(o) = 0.73$, $\rho_{R_u}(o) = 0.20$

(b) $\epsilon_{R_i} = 0.5$ for both recommendations, $\rho_{R_o}(o) = 0.73$, $\rho_{R_u}(o) = 0.20$

Figure 7: Predictions of CR.

the contradicting side of each recommendation: after R_o, the signed update is $(1 - \epsilon_{R_o})(\rho_{R_o}(o) - \mu(o))$, which is larger when the prior is lower, and after R_u the analogous signed update is larger when the prior is higher. The same linear contraction, however, can lead CR to violate two of the testable properties. First, when $\epsilon_R < 1$, CR violates consistency at confirming extreme priors unless the representative belief is at the boundary of the state space. When $\epsilon_R = 1$, the posterior equals the prior, so there is no updating and consistency is satisfied. With the estimated parameters, priors above $\rho_{R_o}(o) = 0.85$ are pulled down after R_o, and priors below $\rho_{R_u}(o) = 0.14$ are pulled up after R_u. Second, when $\epsilon_R < 1$, CR violates threshold updating because the posterior equals the prior only at the representative belief, not on an interval of confirming priors. When $\epsilon_{R_o} = \epsilon_{R_u} = 1$, CR predicts no updating everywhere and satisfies the threshold property, although the threshold locations are unidentified. Thus, CR cannot generate no-updating intervals for both recommendations while remaining responsive elsewhere.

4.4 Weighted Inertial Updating

Dominiak, Kovach, and Tserenjigmid (2026) propose and characterize weighted Inertial Updating (wIU) with general information. The DM first selects the belief in the information set closest to their prior, then mixes that belief with the prior. Based on the recommendation "more likely," we assume that $I(R_o) = \{\pi \in \Delta(S) \mid \pi(o) \in [\frac{1}{2}, 1]\}$ and $I(R_u) = \{\pi \in \Delta(S) \mid \pi(o) \in [0, \frac{1}{2}]\}$, as discussed previously. After an Over 21 recommendation, a DM whose prior assigns probability at least $\frac{1}{2}$ to Over 21 selects their prior. A DM whose prior assigns a lower probability selects the boundary of

$I(R_o)$, so $\pi_{R_o}(o) = \frac{1}{2}$. The DM then mixes π_R with their prior to form their posterior. Formally, the posterior μ_R is given by the formula:

$$\mu_R = w\,\mu + (1-w)\pi_R, \tag{5}$$

where $w \in [0,1]$ and π_R solves the minimization problem:

$$\pi_R = \arg\min_{\pi \in I(R)} d_\mu(\pi) = \arg\min_{\pi \in I(R)} -\sum_{s \in S} \mu(s)\,\sigma\left(\frac{\pi(s)}{\mu(s)}\right),$$

where σ is strictly concave and twice differentiable on $(0,\infty)$.[18] We call $1-w$ the *AI weight* because it is the weight placed on the closest belief satisfying the AI recommendation.

The predictions of wIU in our setting can be summarized as follows. The posterior following R_o is

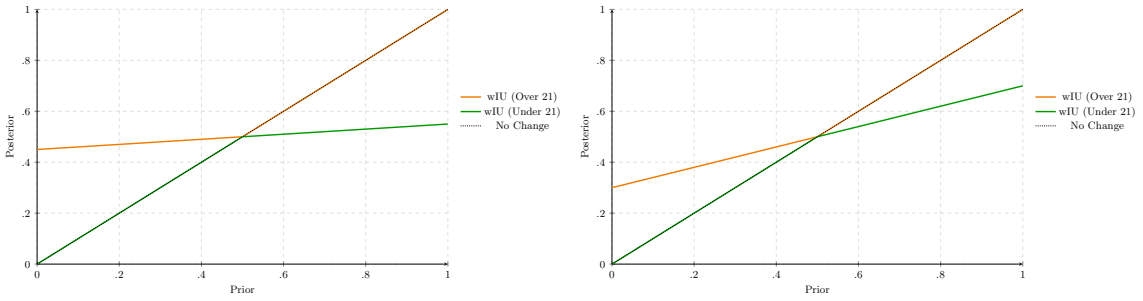
$$\mu_{R_o}(o) = \begin{cases} w\,\mu(o) + (1-w)\,\frac{1}{2} & \text{when } \mu(o) < \frac{1}{2} \\ \mu(o) & \text{when } \mu(o) \geq \frac{1}{2}, \end{cases}$$

while the posterior after R_u is determined similarly.

