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

Matthew Kovach[1], Daniel Martin[2], Gerelt Tserenjigmid[3]

[1]Purdue University
[2]University of California, Santa Barbara
[3]University of California, Santa Cruz

April 2026

Motivation

- How agents **update their beliefs** in light of new information is a foundational problem in economics and game theory
- **Standard approach:** Data-generating process (DGP) or information structure is *known*
- **This paper:** We experimentally study belief updating when **a DM does not know the DGP**, but learns from qualitative statements (intrinsically meaningful information)
- Learning from Artificial Intelligence (AI) recommendations provides an ideal experimental environment
  - DGP is typically unknown to users (but analysts can estimate it)
  - Can align AI incentives with users
  - Recommendations typically provide qualitative information

Overview

- **AI recommendations** can come in many forms
  - Most-likely state / suggested action / alert
  - Likely to be a police officer / "slow down" / icon on maps app
  - Likely to be an engine issue / "fix truck" / blinking light
- Can result from coarsening precise AI output (Hoong & Dreyfuss (2025))
- Common feature: conveys qualitative information about the underlying state
- **Our question:** How do decision makers revise their beliefs in response to AI recommendations?

Contribution

- We designed a controlled experiment to test three possibilities:
  1. Quasi-Bayesian updating (e.g., Grether (1980), Agarwal et al. (2023))
  2. Contraction Rule (a non-Bayesian rule-of-thumb) of Ke et al. (2024)
  3. Weighted Inertial Updating of Dominiak et al. (2021)
- **We find that the non-Bayesian approaches (Contraction Rule and weighted Inertial Updating) have a much better fit than quasi-Bayesian updating**
- The best fit (aggregate/individual level, in-/out-of-sample) is from a flexible version of weighted Inertial Updating that incorporates elements of the Contraction Rule
- Suggests we need new approaches to modeling belief updates in the case of an unknown DGP and also in a world of AI

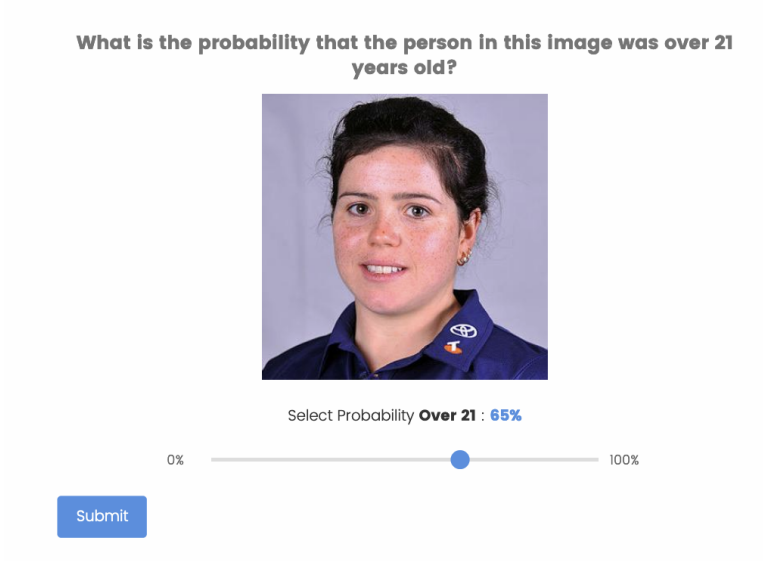
Caveats

- Just one data point!
- Would love to see further testing of these models
  - Other unknown DGPs (other AIs, humans, etc.)
  - Other types of recommendations
  - Other tasks

Our Experimental Task

- To see how people update their beliefs with AI recommendations, we adapt the "Bouncer" task of Caplin et al. (2024)
- In each of 160 rounds, subjects are presented with an image of a human face (on Prolific)
- They are asked to report the probability that the individual was over 21 years old at the time the image was taken
- Report prior $\mu(s)$ where $s_1 =$ Over 21 and $s_2 =$ Under 21

