Supplemental Appendix

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

Matthew Kovach, Daniel Martin, and Gerelt Tserenjigmid

A Preregistration Report

AsPredicted #228821, filed before data collection, identified belief updating after imprecise AI advice as the primary question. Treatment differences and participant heterogeneity were secondary questions. The outcomes were reported probabilities of Over 21 before and after advice. The repeated age classification task, Baseline and AI Info treatments (NOINFO and INFO here), and exclusion of participants who failed attention checks followed the plan. Recruitment stopped when the $5,000 research budget was exhausted, yielding 188 INFO and 189 NOINFO participants rather than the target of 200–300 per treatment. We also excluded 68 timed-out rounds (0.11%), an exclusion that was not preregistered.

The plan called for aggregate and participant-level comparisons on held-out observations of correlation-neglect models, Grether models with belief biases, Inertial Updating, prior-weighted Inertial Updating, and robustness variants. We report qB, CR, and wIU, together with our Bayesian benchmark and a fixed-likelihood alternative. Appendix B.14 also examines weighted subjective Inertial Updating (wsIU). CR and wsIU are additional comparisons. The planned performance measures were accuracy, distance from truth, MSE, and cross-entropy. We emphasize posterior MSE because we compare predictions of reported beliefs rather than the true state, and report binary accuracy descriptively. Additional comparisons use completeness and restrictiveness.

The planned heterogeneity analyses covered confidence in AI and oneself, calibration, age, gender, and response time. Related analyses appear in Table 1, Sections 5 and 4.6, and Appendix B.13. Prior MSE reflects both calibration and informativeness, so the accuracy-based comparisons do not separately implement the planned calibration analysis. Planned secondary analyses of aggregate updating, survey beliefs about AI accuracy, human and AI performance, dynamics, and difficult images appear in Sections 3–5, Figures 5 and S1, and Appendices B.10 and B.13.

Other additions motivated by the observed patterns and model comparisons are the four property tests and inference clustered by participant and image. The threshold criterion and its fixed confirming ranges were specified after examining the data, not preregistered.

B Additional Analyses

B.1 Completeness and Restrictiveness

For prior p_i, posterior q_i, and recommendation R_i, define $e_P(f) = N^{-1}\sum_i [q_i - f(p_i, R_i)]^2$. Let $\overline{\mathcal{F}}$ contain all deterministic updating rules and $\mathcal{F}_\Theta \subseteq \overline{\mathcal{F}}$ be a model class. No updating, $f_0(p, R) = p$, is nested in all three models. The unrestricted fit $f^* = \arg\min_{f \in \overline{\mathcal{F}}} e_P(f)$ gives the mean posterior at each exact prior and recommendation. Following Fudenberg, Kleinberg, Liang, and Mullainathan (2022), completeness is

$$\kappa(\mathcal{F}_\Theta) = \frac{e_P(f_0) - \min_{f_\theta \in \mathcal{F}_\Theta} e_P(f_\theta)}{e_P(f_0) - e_P(f^*)}.$$

Models and cell means are fitted and scored on the same observations with equal observation weights. Participant averages use the common sample in Appendix B.3, omitting normalization denominators at most 10^{-12}.

For restrictiveness (Fudenberg, Gao, and Liang, 2026), we use two eligible classes $\mathcal{F}$. Both require $f(p, R_o) \in [p, 1]$ and $f(p, R_u) \in [0, p]$, and one also imposes weak monotonicity. Priors $0, 0.01, \ldots, 1$ give 202 equally weighted cells. Let $\lambda_\mathcal{F}$ be uniform on the free posterior coordinates of this grid and d be grid MSE. Then

$$r(\mathcal{F}_\Theta, \mathcal{F}) = \frac{\mathbb{E}_{\lambda_\mathcal{F}}[d(\mathcal{F}_\Theta, f)]}{\mathbb{E}_{\lambda_\mathcal{F}}[d(f_0, f)]}, \qquad d(\mathcal{F}_\Theta, f) = \inf_{f_\theta \in \mathcal{F}_\Theta} d(f_\theta, f).$$

We use 100 common draws per class and a ratio of mean losses. Without monotonicity, posteriors are drawn independently from their eligible intervals. Monotone Over curves use 100 sorted uniform draws on $[0, 1]$, rejecting vectors below the corresponding priors $0, 0.01, \ldots, 0.99$ and appending a posterior of one. Independent reflected draws give Under curves. We use MATLAB's Twister generator with seed 12345.

For both measures, models minimize probability MSE using raw beliefs. For participant-level completeness, we estimate each participant's recommendation process using all of their observations and hold it fixed while fitting qB by probability

MSE. For restrictiveness, we instead hold the pooled recommendation process fixed while fitting qB to each synthetic updating rule.

B.2 qB With Signal-Dependence Neglect

For a fixed recommendation process $\delta : S \to \Delta(\mathcal{R})$, conditional independence of recommendations and private visual information given the state yields

$$\mu_R(o) = \frac{\delta(R \mid o)\mu(o)}{\delta(R \mid o)\mu(o) + \delta(R \mid u)\mu(u)}. \tag{S1}$$

Unlike our Bayesian benchmark, this specification imposes signal-dependence neglect by using state-conditional rates without conditioning on private information. INFO supplies these state-conditional rates.

The corresponding qB alternative is

$$\mu_R(o) = \frac{\exp(\alpha)\,\mu(o)^{\beta_2}\delta(R \mid o)^{\beta_1}}{\exp(\alpha)\,\mu(o)^{\beta_2}\delta(R \mid o)^{\beta_1} + \mu(u)^{\beta_2}\delta(R \mid u)^{\beta_1}}, \qquad \beta_1, \beta_2 \geq 0. \tag{S2}$$

The image-sample rates $\delta(R_o \mid o) = 0.825$ and $\delta(R_o \mid u) = 0.300$, with complementary rates for R_u, give likelihood ratios 2.75 and 0.25. Using primary qB's response restrictions, capped log-odds estimation, raw-prior prediction, and original posterior targets gives full-sample estimates $\widehat{\alpha} = 0.139$, $\widehat{\beta}_1 = 0.507$, and $\widehat{\beta}_2 = 0.780$.

Setting $\alpha = 0$ gives the updating rule of Grether (1980), with β_1 and β_2 distorting the recommendation and prior terms. The term *automation neglect* for $\beta_1 < \beta_2$ (Agarwal, Moehring, Rajpurkar, and Salz, 2023) describes the relative weights in this rule rather than an identified mechanism. In primary qB, the prior also enters the advice term, so the same interpretation does not apply.

With fixed positive likelihoods, $\beta_2 \geq 0$ implies monotonicity. When $\beta_2 > 0$, the rule preserves endpoint priors and any nonzero absolute update peaks at an interior prior. For ordinary Bayes with fixed likelihood ratio $L \neq 1$, the peak is $p = 1/(1+\sqrt{L})$, or 0.376 after Over and 0.667 after Under. These peak locations are specific to fixed likelihoods and do not apply to the prior-dependent benchmark. No-updating intervals of positive length for both recommendations require no updating throughout both prior ranges. The fitted rule violates consistency at sufficiently confirming extreme priors.

Table S1 compares the two estimation procedures on common validation observations. The comparison includes differences beyond likelihood dependence. Primary qB estimates its process from training data, whereas fixed-likelihood qB uses rates from the image sample. The procedures also differ in their numerical treatment of nearly zero or unidentified slopes (Appendix B.3).