In words, confirming recommendations do not change beliefs, but when recommendations are disconfirming, the DM takes the closest belief consistent with the recommendation (here 0.5) and mixes it with their prior. Thus, as illustrated in Figure 8, with positive AI weight ($w < 1$), the wIU posterior is a piecewise linear function with a kink at $\frac{1}{2}$.[19] The model has one free parameter, w. Figure 8 shows predictions for $w = 0.1$ and $w = 0.4$, corresponding to AI weights of 0.9 and 0.6. The aggregate estimate is $\hat{w} = 0.44$, with predictions shown in Figure 2(d).

[18] This function d_μ is called an f-divergence in information theory and generalizes many well-known divergences such as KL divergence and α-divergences. In Dominiak, Kovach, and Tserenjigmid (2026), they consider a more general class of divergences and also axiomatically characterize f-divergences. In this experimental setting, the wIU predictions with general divergences match under mild technical conditions. wIU with $w = 0$ nests the pseudo-Bayesian model of Zhao (2022) with KL divergences.

[19] With precise AI advice (i.e., $I(R) = \{\pi\}$), wIU predicts that the posterior is a fixed combination of the prior and the AI prediction (i.e., $\mu_R = w\,\mu + (1-w)\,\pi$). This model then predicts that average posteriors should trace approximately straight lines as a function of the prior for a fixed level of precise AI advice. CR makes the same prediction if the weight on the prior is constant for all AI advice.



(a) AI weight $1 - w = 0.9$ (b) AI weight $1 - w = 0.6$

Figure 8: Predictions of wIU.

While Equations (5) and (4) look similar, there are two distinctions between wIU and CR. First, CR allows the prior weight ϵ_R to depend on each recommendation. Second, in CR, the representative belief ρ_R after recommendation R is independent of the DM's prior, while in wIU, π_R is the closest belief to the DM's prior and it depends on the prior. Hence, the graph for the CR posterior is linear, while the one for wIU is piecewise linear, as illustrated in Figures 2(c) and 2(d). As a result, wIU is not a special case of CR. However, all three models (qB, wIU, and CR) coincide when the DM puts all weight on their prior.

wIU satisfies all four weak updating properties for every $w \in [0, 1]$. Consistency follows because the closest belief in the relevant information set is either the prior itself or the boundary at one half, which lies in the direction of the recommendation. Monotonicity follows from the piecewise-linear form, with slope w outside the information set and 1 inside it. After Over 21, signed updating is $(1-w)\max\{\frac{1}{2} - \mu(o), 0\}$, while after Under 21 it is $(1 - w)\max\{\mu(o) - \frac{1}{2}, 0\}$. Movement therefore weakly increases with contradiction, satisfying reactionary updating. Beliefs remain unchanged on $[\frac{1}{2}, 1]$ after Over 21 and $[0, \frac{1}{2}]$ after Under 21, satisfying threshold updating. When $w = 1$, beliefs remain unchanged for every prior.[20]

[20]The wIU model assumes information is given as a set of distributions, and we operationalize this by assuming that recommendations map to particular sets of distributions. In principle, individuals may have subjective interpretations of each recommendation. For example, an individual might interpret R_o as a stronger recommendation and therefore associate it with the set $\pi(o) \in [\phi_o, 1]$, where $\phi_o > 1/2$. In Appendix B.14, we extend wIU to weighted subjective Inertial Updating (wsIU) and jointly estimate w, ϕ_o, and ϕ_u. This three-parameter extension achieves the highest aggregate completeness in the pooled sample and lowest out-of-sample MSE, outperforming the other three models in all 100 test splits. It is also more restrictive than CR and qB under the monotone reference class.

4.5 Revisiting Non-Updating

Although wIU predicts no updating following confirming recommendations, an important question is whether the high frequency of unchanged reports in our data (78.6%) is consistent with this prediction. To obtain a simple intuition, consider a scenario in which priors and AI recommendations are completely random. In this case, the prior and posterior would be the same under wIU for approximately 50% of observations. However, since priors and AI recommendations are positively correlated, we should observe a higher percentage of non-updating under wIU. Recall that the average accuracy for priors and AI recommendations is 61% and 76%, respectively. For this illustrative calculation, set aside priors of exactly one half and classify each remaining prior according to which state it assigns higher probability. Assuming non-negative correlation between the correctness indicators for this classification and the AI recommendation, marginal accuracies of 61% and 76% imply an agreement rate between 55.7% and 85%. The lower bound corresponds to independent correctness indicators.[21] Hence, the percentage of non-updating observations is well within the wIU prediction, i.e., $78.6\% \in (55.7\%, 85\%)$.