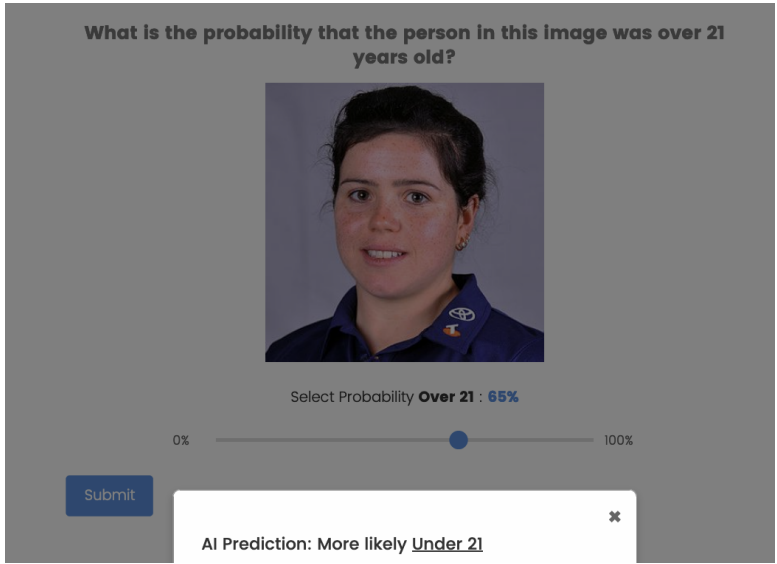
Prior $\mu(s_1)$ Elicitation



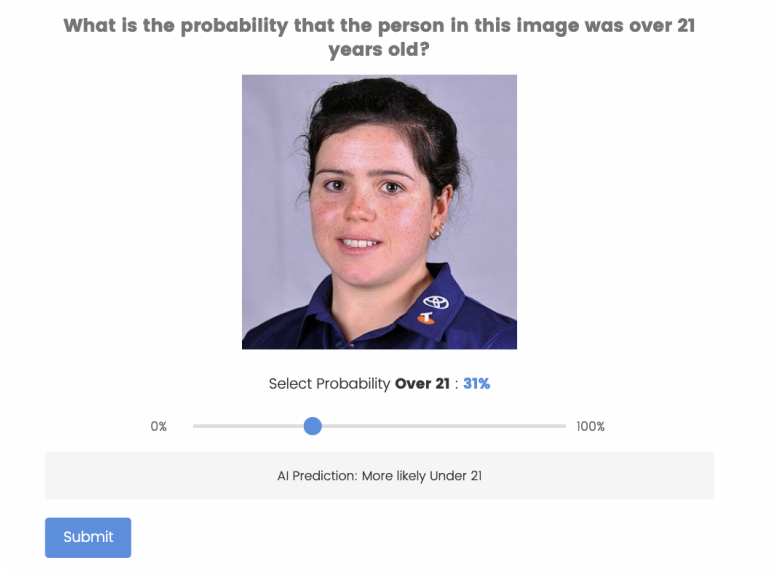
AI Recommendations

- After reporting prior $\mu(s_1)$, subjects are shown an "AI Assistant" recommendation
- This AI recommendation is either the individual is **more likely to be over 21** or **more likely to be under 21**
- After seeing the AI recommendation, subjects can then update their initial report
- Report posterior $\mu_R(s_1)$
- **Our question:** How do reports change after AI recommendations that the person is over or under 21?
- How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?

AI Recommendations



Posterior $\mu_R(s_1)$ Elicitation



How did we generate the AI recommendations?

1. Started with the "Caffe" model from Rothe et al. (2018) that is trained to predict the ages of humans (based just on images of them)
   - Uses a convolutional neural network (CNN) pre-trained on ImageNet (the standard approach to image classification)
   - Trained using a large sample of images scrapped from the internet (e.g., Wikipedia)
2. Output is a confidence score between 0 and 1 for each possible age, where the confidence scores sum to 1 across all possible ages
3. If the sum of confidence scores of ages above 21 is $> .5$, then the AI assistant recommends "more likely to be over 21" otherwise it recommends "more likely to be under 21"

What did subjects know about the AI recommendations?

- **NOINFO** treatment: told nothing about AI
- **INFO** treatment: told who trained it and the following information
  - **Accuracy info**: AI matches age 76% of the time, "better than most humans, but worse than the very best humans"
    - Mimics information in Agarwal et al. (2023) and Caplin et al. (2024)
  - **Probability info**: AI says over 21 when over 21 83% of the time and says under 21 when under 21 70% of the time
    - Needed for the quasi-Bayesian update
- Why vary info?
  1. Information about AI likely to vary in the world
  2. INFO provides a robustness check that favors quasi-Bayesian (QB) updating