Appendix Table S1: Primary qB and qB with signal-dependence neglect.

Specification	Aggregate test MSE	Mean participant test MSE
Primary qB	0.04016	0.02763
qB with signal-dependence neglect	0.04023	0.02629

Notes: Each column uses 100 common 70/30 validation splits. Participant errors are first averaged within participant, then across the common sample of 344 participants defined in Appendix B.3. Both rules use capped log-odds estimation, nonnegative response slopes, raw-prior prediction, and original posterior reports for evaluation. For participant comparisons, primary qB estimates both stages separately from each participant's training observations. Fixed-likelihood qB retains its original fits and has no estimated first-stage coefficients. The aggregate fits are unchanged.

B.3 Primary qB Estimation

Estimation has two stages. We first estimate the recommendation process in Equation (1) without using posterior reports. We then fit the three updating parameters, α, β_1, β_2, by constrained least squares in log odds. Let $\ell(x) = \log[x/(1-x)]$ and $x^c = \min\{\max\{x, 0.005\}, 0.995\}$. The response equation is

$$\ell(\mu_R(o)^c) = \alpha + \beta_2 \ell(\mu(o)^c) + \beta_1 \log \frac{\widehat{\delta}(R \mid o, \mu)}{\widehat{\delta}(R \mid u, \mu)} + e, \qquad \beta_1, \beta_2 \geq 0. \tag{S3}$$

The cap makes log odds finite and changes only endpoint reports on the one percentage point slider grid. It applies to the direct prior and posterior terms, while recommendation probabilities are estimated and evaluated using raw priors. Predictions also use raw priors and are scored against original posterior reports. Appendix B.5 checks robustness to excluding endpoint priors while retaining every posterior report.

Related studies use different conventions. Enke and Graeber (2023) replace posterior endpoints with 0.01 and 0.99 in log-odds regressions, Agarwal, Moehring, Rajpurkar, and Salz (2023) cap probability assessments at 0.05 and 0.95, and Coutts (2019) excludes endpoint observations in the main specification and examines interior

replacement as a robustness check.

For participant-level model comparisons, we exclude the 32 never-updaters because all three models can reproduce their unchanged reports. One additional participant is excluded because the regression estimating their recommendation probabilities does not yield valid estimates in the training samples. The primary comparisons use the resulting common sample of 344 participants and the original validation splits. Never-updaters remain in the aggregate analyses and updating-property tests.

Separation can prevent a finite maximum-likelihood estimate of the recommendation process, while rank deficiency prevents separate identification of its coefficients. We require valid estimates in all 100 training samples and do not replace unavailable fits with pooled or penalized estimates.

Both stages use the same training observations in each of 100 splits for the aggregate analysis. For each retained participant, all models use the same 100 random 70/30 splits. In each replicate, we estimate $(\gamma_0, \gamma_1, \gamma_2)$ and the qB updating parameters separately for each participant using only their training observations. Full-sample fits also estimate both stages separately for each participant. Predictions use priors and recommendations, not test states. CR, wIU, and wsIU retain their original fits and are summarized on the same participants. The aggregate recommendation regressions remain pooled and have full rank and finite, converged likelihood estimates.

Constrained least squares in log odds checks all four combinations of zero or nonnegative slopes, each with a free intercept. Slopes below 10^{-12} in absolute value are treated as numerical zeros. Regressors that are constant in the estimation sample receive zero slopes. When multiple slope estimates give the same fitted values within a case, we select those with the smallest sum of squared slopes. A positive direct-prior coefficient preserves raw endpoints, whereas a zero coefficient need not. When $\beta_2 = 0$, we omit the direct prior term, including at endpoint priors. The prior can still affect predictions through the recommendation likelihood ratio. When the recommendation process is estimable, training samples with no updates receive identity parameters $(\alpha, \beta_2, \beta_1) = (0, 1, 0)$ and predict the raw prior exactly.

Probability fits with MSE use several starting values, including the log-odds estimates and the no-updating rule, and consider cases with coefficients fixed at zero. Fits using raw priors compare a positive direct-prior coefficient, bounded below by 10^{-8}, with a coefficient of zero.

B.4 qB Fitting Objective

Appendix Table S2 compares participant-level qB fitted by minimizing squared log-odds errors with qB fitted by minimizing probability MSE. The direct comparison caps training priors and posteriors at 0.005 and 0.995, while the raw-training alternative uses original beliefs. Both retain an unrestricted intercept and nonnegative prior and recommendation coefficients, predict using raw priors, and score original posterior reports. CR and wIU fits remain fixed.

Table S2 uses the common sample of 344 participants defined in Appendix B.3 and their original 100 random 70/30 splits. In each replicate, every estimator uses the same recommendation process fitted separately to that participant's training observations. These comparisons of estimation procedures show that changing the fitting objective raises qB's winner share from 32.8% to about 38%. Using raw training beliefs also relaxes the capping convention and further reduces qB prediction errors.

Appendix Table S2: Participant-level comparisons by qB fitting objective.

qB estimator	N	Mean qB MSE	Lowest MSE (%)		
			qB	CR	wIU
Log odds	344	0.0276	32.8%	37.5%	29.7%
Probability MSE, same caps	344	0.0273	37.8%	32.0%	30.2%
Probability MSE, raw beliefs	344	0.0193	38.1%	31.1%	30.8%

Notes: Predictions use raw priors in both the recommendation process and the direct prior term and are scored against original held-out posteriors. The same-caps fits differ only in objective. The raw-training fit instead uses original training beliefs. CR and wIU fits remain fixed. Each estimator fits both stages separately for each participant using their training observations. Losses are averaged over each participant's 100 splits. All comparisons use the common sample of 344 participants defined in Appendix B.3.

B.5 Excluding Endpoint Priors

We check robustness to excluding endpoint priors by retaining observations with $0 < \mu(o) < 1$, including every posterior report. This leaves 48,817 observations from 374 participants, including 987 endpoint posteriors. We apply this restriction within the original 100 training and test splits without drawing new assignments. For the refitted comparison, both qB stages and CR and wIU are estimated on the same restricted training observations. Recommendation regressions remain participant-specific for participant comparisons, and qB retains the posterior cap of 0.005 and

0.995 for log-odds estimation. Interior priors require no cap on the reporting grid.

Appendix Table S3 separates changing the scoring sample from changing the estimation sample. All aggregate fits are assessable. Participant comparisons use the 309 of the original 344 participants with assessable results in all 100 restricted splits.[S1] After refitting, CR and wIU retain lower aggregate test MSE than qB. Among the 309 participants, CR has the lowest average test MSE, while qB is the most frequent winner. Five participants tie after refitting. The aggregate ranking therefore persists when endpoint priors are excluded, while participant classifications depend on which priors enter estimation and scoring.

Appendix Table S3: Model comparisons excluding endpoint priors.