4.6 AI Weight

In CR, $1-\epsilon_R$ measures adjustment toward the recommendation's representative belief. In wIU, $1-w$ measures adjustment toward the closest belief in the recommendation set. The estimates are $1-\epsilon_{R_o} = 0.30$ and $1-\epsilon_{R_u} = 0.23$ for CR and $1-w = 0.56$ for wIU. In qB, $\widehat{\beta}_1 = 0.539$ weights the conditional likelihood ratio and $\widehat{\beta}_2 = 0.811$ weights direct prior odds. Because the likelihood ratio also depends on the prior, their ratio is not a separate measure of AI versus private-information weight. As a result, these quantities are not directly comparable across models. The estimates place positive weight on both prior beliefs and recommendation information.

At the participant level, we estimate wIU separately for each participant to examine how AI weight varies with beliefs about AI accuracy and confidence in the prior. In this model, AI weight, $1-w$, determines the adjustment toward 0.5 when advice contradicts the prior. All 377 participants encounter contradicting advice, allowing

[21] Note that $0.557 \approx 0.76 \times 0.61 + 0.24 \times 0.39$ and $0.85 = 0.61 + (1 - 0.76)$. Hence, the minimum (maximum) level can be achieved by assuming zero (maximum) correlation between the two correctness indicators, given the accuracy levels.

us to estimate a weight for each. The mean estimated weight across participants is 0.438, with 52 estimates of zero. Appendix Figure S3(c) shows the distribution.

The estimated AI weight is higher among participants who believe the AI is more accurate and lower among those who are more confident in their prior beliefs. Table 2 reports these descriptive associations using survey responses collected after the experiment. Controlling for treatment, gender, and age, a ten percentage point increase in perceived AI accuracy is associated with 4.5 percentage points more AI weight ($p < 0.001$). A ten point increase in prior confidence is associated with 3.1 percentage points less AI weight ($p = 0.0013$).

Table 2: Survey correlates of wIU AI weight.

	AI weight
NOINFO treatment	0.0289
	(0.0393)
Woman	-0.0403
	(0.0401)
Older (above median)	0.0131
	(0.0399)
Belief about AI accuracy (0–100%)	0.0045
	(0.0010)
Confidence in own prior (0–100)	-0.0031
	(0.0010)
Mean dep. var.	0.438
N	377

Notes: The dependent variable is estimated AI weight, $1 - w$, from each participant's wIU fit. Entries are OLS coefficients with conventional homoskedastic standard errors in parentheses. The regression includes an intercept and all 377 participants, including 52 with zero estimated weight. The intercept is not displayed. Survey responses were collected after the experiment.

5 Model Assessments

We compare how closely qB, CR, and wIU predict reported posteriors and how restrictive their predictions are.

We assess predictive accuracy using out-of-sample mean squared error (MSE). We also report the completeness measure of Fudenberg, Kleinberg, Liang, and Mul-

lainathan (2022) and the restrictiveness measure of Fudenberg, Gao, and Liang (2026). Completeness measures the share of the in-sample MSE improvement over no updating captured by a model, relative to unrestricted cell means. Restrictiveness measures how poorly a model fits synthetic updating rules relative to no updating.

5.1 Out-of-Sample Performance

Our preregistered way to evaluate model fit is via out-of-sample performance. In order to do so, we created 100 random splits of the dataset. For each split, 70% of the observations were used to estimate model parameters. The remaining 30% of observations were used to test the predictions of the estimated model.

We estimate qB by minimizing squared errors in capped log odds, whereas CR and wIU minimize probability MSE. All three models are evaluated using probability MSE. The comparison therefore evaluates models together with their estimation procedures, rather than isolating differences in functional form.

CR and wIU have similar mean test MSEs and both have lower MSE than qB in every split. Table 3 reports MSE averaged across the test splits and the fraction of splits in which each model has the lowest MSE. wIU has the lowest MSE among the three models in 76% of the splits.

5.2 Completeness and Restrictiveness

For completeness, model parameters and the unrestricted cell means are fitted and scored on the same observations. We calculate restrictiveness under two synthetic reference classes. Both require recommendation consistency, and one also imposes weak monotonicity. Higher restrictiveness indicates that a model fits these synthetic rules less well relative to no updating. Table 3 reports both measures.[22]