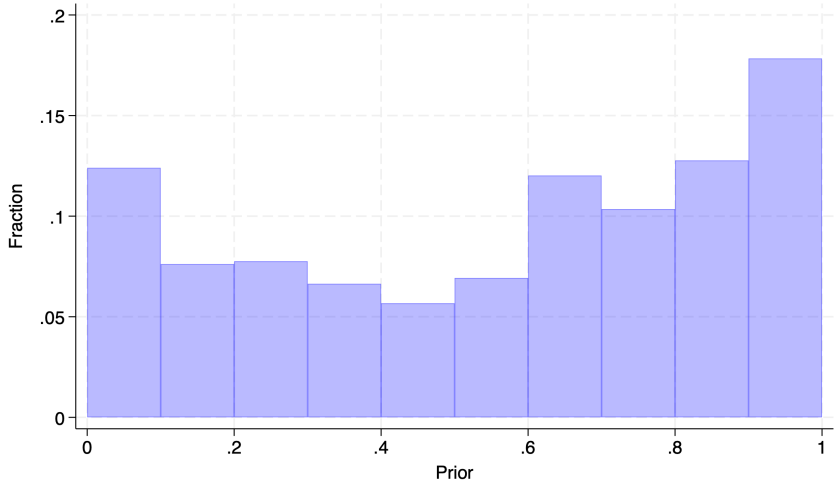
Experiment: Other Details

- 160 total rounds
- No practice rounds (not necessary?)
- No feedback after each choice (minimize learning?)
- Pause after every block of 20 rounds (minimize fatigue?)
- Ran on Prolific (piloted on MTurk)
- Post-experiment questionnaire: confidence in prior and beliefs of AI accuracy
- \$6 participation fee for finishing the experiment
- \$6 bonus payment using binarized scoring rule on randomly selected round and choice (told likelihood of receiving bonus was maximized by truthfully reporting their belief)

Summary Statistics

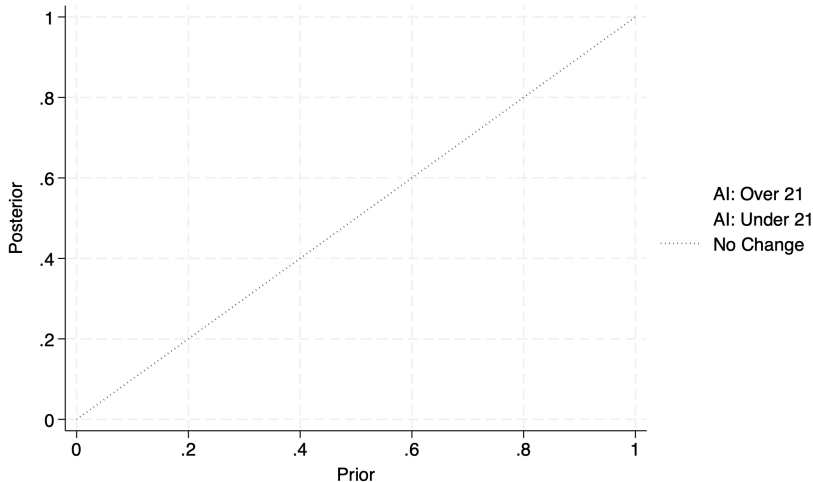
- Observations:
  - 377 subjects across four waves (on Prolific)
    - 189 subjects with NOINFO, 188 with INFO
  - In 68 rounds ($\approx 0.11\%$) time limit hit (no treatment difference)
  - Left with 60,252 total rounds
- Demographics:
  - Required to be ≥ 18 years old, US resident
  - Women: 59%
  - White: 62%
  - Age: median 35, mean 37.7

What is the distribution of the prior $\mu(s_1)$?