	Test MSE			Lowest MSE (%)			
Comparison	qB	CR	wIU	qB	CR	wIU	Tied
Panel A: Aggregate analysis, 100 splits							
Original	0.0402	0.0336	0.0336	0.0	24.0	76.0	0.0
Rescored	0.0334	0.0289	0.0287	0.0	0.0	100.0	0.0
Refitted	0.0317	0.0288	0.0285	0.0	0.0	100.0	0.0
Panel B: Participant comparisons, common sample of 309							
Original	0.0197	0.0154	0.0216	34.3	36.2	29.4	0.0
Rescored	0.0163	0.0154	0.0222	32.7	35.9	31.4	0.0
Refitted	0.0160	0.0153	0.0221	35.6	33.3	29.4	1.6

Notes: Original uses the saved fits and all test observations. Rescored uses those same fits on test observations with interior priors. Refitted restricts both training and scoring to interior priors. Every posterior report is retained. In Panel A, Original uses 60,252 observations from 377 participants, while the other rows use 48,817 from 374. Panel B uses the same 309 participants in every row, with 49,382 observations originally and 44,954 with interior priors, including 849 endpoint posteriors. Thus, its Original row differs from the main comparison of 344 participants. Panel A averages MSE across splits and reports split winner shares. Panel B averages each participant's MSE over 100 splits before averaging across participants and calculating winner shares. Ties within 10^{-10} are reported separately and remain in the denominator. All scores use original posterior reports.

B.6 Methods for the Updating Property Tests

Ranges and restrictions. We compare mean posteriors and signed updating across fixed prior ranges. Let $p = \mu(o)$ and $q = \mu_R(o)$ denote reported prior and posterior

[S1] Of 34,400 participant-by-split comparisons, 1,830 are unavailable. There are 1,443 separated recommendation regressions, 26 rank-deficient regressions, and 361 comparisons with no interior observations in the training or test set. These problems affect 35 participants in total. We do not replace unavailable regressions with pooled or penalized estimates.

probabilities of Over 21. Here $R = 1$ codes Over 21 (R_o) and $R = 0$ codes Under 21 (R_u). Signed updating is $u = (2R - 1)(q - p)$, so positive values indicate movement toward either recommendation. The ranges $B_1 = [0, 0.25)$, $B_2 = [0.25, 0.50)$, $B_3 = [0.50, 0.75)$, and $B_4 = [0.75, 1]$ contain integer slider ticks 0–24, 25–49, 50–74, and 75–100. These are fixed probability ranges shared across recommendations, participants, and splits, not empirical quartiles.

Integer slider ticks determine range membership and equality of reports. Stored probabilities differ from the slider reports expressed in probability units by less than 3×10^{-8}, consistent with floating-point storage. We use the original probabilities without recoding to calculate updates and give each observation equal weight.

We group observations by recommendation and prior range. Within each group, $\bar{p}_{rj}$, $\bar{q}_{rj}$, and $\bar{u}_{rj}$ are the averages of the reported priors, their paired posteriors, and signed updates, respectively. Group membership depends only on the prior and recommendation, not on the posterior. For each recommendation, we compare these averages across successive occupied prior ranges. We skip empty ranges but not singleton cells, which make the corresponding comparison unassessable because each contributing cell needs at least two observations. Consistency, monotonicity, and reactionary updating imply the weak restrictions

$$\begin{aligned} C_{rj} &= \bar{u}_{rj} \geq 0, \\ M_{r,jk} &= \bar{q}_{rk} - \bar{q}_{rj} \geq 0, \\ A_{r,jk} &= (\bar{p}_{rk} - \bar{p}_{rj}) - (\bar{q}_{rk} - \bar{q}_{rj}) \geq 0, \qquad j < k. \end{aligned}$$

The three properties yield at most 8, 6, and 6 component tests, respectively. Each tests against a negative alternative and allows equality. Since $A_{1,jk} = \bar{u}_{1j} - \bar{u}_{1k}$ and $A_{0,jk} = \bar{u}_{0k} - \bar{u}_{0j}$, $A \geq 0$ requires that movement not shrink as contradiction increases. We estimate A from differences in mean $p - q$, retaining variation in both beliefs and their covariance. As a numerical check, we verify $M_{r,jk} + A_{r,jk} = \bar{p}_{rk} - \bar{p}_{rj}$.

Inference. Aggregate means come from intercept-only regressions. For contrasts, we add an indicator for the higher range and no other controls. We account for dependence within participants and images by adding the participant and image cluster covariance estimates and subtracting their intersection (Cameron, Gelbach, and Miller, 2011). This includes covariance between cell means. We use no finite-sample correction and a standard normal reference distribution. For each participant's

cell mean, the standard error is $s/\sqrt{n}$ and the reference distribution is Student-t_{n-1}, where s is the sample standard deviation. Contrasts use Welch inference:

$$\text{se} = \sqrt{v_j + v_k}, \qquad \nu = \frac{(v_j + v_k)^2}{v_j^2/(n_j - 1) + v_k^2/(n_k - 1)}, \qquad v_j = s_j^2/n_j.$$

Participant-level inference is unclustered and assumes independent images within each participant, with at most one observation per image. The two-sided 95% intervals are pointwise rather than simultaneous bands. To account for multiple comparisons, Bonferroni-adjusted p-values are $\min\{1, Kp\}$. At both levels, the fixed family sizes are $K = 8, 6, 6$ for the first three properties and $K = 2$ for the threshold criterion. Missing components do not reduce K. A family rejects at 5%, with no further adjustment across properties, participants, or splits.

Degenerate and missing tests. When a standard error is zero, we treat estimated means or differences within numerical rounding tolerance of zero as zero. For our bounded outcomes, this tolerance is approximately 1.4×10^{-14} in probability units, equal to 64 times double-precision machine epsilon. Deterministic weak tests reject only negative coefficients, while deterministic two-sided zero-mean tests reject only nonzero means. At the aggregate level, a test is treated as deterministic only with full rank, numerically vanishing residuals, and at least two clusters in each dimension.

For the weak tests and the threshold criterion, a valid component rejection counts even when another component is missing. If no component rejects, at least one assessable component gives *non-rejection on observed support*. With none, the criterion is unassessable, subject to the never-updater convention below. We neither impute missing cells nor count them as passing. Because support and precision differ across participants, sparse data can favor non-rejection. The all-four and exactly-three classifications require all criteria to be assessable and use 377 participants as the denominator.

Threshold criterion. At both the aggregate and participant levels, we test whether mean signed updating differs from zero in the most confirming range. The two-sided tests are $\mathbb{E}[u \mid R = 1, p \in B_4] = 0$ and $\mathbb{E}[u \mid R = 0, p \in B_1] = 0$, using the inference procedures above and $K = 2$. *Compatibility with zero* means non-rejection on observed support. No other range is substituted when these are unavailable. The test concerns average updating across the specified prior range. Failure to reject zero does not establish a no-updating interval. Conversely, average updating can

differ from zero across the full range even when conditional mean updating is zero for the most confirming priors within it. For participant classifications only, we count observed never-updaters as meeting the threshold criterion by convention even when their ranges are unassessable, without assigning them a threshold.

Coverage. All 20 components for the first three properties and both components of the threshold criterion are assessable in the 23 aggregate samples and five additional exclusion samples. At the participant level, 133 participants have an empty cell and 192 have at least two observations in every cell. The threshold criterion has two assessable components for 311 participants, one for 61, and none for five. Among participants who sometimes update, 44 meet the threshold criterion with only one assessable range. All 32 never-updaters meet it by convention, but only 28 meet all four criteria because four lack assessable monotonicity and reactionary comparisons. Of the 192 participants with full cell coverage, 151 (78.6%) meet all four criteria.