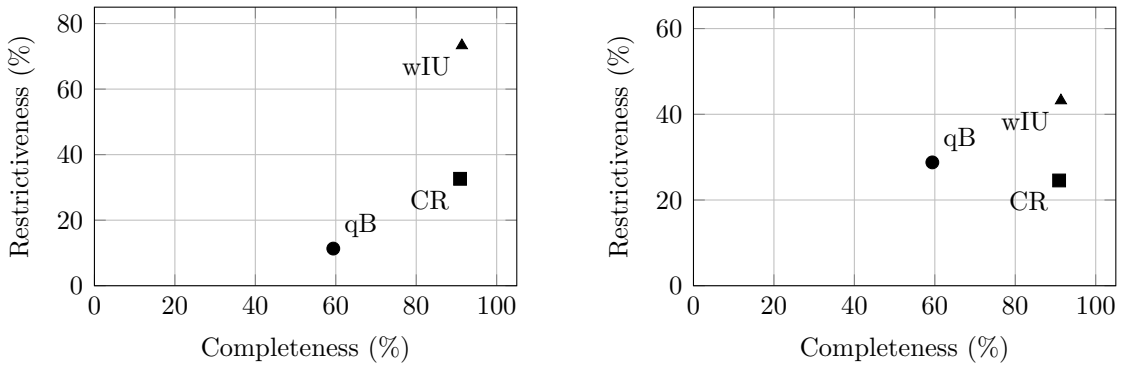
CR and wIU capture more than 90% of the attainable in-sample MSE improvement over no updating, compared with 59.4% for qB. Figure 9 plots completeness against restrictiveness under the two synthetic reference classes. With monotonicity imposed, wIU is the most restrictive, followed by CR and qB. Without imposing monotonicity, wIU remains the most restrictive, followed by qB and CR.

[22] Appendix B.1 defines the measures and describes their calculation.

Table 3: Aggregate model fit and flexibility comparisons.

Model	Test MSE	Test wins	Completeness	Restrictiveness	
				Monotone	No monotonicity
qB	0.0402	0.0%	59.4%	11.3%	28.8%
CR	0.0336	24.0%	90.9%	32.6%	24.6%
wIU	0.0336	76.0%	91.4%	73.3%	43.2%

Notes: Test MSE is probability MSE averaged over 100 random 70/30 training/test splits. Test wins is the share of splits with the lowest MSE. For these comparisons, qB is fitted in capped log odds, whereas CR and wIU minimize probability MSE. Completeness is the share of the in-sample MSE improvement over no updating captured relative to unrestricted prior-recommendation cell means. Restrictiveness compares model loss with no updating across 100 common synthetic rules per reference class. Both classes impose recommendation consistency. The Monotone class also imposes monotonicity. Completeness and restrictiveness use probability MSE, with qB's estimated recommendation process held fixed. Higher completeness indicates better fit, and higher restrictiveness indicates less flexibility.



(a) Monotonic eligible class.

(b) Recommendation-consistent eligible class.

Figure 9: Completeness and restrictiveness of qB, CR, and wIU.

5.3 Updating Types

The aggregate model comparisons mask heterogeneity in updating behavior. We therefore also compare models using data for each participant. All comparisons use the common sample of 344 participants defined in Appendix B.3.

For each participant, we use 100 common random splits, estimating all three models on 70% of observations and evaluating their predictions against the original posterior beliefs on the remaining 30%. Both the recommendation coefficients $(\gamma_0, \gamma_1, \gamma_2)$ and the qB updating parameters are estimated separately for each participant using only their training observations. We calculate each participant's mean test MSE over the 100 splits. Appendix B.3 reports the unavailable fits and estimation

procedures.

Table 4: Model fit among classified participants.

Model	Free parameters	Mean MSE	Median MSE
qB	4	0.0276	0.0144
CR	4	0.0174	0.0127
wIU	1	0.0243	0.0164

Notes: MSE is averaged over each participant's 100 splits and then summarized across the 344 participants. Both stages of qB estimation use participant-specific training observations. The exclusion leaves the validation splits and CR and wIU fits unchanged.

CR has lower mean MSE across participants than qB and wIU (both comparisons $p < 0.001$). The mean difference between qB and wIU is 0.003354 (95% CI $[-0.000115, 0.006823]$, $p = 0.058$).

Mean MSE reflects the size of prediction errors, whereas winner shares count each participant once regardless of the size of the difference. Table 5 reports classifications by lowest mean test MSE for the three models together and for each head-to-head comparison with qB. Among the three models, CR is the most frequent classification (37.5%), followed by qB (32.8%) and wIU (29.7%).[23] In the head-to-head comparisons, qB has lower mean test MSE for 51.5% of participants against wIU and 41.9% against CR. Appendix Table S10 reports the three-model classifications separately for INFO and NOINFO. The share classified as qB is 35.5% in INFO and 30.2% in NOINFO, a difference that is not statistically significant ($p = 0.348$).

Table 5: Model classification by lowest MSE.

Model	Three models	qB vs. wIU	qB vs. CR
qB	32.8%	51.5%	41.9%
CR	37.5%	—	58.1%
wIU	29.7%	48.5%	—

Notes: Each participant's MSE is averaged across 100 common validation splits. The three-model comparison includes qB, CR, and wIU. Head-to-head columns compare only the two named models.