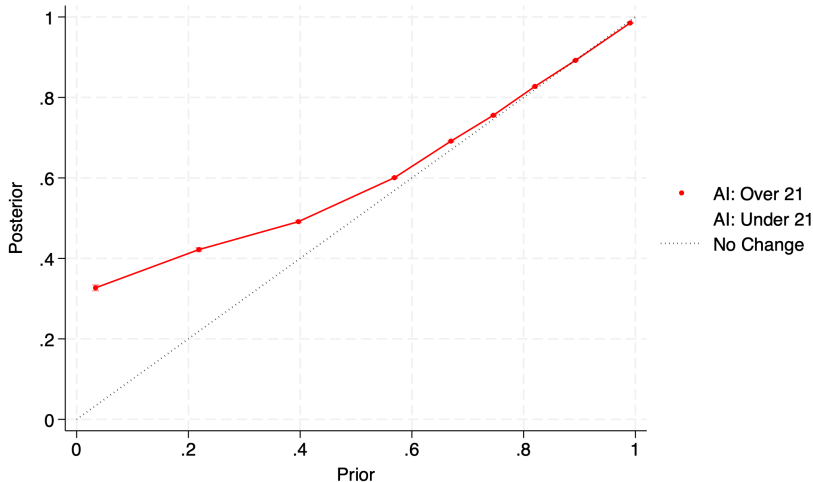


Detailed Subject-Level

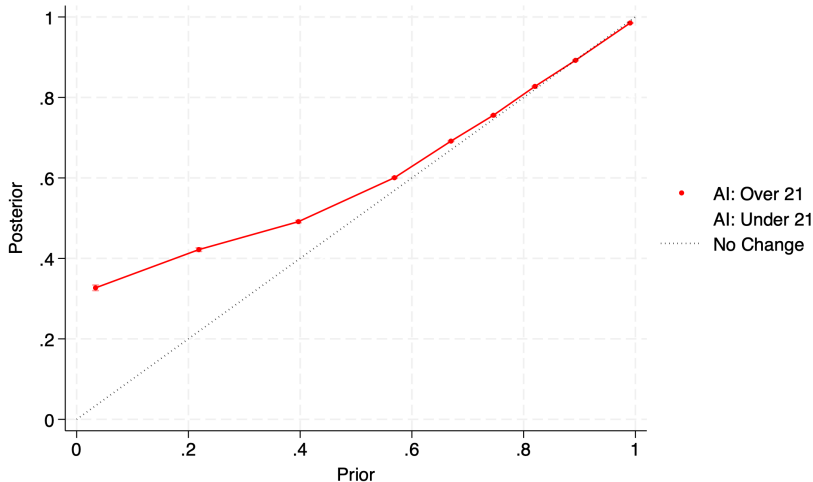
Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?



Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?

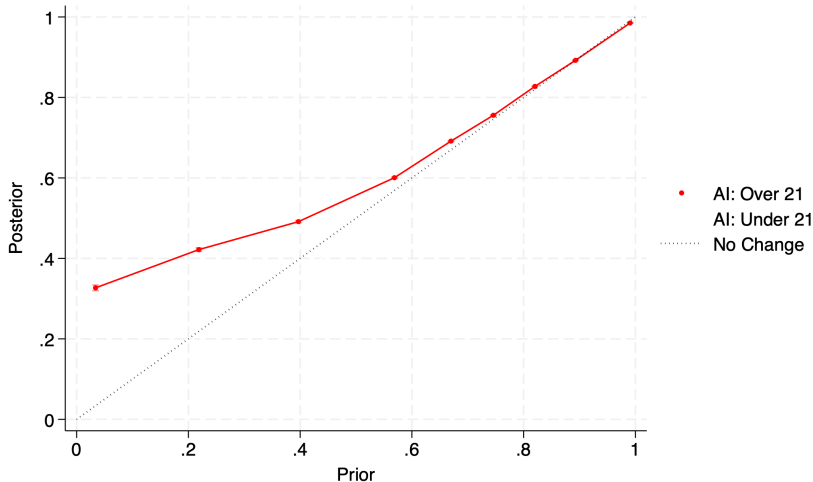


Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?



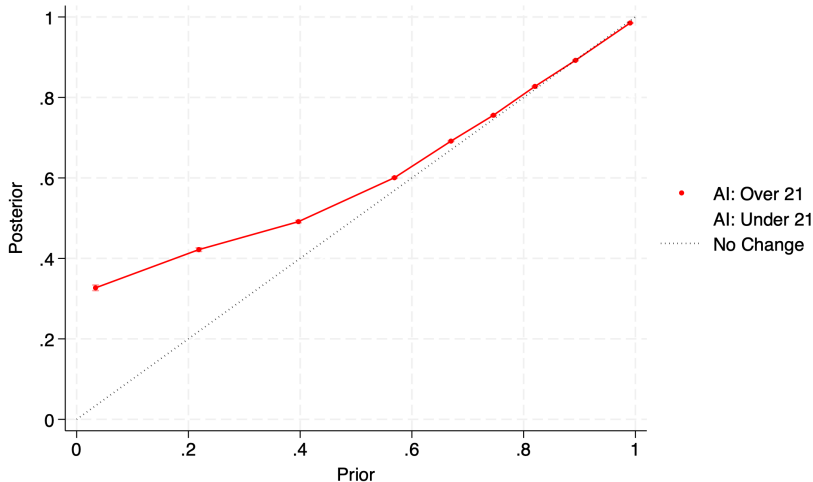
Feature #1: Little updating when prior strong and AI agrees

Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?