Appendix Table S4: Aggregate means and contrasts in fixed prior ranges.

Recommendation and range	Estimate	95% interval	Adjusted p	Result
Consistency				
Over 21, range 1	0.267	[0.223, 0.311]	1.000	NR
Over 21, range 2	0.113	[0.096, 0.130]	1.000	NR
Over 21, range 3	0.023	[0.017, 0.029]	1.000	NR
Over 21, range 4	0.001	[-0.002, 0.003]	1.000	NR
Under 21, range 1	0.000	[-0.003, 0.002]	1.000	NR
Under 21, range 2	0.021	[0.015, 0.028]	1.000	NR
Under 21, range 3	0.098	[0.082, 0.114]	1.000	NR
Under 21, range 4	0.191	[0.161, 0.222]	1.000	NR
Monotonicity				
Over 21, range 1 to 2	0.114	[0.070, 0.158]	1.000	NR
Over 21, range 2 to 3	0.189	[0.173, 0.205]	1.000	NR
Over 21, range 3 to 4	0.239	[0.228, 0.249]	1.000	NR
Under 21, range 1 to 2	0.244	[0.232, 0.256]	1.000	NR
Under 21, range 2 to 3	0.197	[0.182, 0.213]	1.000	NR
Under 21, range 3 to 4	0.170	[0.140, 0.201]	1.000	NR
Reactionary updating				
Over 21, range 1 to 2	0.154	[0.112, 0.196]	1.000	NR
Over 21, range 2 to 3	0.090	[0.074, 0.106]	1.000	NR
Over 21, range 3 to 4	0.023	[0.016, 0.029]	1.000	NR
Under 21, range 1 to 2	0.022	[0.016, 0.027]	1.000	NR
Under 21, range 2 to 3	0.077	[0.062, 0.092]	1.000	NR
Under 21, range 3 to 4	0.093	[0.064, 0.122]	1.000	NR

Notes: Ranges and contrasts are defined above. Estimates are in probability units. Intervals use participant and image clustered standard errors and a normal reference. Intervals are pointwise, while p-values use the fixed Bonferroni family sizes. NR denotes non-rejection.

Appendix Table S5: Participant classifications on observed support within prior ranges.

Criterion	Meets	Does not meet	Unassessable	Share meeting
Consistency	369	8	0	97.9%
Monotonicity	354	11	12	93.9%
Reactionary updating	354	11	12	93.9%
Threshold criterion	322	50	5	85.4%
All four criteria	292	68	17	77.5%
Exactly three criteria	59	301	17	15.6%

Notes: All shares use 377 participants. Meeting a criterion means non-rejection on observed support, with the stated classification convention for observed never-updaters. The threshold criterion tests whether average signed updating differs from zero in the most confirming prior ranges. Non-rejection does not establish the theoretical threshold property. A valid rejection is retained even when another component is unavailable. In the joint rows, *Does not meet* counts participants with all criteria assessable who do not belong to that joint category. The 17 participants with an unassessable criterion are excluded from both joint numerators.

Supporting diagnostics. For paired observations in a cell of size n, with $e_i = y_i - \bar{y}$, the participant and image clustered variance of the estimated mean is

$$\widehat{\mathrm{Var}}(\bar{y}) = \frac{1}{n^2}\left[\sum_a \left(\sum_{i:\,\mathrm{participant}(i)=a} e_i\right)^2 + \sum_b \left(\sum_{i:\,\mathrm{image}(i)=b} e_i\right)^2 - \sum_{a,b}\left(\sum_{i:\,(a,b)} e_i\right)^2\right].$$

We use a standard normal reference and no finite-sample correction. Unlike the property tests' conventions for zero variance, these diagnostics require at least two observations and two clusters per dimension, valid identifiers, full-rank fits, finite nonnegative variances, and nondegenerate standard errors. When these conditions fail, including for all-zero updates, we report descriptive means and support but mark inference unassessable.

The diagnostics describe absolute changes and unchanged reports rather than impose additional threshold requirements. Absolute changes cannot cancel as signed changes can, while unchanged reports need not imply unchanged latent beliefs. We apply the covariance formula to $y = |q - p|$ and $y = \mathbf{1}\{q = p\}$ to report one-sided 95% bounds $\bar{y} + z_{0.95}\,\mathrm{se}(\bar{y})$ for absolute change and $\bar{y} - z_{0.95}\,\mathrm{se}(\bar{y})$ for the unchanged fraction, where $\mathrm{se}(\bar{y}) = \sqrt{\widehat{\mathrm{Var}}(\bar{y})}$. In the two confirming ranges of the pooled sample, mean absolute changes are 1.15 and 1.30 percentage points, with upper bounds of 1.41 and 1.54 points, respectively. About 91% of reports are unchanged (Appendix Table S6).

Appendix Table S6: Absolute changes and unchanged reports in confirming ranges of the pooled sample.

Advice	Outcome	Mean	SE	95% bound	Direction
Over	Absolute change	0.0115	0.0016	0.0141	Upper
Over	Unchanged fraction	0.9156	0.0113	0.8969	Lower
Under	Absolute change	0.0130	0.0015	0.0154	Upper
Under	Unchanged fraction	0.9141	0.0110	0.8961	Lower

Notes: The confirming ranges are ticks 75–100 after Over 21 and 0–24 after Under 21, using the same observations as the threshold criterion for the pooled sample. Absolute change is in probability units and unchanged reports are a fraction. Bounds are one-sided, using participant and image clustered standard errors and a normal reference.

B.7 Robustness of Aggregate Property Tests

Appendix Table S8 assesses sensitivity to sample composition by excluding participants based on response times, reported beliefs, and updating behavior.

B.8 Updating by Subgroup

Figure S1 shows the participant and task splits defined in Section 3.3.

B.9 Binary Accuracy

We measure binary accuracy by classifying beliefs above 50% as Over 21 and beliefs below 50% as Under 21. A classification is correct when it matches the person's age category. Reports of exactly 50% receive an accuracy score of 0.5. AI recommendations are scored by whether they match the true age category. Binary accuracy is 61.8% for participants' priors, 66.2% for their posteriors, and 76.3% for AI recommendations.

The average binary accuracy of priors and posteriors is not significantly different across treatments ($p = 0.321$ and $p = 0.822$, respectively). We find no significant differences in binary accuracy by demographic characteristics, even when accounting for the gender and race of the individual in the image. Joint tests of participant gender, age, and race yield $p = 0.686$ for prior accuracy and $p = 0.474$ for posterior accuracy. AI recommendations have higher binary accuracy than 98.4% of participants' prior beliefs and 91.0% of participants' posterior beliefs, and lower binary accuracy than 1.6% and 6.6%, respectively.

Appendix Table S7: Support in the confirming prior ranges by sample.