Using separate fits to each participant's full sample that minimize probability MSE, mean completeness across participants is 41.9% for qB, 47.6% for CR, and 35.7% for wIU.

[23] Participant classifications depend on the fitting objective. When qB is estimated by probability MSE with the same capped endpoints, its classification share rises to 37.8%, making it the most frequent classification (Appendix Table S2).

6 Discussion

6.1 Related Literature

Our paper contributes to the experimental literature on non-Bayesian updating by providing evidence from a controlled experiment in which information is payoff relevant, but the information structure is not transparent to participants.[24] This distinguishes our work from recent papers in which the DGP is ambiguous or uncertain, but where the DM is told a set of possible signal structures, source accuracies, or interpretive models (Liang, 2025; Shishkin and Ortoleva, 2023; Aina and Schneider, 2025; Farina and Herman, 2026) and work on interpreting ambiguous evidence (Epstein and Halevy, 2024; Fryer, Harms, and Jackson, 2019) or updating ambiguous or uncertain prior beliefs (Ngangoué, 2021; Chan, Charness, Dave, and Reddinger, 2024; Cohen, Gilboa, Jaffray, and Schmeidler, 2000). Reactions to an unknown DGP are studied in an experiment by Afrouzi, Kwon, Landier, Ma, and Thesmar (2023), where participants observe a series of 40 past realizations from a simple AR(1) process and are then asked to make forecasts for future values without knowing the DGP. They find overreaction to recent observations. Kelley and Friedman (2002) study price forecasting from a stationary stochastic process whose functional form and coefficients are not disclosed to participants.

Esponda, Vespa, and Yuksel (2024) also include a *NoPrimitives* treatment that withholds both the base rate and signal accuracy. Participants learn from repeated feedback about states and signals, with final beliefs closer to the Bayesian benchmark than when these probabilities are disclosed. Our focus is instead on how participants revise a visually formed prior after a qualitative AI recommendation.

In addition, we contribute to a growing literature on the interaction between humans and AI by studying how humans update their beliefs after receiving AI recommendations. Our focus is on systems that communicate through qualitative recommendations or alerts about an underlying state without reporting a probability for that state. However, there is limited evidence on how decision makers update their beliefs based on the general information conveyed by qualitative AI recommendations.

Our work is related to papers that consider the impacts of AI advice on beliefs

[24] In a sequential experiment with a known signal structure, Sarnoff (2026) finds that a minority of participants overreact to conflicting signals. Our experiment instead examines updating from a visually formed prior after a single qualitative AI recommendation.

(Agarwal, Moehring, Rajpurkar, and Salz, 2023; Caplin, Deming, Li, Martin, Marx, Weidmann, and Ye, 2026), and the design of coarsened recommendations (Hoong and Dreyfuss, 2025). Agarwal, Moehring, Rajpurkar, and Salz (2023) study how professional radiologists combine their own information with AI predictions, and they find that AI predictions do not improve radiologist performance on average. They explain the failure to benefit from AI assistance as arising from errors in belief updating: radiologists underweight AI predictions and neglect the statistical dependence between their own information and the AI prediction. Our experiment complements their study of precise AI advice with evidence on qualitative recommendations, although differences in tasks and populations prevent isolating the role of advice format. Their analysis of belief updating focuses on variants of Bayesian updating, while we compare qB to CR and wIU, models designed for general or black box information. Their revised policy analysis does not depend on that parametric specification. Hoong and Dreyfuss (2025) derive conditions under which coarsening AI predictions improves decisions and show that a calibrated binary recommendation improves loan specialists' decisions relative to an uncoarsened AI score.

Our results also relate to Caplin, Deming, Li, Martin, Marx, Weidmann, and Ye (2026), who use the Bouncer task to study who benefits from AI assistance. They show that AI increases payoffs more for low-ability participants, but that conditional on ability, the value of AI assistance depends on whether participants are calibrated about their own ability. Our results suggest a belief-updating channel for this finding. If participants who are more confident in their own assessments place more weight on their priors, then miscalibration can reduce the value of AI even when the AI is informative. Consistent with this interpretation, our wIU estimates suggest that AI weight is increasing in beliefs about AI accuracy and decreasing in confidence in their own prior. Thus, overconfidence may matter twice: first by affecting the initial prior, and second by affecting the weight placed on that prior after observing the AI recommendation.