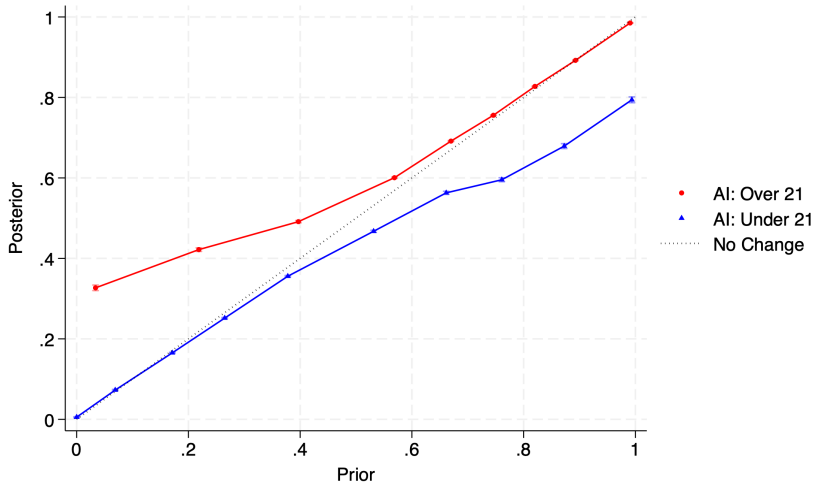
Feature #2: Large updating when prior strong and AI disagrees

Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?



Feature #3: Little updating when prior uncertain

Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?



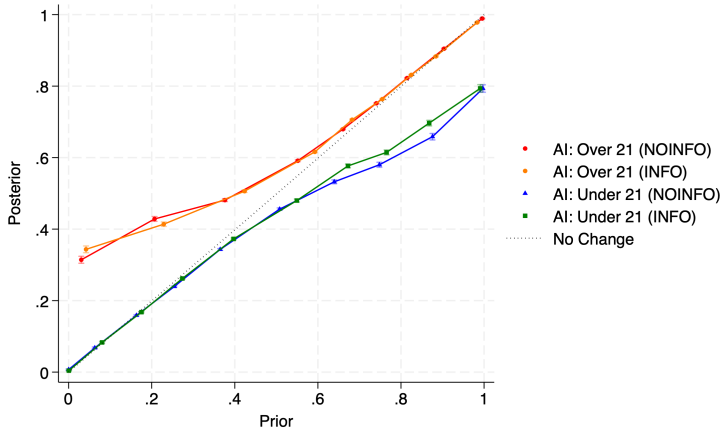
Similar updating pattern in other direction

Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?

Quantitatively similar average updating for many many splits of the data:

1. AI correctness: Advice correct or not
2. AI confidence: Above or below median confidence in advice
3. Image difficulty: MSE on image (across subjects) above or below median
4. Round: Round 1-80 or 81-160
5. Response time: Above or below median response time
6. Gender: Women or not women
7. Age: < 35 or $>= 35$
8. Subject ability: Above or below median accuracy

Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?



Also quantitatively similar average updating across two treatments, NOINFO and INFO

Updating: How are prior $\mu(s_1)$ and posterior $\mu_R(s_1)$ related?

- Possible reason: INFO does correct beliefs, but beliefs about AI are correct on average in NOINFO
- While belief differences cancel out on average, they do matter
- Two things matter for updating:
  1. Belief of AI Accuracy: Above or below median guess of average AI correctness
  2. Decision confidence: Above or below median confidence in prior beliefs

Even with these quantitative differences, same robust features...
what explains this?

Quasi-Bayesian (QB) Updating

- In Quasi-Bayesian (QB) updating rule[1]:

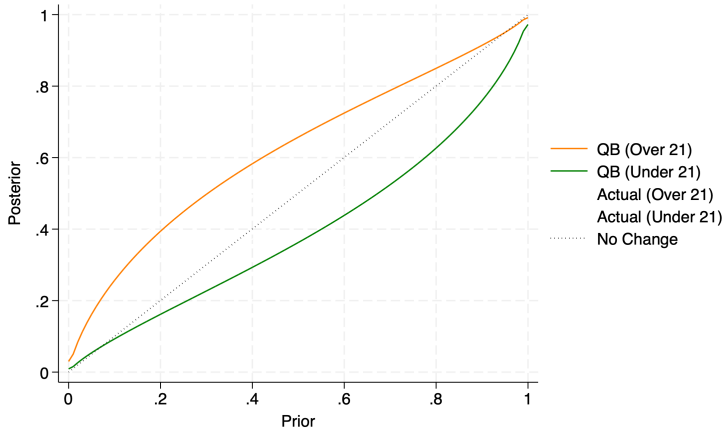
$$\mu_R(s_1) = \frac{Pr(R|s_1)^{\beta_1}\mu(s_1)^{\beta_2}(1+a)}{Pr(R|s_1)^{\beta_1}\mu(s_1)^{\beta_2}(1+a) + Pr(R|s_2)^{\beta_1}\,\mu(s_2)^{\beta_2}}$$

- Allows for many biases and pull-to-state ($\alpha \neq 0$)
- Altogether, three parameters to estimate: $\beta_1, \beta_2, \alpha = \ln(1+a)$