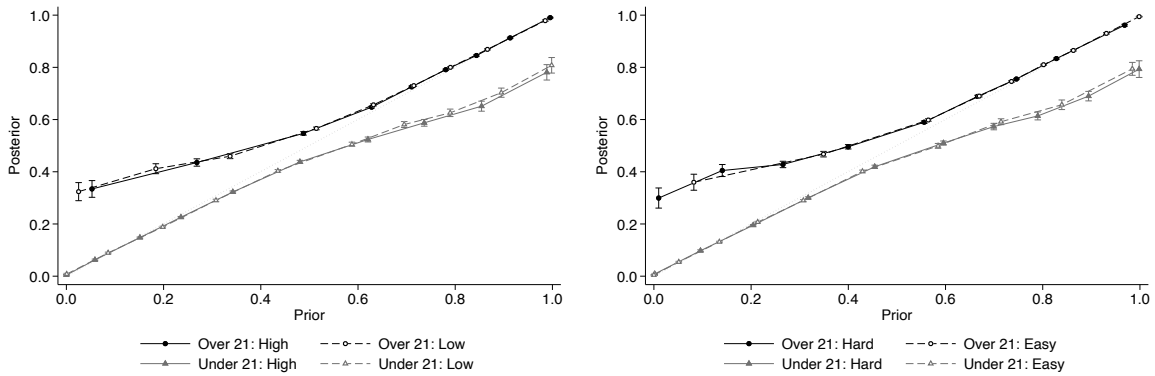
Sample	N: O/U	Participants: O/U	Images: O/U	Endpoints (%): O/U
Samples and splits				
All	15,127/8,812	365/342	90/70	32.1/31.0
NOINFO	7,376/4,524	180/171	90/70	36.7/32.1
INFO	7,751/4,288	185/171	90/70	27.8/29.8
AI correct	11,828/7,720	364/341	66/56	32.5/31.0
Not AI correct	3,299/1,092	344/271	24/14	30.8/30.8
AI high confidence	7,807/5,207	363/334	42/38	32.2/30.8
Not AI high confidence	7,320/3,605	358/324	48/32	32.0/31.3
Hard image	5,333/4,268	352/323	41/39	30.6/31.7
Not Hard image	9,794/4,544	364/335	49/31	33.0/30.3
Second half	7,489/4,545	354/318	90/70	31.7/30.3
Not Second half	7,638/4,267	362/329	90/70	32.5/31.7
Fast prior report	7,699/4,597	180/175	90/70	32.4/32.3
Not Fast prior report	7,428/4,215	185/167	90/70	31.9/29.5
Woman participant	9,217/5,356	218/205	90/70	33.3/32.4
Not Woman participant	5,910/3,456	147/137	90/70	30.4/28.9
Older participant	7,466/4,059	180/168	90/70	28.6/29.2
Not Older participant	7,661/4,753	185/174	90/70	35.5/32.5
High ability	6,324/3,959	186/174	90/70	17.0/9.7
Not High ability	8,803/4,853	179/168	90/70	43.0/48.4
High belief about AI	7,594/4,509	181/165	90/70	34.9/30.3
Not High belief about AI	7,533/4,303	184/177	90/70	29.3/31.7
High prior confidence	8,170/4,721	181/174	90/70	40.8/39.0
Not high prior confidence	6,957/4,091	184/168	90/70	22.0/21.8
Exclusions				
Exclude never-updaters	13,835/7,955	336/314	90/70	30.0/27.8
Exclude rate $\geq$95%	10,309/5,939	251/234	90/70	27.5/25.9
Exclude median RT $<$ 2s	14,543/8,610	354/330	90/70	31.5/31.4
Exclude INFO discrepancy	13,832/8,150	336/313	90/70	32.4/31.4
Exclude concentrated priors	14,767/8,601	360/337	90/70	30.7/29.3

Notes: Each entry reports Over/Under 21 support in the fixed confirming ranges, ticks 75–100 and 0–24. Every sample contains all 25 interior ticks 75–99 after Over and all 24 interior ticks 1–24 after Under. Counts refer to each cell, not the entire sample. Endpoints are prior reports of 100 and 0, respectively. The first row reports the full analysis sample before the additional exclusions. Outcomes and inference appear in Section 3.2 and Tables 1 and S8.

Appendix Table S8: Updating property tests using fixed ranges under exclusion rules.

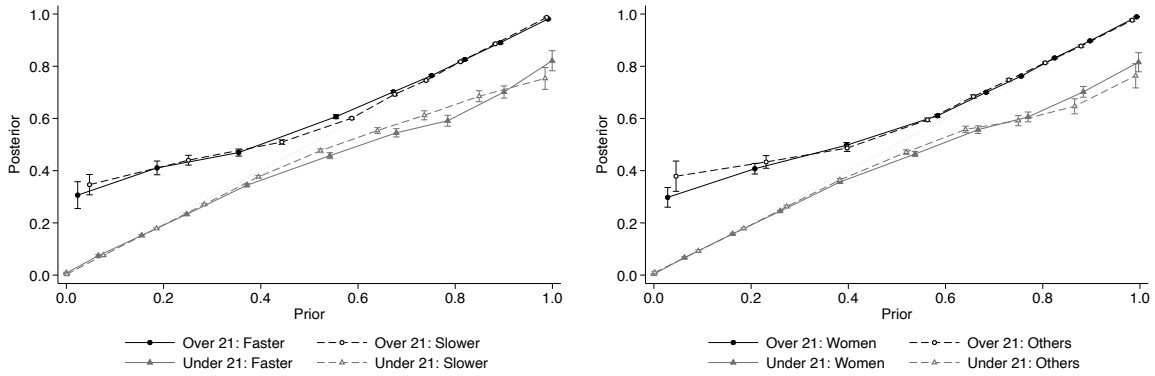
Sample	N	Cons.	Mono.	React.	Thresh.
Main sample	60,252	NR	NR	NR	NR
Exclude never-updaters	55,134	NR	NR	NR	NR
Exclude no-update rate $\geq 95\%$	40,427	NR	NR	NR	NR
Exclude median prior RT below 2 seconds	58,174	NR	NR	NR	NR
Exclude large INFO belief discrepancy	55,455	NR	NR	NR	NR
Exclude concentrated priors	58,973	NR	NR	NR	NR

Notes: NR denotes non-rejection on observed support, as in Table 1. All planned components are assessable in every row. The final column uses two-sided tests of zero mean signed updating in the two most confirming ranges, with a Bonferroni adjustment for two comparisons. Non-rejection under the threshold criterion does not establish an exact plateau. Appendix B.6 provides the methods. Never-updaters are participants whose posterior belief equals their prior belief in every analyzed task. The exclusion based on extreme rates of no updating removes participants whose rate of no updating is at least 95%. The response-time exclusion uses the fifth percentile (2 seconds) of participants' median prior response times and excludes only medians strictly below it. It removes 13 of 377 participants (3.45%) and 2,078 observations. All 67 participants tied at 2 seconds remain. The accuracy-belief discrepancy exclusion removes INFO participants whose stated overall AI accuracy differs by more than 25 percentage points from the empirical AI accuracy calculated over all analyzed participant-image observations. The exclusion based on concentrated priors removes participants whose prior standard deviation is below 0.05 or whose priors fall in a single bin spanning ten percentage points in more than 90% of analyzed tasks. The ten-percentage-point bins contain integer reports of 0–9%, 10–19%, …, 90–100%.