6.2 Behavioral Microfoundations

CR and wIU provide parsimonious descriptions of updating that admit several behavioral interpretations. Inertia, attention to contradiction, and interpretive uncertainty offer three possible explanations. First, inertia with qualitative information maps di-

rectly into wIU: DMs may select the closest belief satisfying the information constraint and then mix that intermediate belief with their prior (Dominiak et al., 2026). Second, contradictory recommendations may attract attention and contribute to larger average updates. Limited attention and cognitive imprecision can distort perceived information (Ba et al., 2026). This possible interpretation connects to cue weighting and surprise (Lynch and Ofir, 1989; Friston, 2008). We find longer response times following contradicting recommendations, which is consistent with greater deliberation, although the measure also includes the time needed to revise a report (see Appendix Figure S4). Third, uncertainty about how to interpret information may shape updating. Enke and Graeber (2023) associate cognitive uncertainty with compression of reported probabilities toward one half. Ortoleva (2012) models reassessment of prior beliefs following sufficiently unexpected news. Chan, Charness, Dave, and Reddinger (2024) manipulate confidence in priors by varying how long participants view a black-and-white grid representing the prior probability. Longer viewing increases confidence and the weight placed on priors relative to subsequent signals. Limited sensitivity to signal strength and difficulties with contingent thinking may help explain INFO's limited impact (Augenblick et al., 2025; Aina et al., 2023).

6.3 Human vs. AI Recommendations

Comparing human and AI advice raises two questions. Do people respond differently to identical advice attributed to different sources? And do the sources produce different advice?

There is a large existing literature on how humans respond to human recommendations. For example, De Filippis, Guarino, Jehiel, and Kitagawa (2022) experimentally study updating from a predecessor's belief and a private signal, while Chen (2026) theoretically studies social learning when signal distributions are ambiguous. The review by Bonaccio and Dalal (2006) discusses the tendency to discount advice relative to one's own opinion, while Sniezek and Van Swol (2001) study the roles of trust, confidence, and expertise.[25]

On the first question, Vodrahalli, Daneshjou, Gerstenberg, and Zou (2022) use

[25]Our design shares the basic judge-advisor structure. Many tasks in that literature elicit continuous point estimates rather than probabilities over a common binary state, although the literature also includes binary choices and confidence judgments. Our elicitation of both prior and posterior probabilities permits direct comparison of updating rules in this setting.

deception around the stated source of advice, and distinguish whether people use advice (activation) from how much they adjust conditional on using it (integration). The source label affects activation, but they find no statistically significant difference in integration conditional on activation. In their four layperson tasks, participants first reported a continuous belief about the correct state, received precise advice experimentally labeled as either human or AI, and then reported a posterior belief. For these layperson tasks the underlying advice was held fixed across the human- and AI-labeled conditions: the comparison therefore identifies the effect of the perceived advice source rather than differences in the actual production of advice.

The second question concerns differences in the advice itself. In our setting, participants' initial judgments and AI recommendations are related but not highly correlated. Defining the participant's binary judgment as whether the prior belief is above 50%, the overall correlation between the participant's binary judgment and the AI recommendation is 0.233, with agreement in 62.5% of observations.[26] Among the 369 participants with defined correlations, the median is 0.266 and the 95th percentile is 0.434. Only five participants have correlations above 0.5, with a maximum of 0.762. Thus, even the participants most aligned with AI are not simply reproducing the AI recommendation in binary form.

This distinction matters for future comparisons. Human and AI sources may produce different recommendations, and participants' beliefs about those differences may affect how they use advice. Deception about the source of advice, or selecting human advisors whose recommendations mimic the AI, can hold the recommendation fixed across perceived sources. However, an advisor selected to resemble the AI need not be representative of the human advisors participants expect.

A related question is whether responses to human and AI advice differ by participant gender. Survey evidence indicates that men adopt generative AI more often than women (Stephany and Duszynski, 2026), although adoption is distinct from the weight placed on advice. In our data, Appendix Figure S1(d) shows smaller average updating at extreme contradicting priors among participants who self-reported being women. See et al. (2011) find that women receive higher coworker ratings of advice taking than men in a study that does not compare responses to human and AI recommendations.

[26] Exact 50% priors are coded as not Over 21 for this calculation. Excluding them gives an overall correlation of 0.234.

7 Conclusion

We study how individuals update their beliefs after receiving qualitative AI recommendations when the underlying data-generating process is not fully specified. Our experiment documents three systematic behavioral patterns: average updating is close to zero when recommendations confirm extreme priors, larger when recommendations contradict extreme priors, and smaller at intermediate priors. Motivated by these patterns, we examine four properties of belief updating: consistency, monotonicity, reactionary updating, and threshold updating. Our aggregate tests detect no violations of the first three properties across the pooled sample and reported splits, while updating following strongly confirming recommendations is generally statistically indistinguishable from zero. At the participant level, 77.5% meet all four empirical criteria, and 15.6% meet exactly three, although there is considerable heterogeneity in updating behavior. The estimated aggregate updating curves are similar across INFO and NOINFO.