	β_1	β_2	α
Here	0.51	0.78	0.14
Agarwal et al. (2023)	0.26	0.87	Not reported

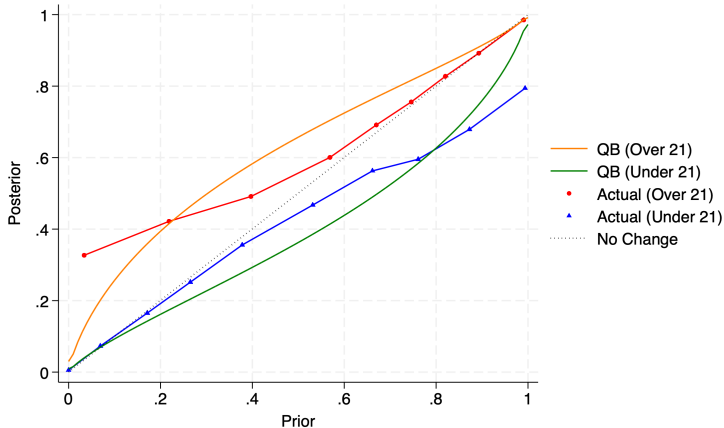
[1] Grether (1980), Kovach (2020), and Agarwal et al. (2023)

QB Predictions



Bayesian Update Other Parameters

QB Predictions + Actual



Contraction Rule (CR) of Ke et al. (2024)

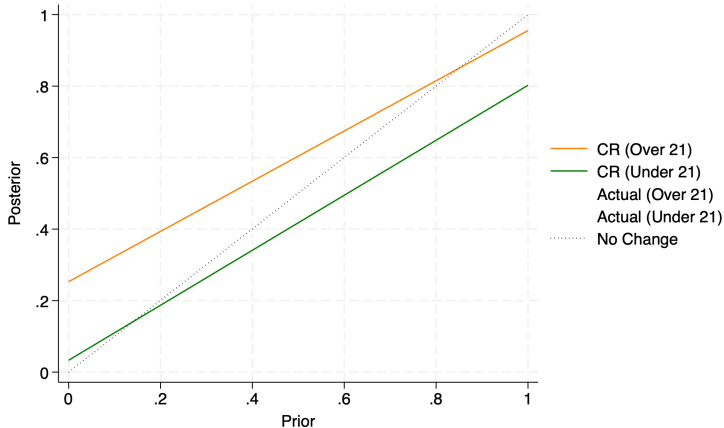
- Decision-maker combines prior μ with belief ρ consistent with recommendation R

$$\mu_R = \epsilon_R \, \mu + (1 - \epsilon_R) \, \rho_R$$

- Parameters to estimate:
  - ϵ_R = weight on prior given recommendation
  - ρ_R = belief given recommendation

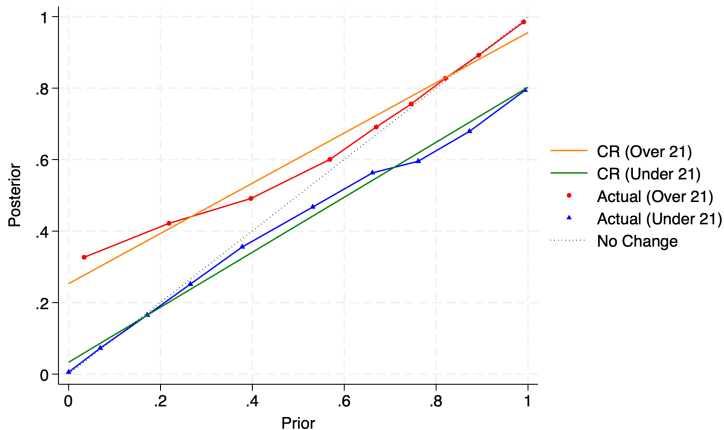
ϵ_{Over}	ϵ_{Under}	$\rho_{Over}(s_1)$	$\rho_{Under}(s_1)$
0.70	0.77	0.85	0.14

CR Predictions



Other Parameters

CR Predictions + Actual



Inertial Updating of Dominiak et al. (2021)