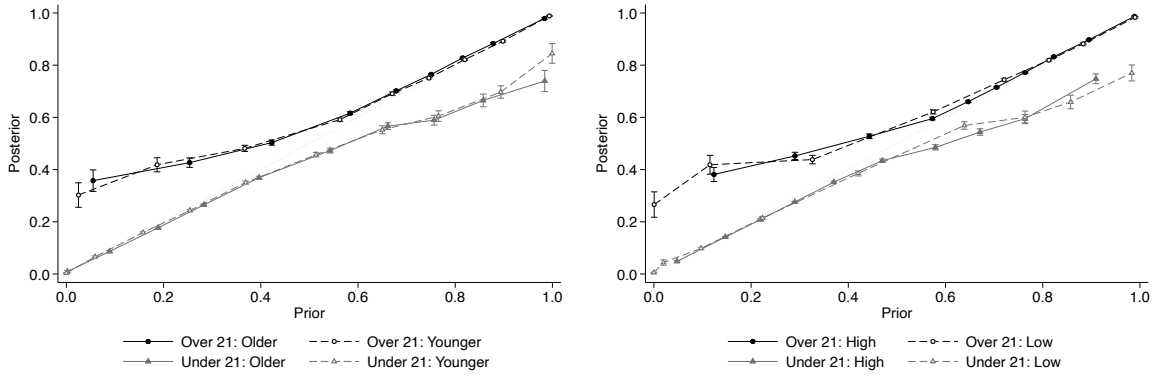
(a) AI confidence

(b) Image difficulty



(c) Prior response time

(d) Gender

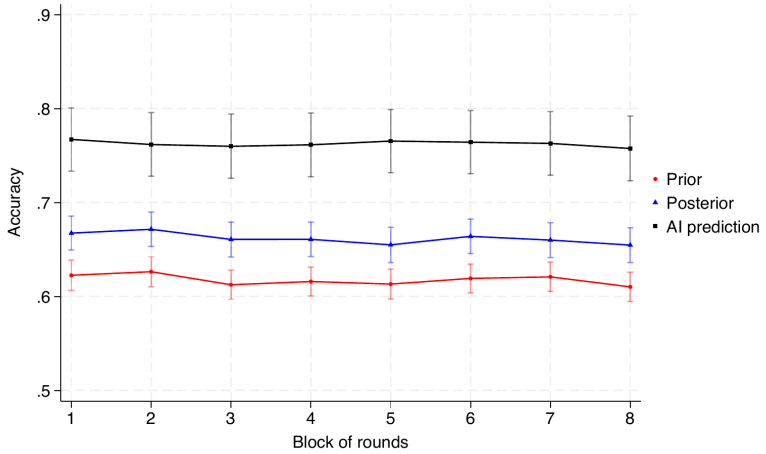


(e) Age

(f) Participant ability

Appendix Figure S1: Updating by subgroup. Points show mean posterior against mean prior beliefs within prior deciles by recommendation and subgroup. Black circles represent Over 21 and gray triangles Under 21. Filled markers and solid lines identify the first legend group, and open markers and dashed lines the second. The dotted diagonal represents no change. Bars show pointwise ±1 standard error clustered by participant and image, conditional on the observed bins.

Figure S2 shows accuracy by round block. Regressions on round block detect no linear trend at the 5% level for prior accuracy ($p = 0.192$), posterior accuracy ($p = 0.057$), or AI accuracy ($p = 0.489$). Together with the randomized image order and similar updating curves across round halves in Figure 1(b), this motivates pooling observations across rounds.



Appendix Figure S2: Dynamics of binary accuracy in the pooled sample. Bars show ±1 standard error clustered by participant and image.

B.10 Additional Summary Tables

Table S9 summarizes participant characteristics by treatment (Panel A), images and AI performance (Panel B), and beliefs about AI accuracy (Panel C).

B.11 Predictive Winner Shares by Treatment

Appendix Table S10 compares classifications by lowest mean validation MSE among qB, CR, and wIU. The common sample of 344 participants defined in Appendix B.3 contains 172 participants in each treatment. The qB shares are 35.5% in INFO and 30.2% in NOINFO. The NOINFO minus INFO difference is −5.2 percentage points ($p = 0.348$).

Appendix Table S9: Summary statistics, treatment balance, and survey beliefs.

Panel A: Participant summary and treatment balance

	All	INFO	NOINFO	p-value
N	377	188	189	
Age, mean (sd)	37.7 (13.3)	39.0 (13.9)	36.5 (12.5)	0.072
Age, median	35.0	37.0	34.0	
Woman	59.4%	61.2%	57.7%	0.490
White	61.5%	63.3%	59.8%	0.485
Prior binary accuracy	61.8%	62.1%	61.4%	0.321
Posterior binary accuracy	66.2%	66.3%	66.1%	0.822
Belief about AI accuracy, mean (sd)	67.6 (20.6)	68.2 (19.0)	67.0 (22.1)	0.544
Confidence in own prior, mean (sd)	73.9 (20.8)	72.5 (21.7)	75.4 (19.9)	0.178
Confidence in posterior, mean (sd)	74.3 (20.9)	73.0 (22.1)	75.6 (19.6)	0.239
Prior RT, mean of medians (s)	3.8 (1.6)	3.8 (1.7)	3.7 (1.6)	0.512
Posterior RT, mean of medians (s)	6.9 (2.5)	7.0 (2.6)	6.8 (2.5)	0.507
No-update rate	78.6%	78.6%	78.6%	0.984
Above-median Prolific approvals	49.9%	52.7%	47.1%	0.281

Panel B: Image and task-instance summary

Measure	Value
Images	160
Over 21 images	80 (50.0%)
Under 21 images	80 (50.0%)
Age labels (years), Under / Over 21	16–19 / 23–26
AI recommends Over 21	56.2%
AI binary accuracy	76.2%
AI model confidence, mean (sd)	0.76 (0.13)
White image	78.8%
Woman image	50.0%

Panel C: Survey beliefs about overall AI accuracy

Measure	Value
Empirical overall AI accuracy	76.3 %
Mean belief: INFO / NOINFO (sd)	68.2 (19.0) / 67.0 (22.1)
INFO minus NOINFO (Welch p)	1.3 (0.544)
Mean absolute error: INFO / NOINFO	13.8 / 18.4

Notes: Panel A uses participants, with standard deviations in parentheses. RT rows are means of participant medians in seconds. Both times start at the beginning of the round, so posterior RT is not incremental. No updating means equal reports. Prolific approvals proxy experience. Fatigue and previous paid classification experience are unavailable in the cleaned data. Panel B uses images, with AI confidence $\max\{p, 1-p\}$ for model probability p. Panel C uses survey reports on a 0–100 scale, with standard deviations in parentheses. Its empirical accuracy is observation weighted after exclusions, unlike Panel B. Absolute errors compare each participant's belief with that rate. All treatment p-values use two-sided Welch tests.

Appendix Table S10: Model classifications by treatment.