We also compare three models of belief updating: quasi-Bayesian updating (qB), the Contraction Rule (CR), and weighted Inertial Updating (wIU). In the pooled sample, CR and wIU outperform qB in out-of-sample prediction. CR and wIU capture more than 90% of the attainable in-sample MSE improvement over no updating. Moreover, wIU is substantially more restrictive than both CR and qB under the two benchmarks we consider. These findings suggest that updating rules incorporating contraction toward recommendations or inertia under confirming information provide useful descriptions of how individuals respond to qualitative information.

Our results do not identify a unique behavioral microfoundation. In particular, our binary design with two qualitative recommendations assesses selected implications of wIU and CR, not their full axiomatic foundations, which require richer variation in information sets. Our analysis focuses on monotonicity, movement toward beliefs consistent with the recommendation, linear contraction under CR, and threshold inertia under wIU.

The central asymmetry is between little updating after strongly confirming recommendations and larger adjustments after strongly contradicting recommendations. Capturing that asymmetry helps explain the aggregate performance of CR and wIU.

References

Afrouzi, H., S. Y. Kwon, A. Landier, Y. Ma, and D. Thesmar (2023): "Overreaction in expectations: Evidence and theory," *The Quarterly Journal of Economics*, 138, 1713–1764.

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2023): "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," NBER Working Paper 31422, National Bureau of Economic Research, revised August 12, 2026.

Aina, C., A. Amelio, and K. Brütt (2023): "Contingent belief updating," ECONtribute Discussion Paper 263, ECONtribute, revised June 16, 2026.

Aina, C. and F. H. Schneider (2025): "Weighting Competing Models," CESifo Working Paper 11757, CESifo.

Augenblick, N., E. Lazarus, and M. Thaler (2025): "Overinference from weak signals and underinference from strong signals," *The Quarterly Journal of Economics*, 140, 335–401.

Ba, C., J. A. Bohren, and A. Imas (2026): "Over- and Underreaction to Information: Belief Updating with Cognitive Constraints," Working paper, July 2026. Accepted at the Quarterly Journal of Economics.

Benjamin, D. J. (2019): "Errors in probabilistic reasoning and judgment biases," in *Handbook of Behavioral Economics: Applications and Foundations 1*, ed. by B. D. Bernheim, S. DellaVigna, and D. Laibson, North-Holland, vol. 2, 69–186.

Bonaccio, S. and R. S. Dalal (2006): "Advice Taking and Decision-Making: An Integrative Literature Review, and Implications for the Organizational Sciences," *Organizational Behavior and Human Decision Processes*, 101, 127–151.

Caplin, A., D. J. Deming, S. Li, D. Martin, P. Marx, B. Weidmann, and K. J. Ye (2026): "The ABCs of Who Benefits from Working with AI: Ability, Beliefs, and Calibration," *Management Science*, 72, 5843–5852.

Caplin, A. and D. Martin (2015): "A testable theory of imperfect perception," *The Economic Journal*, 125, 184–202.

Chan, K., G. Charness, C. Dave, and J. L. Reddinger (2024): "On Prior Confidence and Belief Updating," *arXiv preprint arXiv:2412.10662*, revised May 2025.

Charness, G., R. Oprea, and S. Yuksel (2021): "How Do People Choose

Between Biased Information Sources? Evidence from a Laboratory Experiment," *Journal of the European Economic Association*, 19, 1656–1691.

Chen, J. Y. (2026): "Sequential Learning under Informational Ambiguity," *American Economic Review*, 116, 209–245.

Cohen, M., I. Gilboa, J.-Y. Jaffray, and D. Schmeidler (2000): "An experimental study of updating ambiguous beliefs," *Risk, Decision and Policy*, 5, 123–133.

Coutts, A. (2019): "Good News and Bad News Are Still News: Experimental Evidence on Belief Updating," *Experimental Economics*, 22, 369–395.

Danz, D., L. Vesterlund, and A. J. Wilson (2022): "Belief Elicitation and Behavioral Incentive Compatibility," *American Economic Review*, 112, 2851–2883.

De Filippis, R., A. Guarino, P. Jehiel, and T. Kitagawa (2022): "Non-Bayesian Updating in a Social Learning Experiment," *Journal of Economic Theory*, 199, 105188.

Dominiak, A., M. Kovach, and G. Tserenjigmid (2026): "Inertial Updating with General Information," *arXiv preprint arXiv:2502.00958v2*, October 6, 2026 revision.

Dreyfuss, B. and R. Raux (2026): "Human Learning about AI," Working paper, arXiv:2406.05408.