- The Inertial Update (IU) is

$$\pi_R = \arg\min_{\pi \in I_R} d_\mu(\pi),$$

  and I_R is the set of probability distributions consistent with R

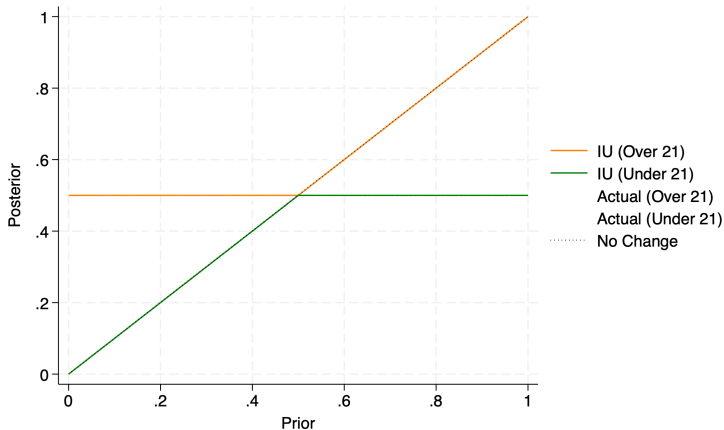
- The prediction for the IU (π_R) is stark if $I_{Over}(s_1) > 0.5$:

$$\pi_{Over}(s_1) = \begin{cases} .5 & \text{when } \mu(s_1) < 0.5 \\ \mu(s_1) & \text{when } \mu(s_1) \geq 0.5 \end{cases}$$

- And if $I_{Under}(s_1) < 0.5$:

$$\pi_{Under}(s_1) = \begin{cases} .5 & \text{when } \mu(s_1) > 0.5 \\ \mu(s_1) & \text{when } \mu(s_1) \leq 0.5 \end{cases}$$

IU Predictions: π_R



Weighted Inertial Updating of Dominiak et al. (2021)

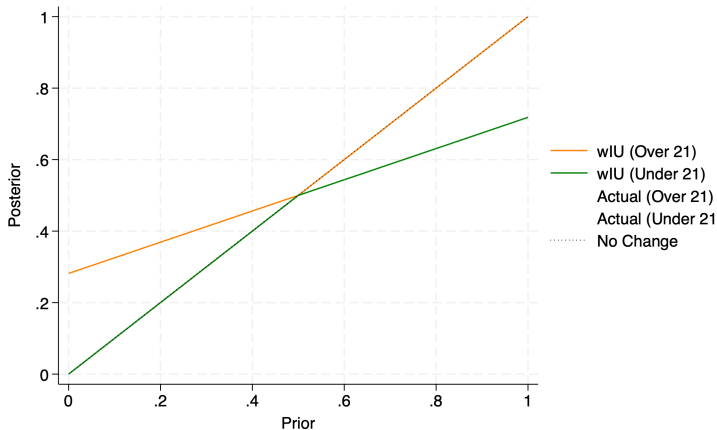
- But with **weighted** Inertial Updating (wIU), mix in prior with the IU:

$$\mu_R = w\,\mu + (1-w)\pi_R$$

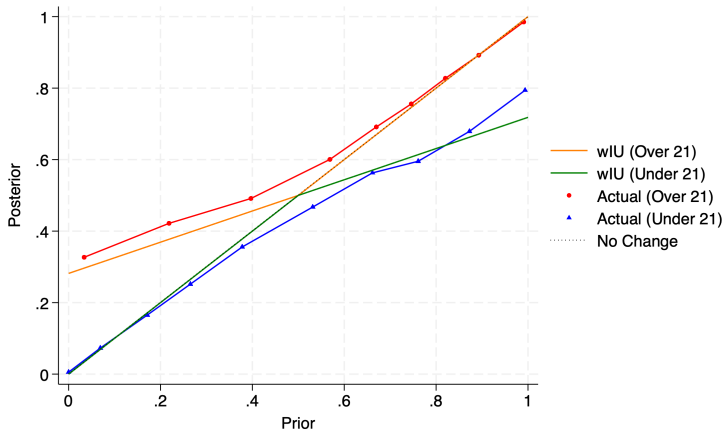
- There is only one parameter to estimate in wIU:

w
0.44

wIU Predictions



wIU Predictions + Actual

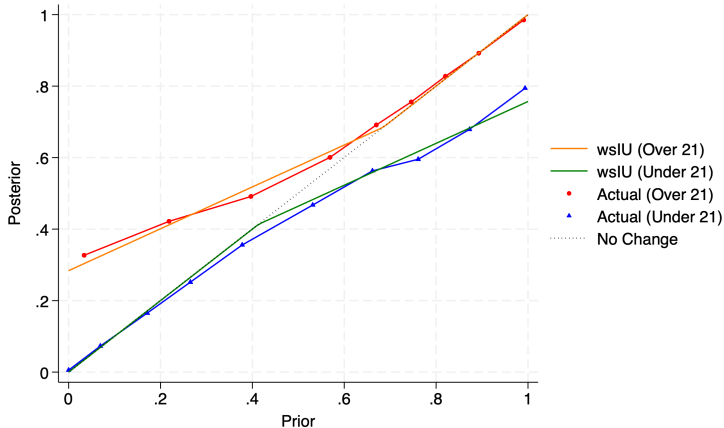


Weighted Subjective Inertial Updating (wsIU)