Treatment	N	qB	CR	wIU
INFO	172	35.5%	34.3%	30.2%
NOINFO	172	30.2%	40.7%	29.1%

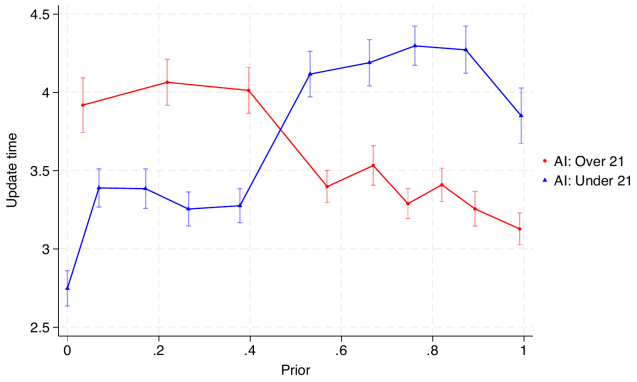
Notes: Each participant's MSE is averaged over 100 common validation splits. All comparisons use the common sample of 344 participants defined in Appendix B.3. Shares use the 172 participants in each treatment as the denominator. The reported p-value tests the difference in qB shares using 5,000 permutations of treatment labels among these 344 participants, holding classifications fixed. This comparison conditions on the retained participants, observed images, and fitted participant-specific recommendation processes.

B.12 Survey Responses

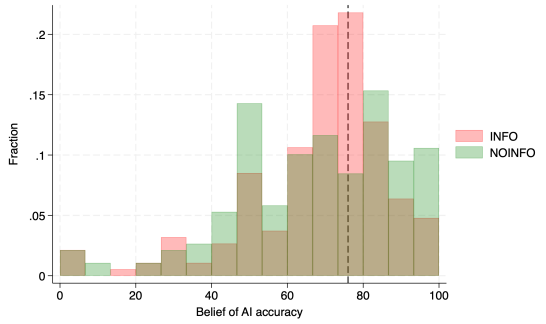
Panel C of Table S9 reports survey beliefs and absolute errors by treatment. Figure S3 shows the distributions of beliefs about AI accuracy, prior confidence, and estimated wIU AI weight. All three panels use the full participant sample. Section 4.6 relates the weights to survey beliefs.

B.13 Response Time

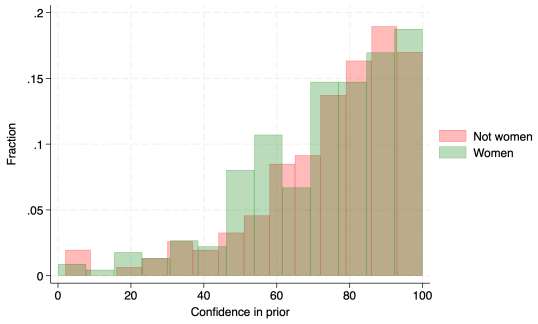
Appendix Figure S4 describes response time after advice, measured as posterior minus prior submission time, without identifying why reports remain unchanged. Among observations with extreme confirming priors and unchanged reports, the median was 2 seconds and the mean was 3.0 seconds.



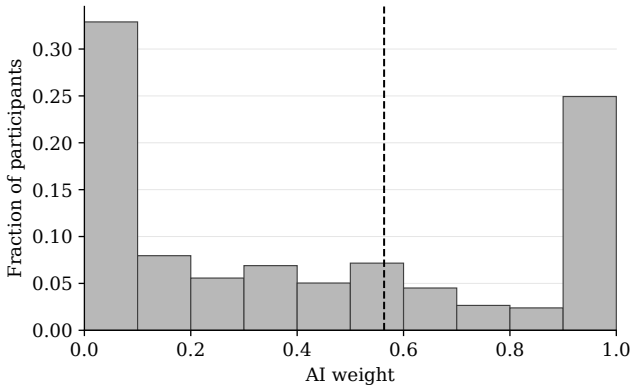
Appendix Figure S4: Update time by prior belief $\mu(o)$. Bars show ± 1 standard error clustered by participant and image.



(a) Beliefs about AI accuracy



(b) Confidence in own prior by gender



(c) Estimated AI weight

Appendix Figure S3: Survey beliefs and estimated wIU AI weight. All panels use 377 participants. Panel (a) separates INFO (188) and NOINFO (189), with a dashed reference at 76%, the rounded AI accuracy. Panel (b) separates 224 women from 153 others (150 men and three who preferred not to say), without a reference line. Panel (c) includes all 52 estimates of zero AI weight. Its dashed line marks the aggregate wIU estimate of 0.563, while the mean estimate across participants is 0.438.

B.14 Weighted Subjective Inertial Updating

Motivation and Definition

wIU, like CR, assumes that information is received as a set of possible probability distributions. To apply these models to our experiment, we must specify how AI recommendations map to sets of distributions. In Section 4, we identify "More likely Over 21" (R_o) with the set $I(R_o) = \{\pi \in \Delta(S) : \pi(o) \in [1/2, 1]\}$, and "More likely Under 21" (R_u) with the symmetric set. These are natural benchmarks, representing the broadest (objective) interpretation of the information conveyed by each recommendation. However, these are not the only possible interpretations of the recommendations. For example, a DM who expects AI recommendations to reflect relatively confident assessments might interpret R_o as implying a probability of at least $\phi_o > 1/2$. Further, there may be heterogeneity in how individuals interpret the same recommendation, and the perceived strengths of the two recommendations need not be symmetric. Studies of verbal probability expressions provide related evidence that people associate different numerical probabilities and bounds with the same wording (Shying, n.d.; Irwin and Mandel, 2023).[S2]

We therefore introduce and estimate a simple extension of wIU that allows for subjective interpretations of recommendations, which we call *weighted subjective Inertial Updating* (wsIU). The extension retains the closest-belief and mixing steps of wIU within the general-information framework of Dominiak, Kovach, and Tserenjigmid (2026), but allows the effective information constraint associated with the recommendations to be subjective:

$$\text{(S4)} \qquad \begin{aligned} \widetilde{I}(R_o) &= \{\pi \in \Delta(S) : \pi(o) \in [\phi_o, 1]\}, \\ \widetilde{I}(R_u) &= \{\pi \in \Delta(S) : \pi(o) \in [0, \phi_u]\}, \end{aligned} \qquad \phi_o \in [1/2, 1], \quad \phi_u \in [0, 1/2].$$

We restrict these sets to be subsets of the corresponding information sets in wIU. Thus, subjective interpretations may strengthen the information conveyed by a recommendation without reversing its direction.

wsIU has three free parameters: w, ϕ_o, ϕ_u, and includes wIU as the case $\phi_o = \phi_u =$

[S2] Shying asks 55 auditors to report the minimum probability they associate with "more likely than not," finding an average threshold of 55.9%. The elicitation of a minimum probability makes this evidence particularly relevant to subjective recommendation boundaries, although the study examines expressions in isolation rather than responses to AI advice.