Enke, B. and T. Graeber (2023): "Cognitive uncertainty," *The Quarterly Journal of Economics*, 138, 2021–2067.

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

Esponda, I., E. Vespa, and S. Yuksel (2024): "Mental Models and Learning: The Case of Base-Rate Neglect," *American Economic Review*, 114, 752–782.

Farina, A. and C. Herman (2026): "Communicating with Data-Generating Processes: An Experimental Analysis," *Available at SSRN 6482578*.

Friston, K. (2008): "Hierarchical models in the brain," *PLoS Computational Biology*, 4, e1000211.

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

Fudenberg, D., W. Gao, and A. Liang (2026): "How Flexible Is That Functional

Form? Quantifying the Restrictiveness of Theories," *The Review of Economics and Statistics*, 108, 194–209.

Fudenberg, D., J. Kleinberg, A. Liang, and S. Mullainathan (2022): "Measuring the completeness of economic models," *Journal of Political Economy*, 130, 956–990.

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

Holt, C. A. and A. M. Smith (2009): "An Update on Bayesian Updating," *Journal of Economic Behavior & Organization*, 69, 125–134.

Hoong, R. and B. Dreyfuss (2025): "Calibrated Coarsening: Designing Information for AI-Assisted Decisions," Working Paper 7407-25, MIT Sloan School of Management, previously circulated as "Improving AI-Assisted Decision-Making Through Calibrated Coarsening".

Hossain, T. and R. Okui (2013): "The binarized scoring rule," *Review of Economic Studies*, 80, 984–1001.

Ke, S., B. Wu, and C. Zhao (2024): "Learning from a black box," *Journal of Economic Theory*, 221, 105886.

Kelley, H. and D. Friedman (2002): "Learning to Forecast Price," *Economic Inquiry*, 40, 556–573.

Liang, Y. (2025): "Learning from Unknown Information Sources," *Management Science*, 71, 3873–3890, published online August 22, 2024.

Lynch, Jr., J. G. and C. Ofir (1989): "Effects of cue consistency and value on base-rate utilization," *Journal of Personality and Social Psychology*, 56, 170–181.

McKelvey, R. D. and T. Page (1990): "Public and Private Information: An Experimental Study of Information Pooling," *Econometrica*, 58, 1321–1339.

Molavi, P. (2026): "The Empirical Content of Bayesian Updating under Misspecification," Working paper, March 2026. Previously circulated as "The Empirical Content of Bayesianism".

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

Ortoleva, P. (2012): "Modeling the change of paradigm: Non-Bayesian reactions to unexpected news," *American Economic Review*, 102, 2410–2436.

——— (2024): "Alternatives to Bayesian Updating," *Annual Review of Economics*,

16, 545–570.

Rothe, R., R. Timofte, and L. Van Gool (2018): "Deep expectation of real and apparent age from a single image without facial landmarks," *International Journal of Computer Vision*, 126, 144–157.

Sarnoff, K. (2026): "The Structure of Sequential Updating," Working paper, February 26.

Savage, L. J. (1971): "Elicitation of Personal Probabilities and Expectations," *Journal of the American Statistical Association*, 66, 783–801.

See, K. E., E. W. Morrison, N. B. Rothman, and J. B. Soll (2011): "The detrimental effects of power on confidence, advice taking, and accuracy," *Organizational Behavior and Human Decision Processes*, 116, 272–285.

Shishkin, D. and P. Ortoleva (2023): "Ambiguous Information and Dilation: An Experiment," *Journal of Economic Theory*, 208, 105610.

Shmaya, E. and L. Yariv (2016): "Experiments on Decisions under Uncertainty: A Theoretical Framework," *American Economic Review*, 106, 1775–1801.

Smith, C. A. B. (1961): "Consistency in Statistical Inference and Decision," *Journal of the Royal Statistical Society: Series B (Methodological)*, 23, 1–25.

Sniezek, J. A. and L. M. Van Swol (2001): "Trust, Confidence, and Expertise in a Judge-Advisor System," *Organizational Behavior and Human Decision Processes*, 84, 288–307.

Stephany, F. and J. Duszynski (2026): "Women Worry, Men Adopt: How Gendered Perceptions Shape the Use of Generative AI," *arXiv preprint arXiv:2601.03880v1*, January 7, 2026.

Vodrahalli, K., R. Daneshjou, T. Gerstenberg, and J. Zou (2022): "Do Humans Trust Advice More if it Comes from AI?: An Analysis of Human-AI Interactions," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, New York, NY, USA: Association for Computing Machinery, AIES '22, 763–777.

Zhao, C. (2022): "Pseudo-Bayesian updating," *Theoretical Economics*, 17, 253–289.