1/2. The subjective thresholds determine when a prior is sufficiently consistent with the recommendation to remain unchanged, while w determines the weight placed on the prior when updating occurs. For example, under wIU, a prior of 0.60 remains unchanged following an Over 21 recommendation. Under wsIU, the same prior may increase if $\phi_o > 0.60$. A sufficiently high prior, however, remains unchanged. Thus, wsIU allows updating even when a recommendation confirms the direction of the prior, while preserving non-updating for sufficiently strong confirming priors. Formally, the posterior after R_o is

$$\mu_{R_o}(o) = \begin{cases} w\,\mu(o) + (1-w)\,\phi_o & \text{when } \mu(o) < \phi_o \\ \mu(o) & \text{when } \mu(o) \geq \phi_o, \end{cases} \tag{S5}$$

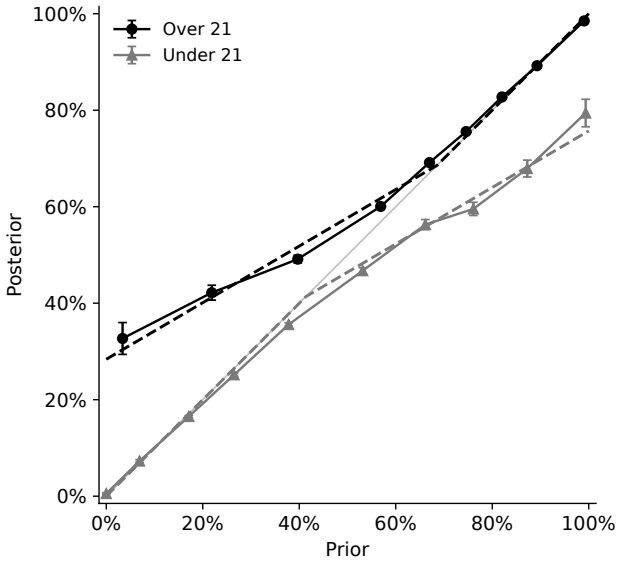
while the posterior after R_u is determined similarly.

Consistency, monotonicity, and reactionary updating hold throughout the stated parameter ranges. If $w < 1$, threshold updating also holds when $\phi_o < 1$ and $\phi_u > 0$, giving no-updating intervals $[\phi_o, 1]$ and $[0, \phi_u]$. At $\phi_o = 1$ or $\phi_u = 0$, only the corresponding confirming endpoint remains unchanged, which does not satisfy the positive-length interval requirement. When $w = 1$, every prior remains unchanged and all four weak properties hold regardless of the thresholds. In that case, the thresholds are unidentified.

Aggregate Fit

We estimate wsIU by minimizing probability MSE over the full closed parameter ranges. For each pair of thresholds, we calculate the AI weight that minimizes training MSE. We first evaluate a grid of threshold pairs, then allow the thresholds to vary continuously in numerical searches starting from several grid points and additional starting values. We retain the parameters with the lowest training MSE. The comparison uses the same 60,252 observations and 100 random 70/30 training and test splits as Section 5.1. Each split estimates parameters using training observations only.

In the pooled sample, the estimates are $\widehat{\phi}_o = 0.685$, $\widehat{\phi}_u = 0.414$, and $\widehat{w} = 0.586$. Figure S5 plots the fitted rule alongside the observed averages. The fitted wsIU rule leaves beliefs unchanged for priors at or above 0.685 after Over 21 advice and at or below 0.414 after Under 21 advice.



Appendix Figure S5: Observed beliefs and the fitted wsIU rule in the pooled sample. Black circles represent Over 21 and gray triangles Under 21. Points and solid lines show observed means within the recommendation-specific prior deciles used in Figure 1. Dashed lines show predictions from the fitted wsIU rule at each prior. The light-gray diagonal represents no updating. Bars show ±1 standard error clustered by participant and image, conditional on the observed bins.

Table S11 adds wsIU to the aggregate comparisons. Its mean test MSE is 0.0330, compared with 0.0336 for CR and wIU and 0.0402 for qB. wsIU has the lowest MSE in all 100 splits. Its in-sample completeness is 95.3%, compared with 91.4% for wIU. The additional thresholds improve fit but also reduce restrictiveness relative to wIU under both synthetic reference classes.

Appendix Table S11: Aggregate comparisons including wsIU.

Model	Test MSE	Completeness	Restrictiveness (%)	
			Monotone	No monotonicity
qB	0.0402	59.4%	11.3 (0.6)	28.8 (0.3)
CR	0.0336	90.9%	32.6 (1.0)	24.6 (0.3)
wIU	0.0336	91.4%	73.3 (1.0)	43.2 (0.3)
wsIU	0.0330	95.3%	35.2 (1.0)	24.7 (0.3)

Notes: Test MSE averages probability MSE over the same 100 splits as Table 3. qB uses capped log-odds estimation for these predictions. CR, wIU, and wsIU minimize training probability MSE. Completeness and restrictiveness use probability MSE for all four models. Restrictiveness uses 100 common synthetic rules per class, with Monte Carlo standard errors in percentage points in parentheses. qB's empirical recommendation process is held fixed for these measures. Appendix B.1 defines the benchmarks and simulation procedure.

Participant Fit

We also fit wsIU separately for each participant using the same 100 training and test splits as the three main models. All comparisons use the common sample of 344 participants defined in Appendix B.3. Table S12 therefore uses the same participants as Tables 4 and 5.

After omitting six fits with zero AI weight and two with insufficient variation to distinguish the thresholds, we summarize the full-sample wsIU estimates for 336 participants. The median Over threshold is 0.755, with an interquartile range of 0.592–0.946. The median Under threshold is 0.334, with an interquartile range of 0.152–0.500. Both thresholds equal 0.5 for 14 participants, so their fitted wsIU rule coincides with wIU.

Mean participant test MSE is 0.0173 for wsIU, compared with 0.0174 for CR, 0.0243 for wIU, and 0.0276 for qB. Paired tests across participants find lower MSE for wsIU than for qB and wIU (both $p < 0.001$). The difference from CR is not statistically significant ($p = 0.217$). Average completeness is 46.1% for wsIU, above wIU's 35.7% but below CR's 47.6%.

Appendix Table S12: Participant comparisons including wsIU.

Model	Mean MSE	Median MSE	Mean completeness	Lowest MSE among four models
qB	0.0276	0.0144	41.9%	26.7%
CR	0.0174	0.0127	47.6%	14.8%
wIU	0.0243	0.0164	35.7%	25.6%
wsIU	0.0173	0.0122	46.1%	32.8%

Notes: Each participant's test MSE is averaged over 100 splits before calculating the displayed means, medians, and classifications. All comparisons use the common sample of 344 participants defined in Appendix B.3. Fits within 10^{-10} of the minimum are tied. Three participants tie for the lowest MSE under wIU and wsIU. Both models assign zero weight to AI recommendations in every split for these participants, so we assign these ties to wIU as the more parsimonious model. Mean completeness instead uses separate probability-MSE fits to each participant's full sample, with participant-specific qB recommendation processes. Paired tests use one mean loss difference per participant and a Student-t reference, conditional on the fitted models and observed images.

Among the four models, wsIU has the lowest mean test MSE for 32.8% of participants, with ties handled as described in the table. In a head-to-head comparison with qB, wsIU has lower MSE for 65.7% of these participants. The extension therefore accommodates additional heterogeneity in where participants begin to leave reports unchanged. Allowing recommendation boundaries to vary improves predictive fit, although our experiment does not directly elicit participants' interpretations of the recommendations.

References

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.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011): "Robust Inference With Multiway Clustering," *Journal of Business & Economic Statistics*, 29, 238–249.

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

Dominiak, A., M. Kovach, and G. Tserenjigmid (2026): "Inertial Updating

with General Information," *arXiv preprint arXiv:2502.00958v2*, October 6, 2026 revision.

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

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.

Irwin, D. and D. R. Mandel (2023): "Communicating Uncertainty in National Security Intelligence: Expert and Nonexpert Interpretations of and Preferences for Verbal and Numeric Formats," *Risk Analysis*, 43, 943–957.

Shying, M. (n.d.): "Auditors Interpretations of "In-Isolation" Verbal Probability Expressions: A Cross-National Study," Undated manuscript, SEC comment archive.