- But subjects might not interpret "more likely over 21" as $> .5$ and "more likely under 21" as $< .5$
- What if we allow for subjective interpretations?
  - "more likely over 21" means $> \phi_{Over}$ where $\phi_{Over}(s_1) > .5$
  - "more likely under 21" means $< \phi_{Under}$ where $\phi_{Under}(s_1) < .5$
- Keeps form of wIU, but allows kink point to vary (subjectively)
- Technically, ϕ_R operates similar to the ρ_R in CR
- There are three parameters to estimate in wsIU:

$\phi_{Over}(s_1)$	$\phi_{Under}(s_1)$	w
0.68	0.41	0.59

wsIU Predictions + Actual



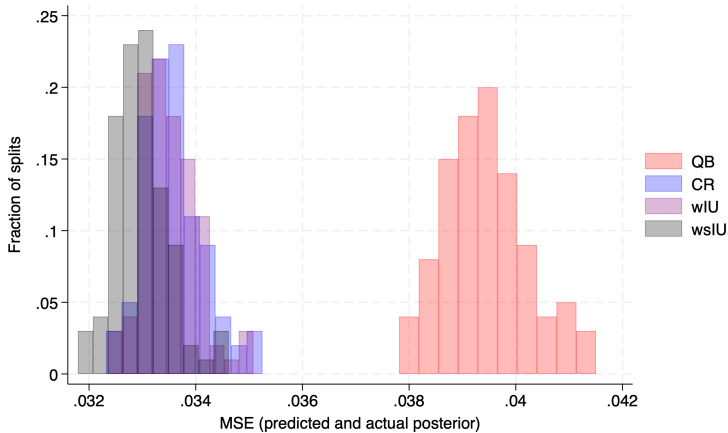
Model Fit: Out-of-Sample Performance

- Our pre-registered way to evaluate fit is with "out-of-sample" performance
  - 100 random splits of dataset into training (70%) and test (30%) observations
  - Estimate model on training and evaluate fit on test
- What is the mean squared error across test splits?

Model	Params	Avg. MSE	Median MSE
QB	3	0.0394	0.0394
CR	4	0.0336	0.0335
wIU	1	0.0335	0.0334
wsIU	3	0.0330	0.0329

CR, wIU, and wsIU have better fit than QB

Model Fit: Out-of-Sample Performance



CR, wIU, and wsIU have better fit than QB for 100% of splits

Model Fit: Out-of-Sample Performance

- What is the fraction of test splits a model has the lowest mean squared error?

Model	Lowest MSE	Lowest MSE
QB	0%	0%
CR	28%	0%
wIU	72%	0%
wsIU		100%

wsIU best fits 100% of splits

Model Fit: Completeness

- Another way to assess model fit is with its "completeness" (Fudenberg et al. (2022))
- Measures how close a model comes to the lowest possible error
  - Normalize using benchmark: no updating (nested in all three models)
- Error measure is squared error $= (predicted - actual\ posterior)^2$

Model	Completeness
QB	47.6%
CR	90.9%
wIU	91.4%
wsIU	95.3%

CR, wIU, and wsIU have a much higher level of completeness than QB

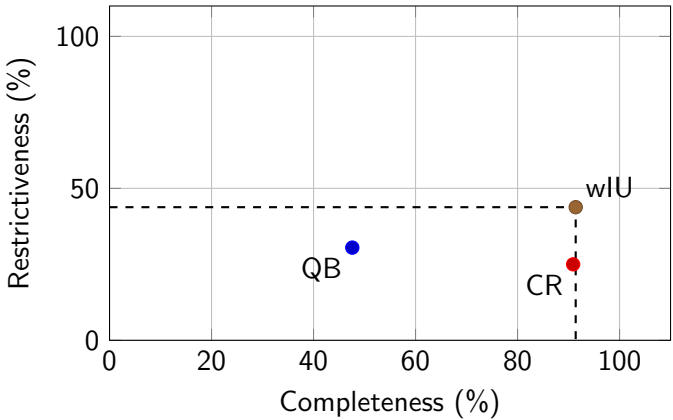
Model Fit: Restrictiveness

- Also want a model that is "restrictive" – isn't so flexible it could explain anything (Fudenberg et al. (2023))
- Measures the (normalized) average error for predicting synthetic data sets
  - Synthetic data: Uniform random updating in the direction of the AI recommendation
- Error measure is squared error $= (predicted - actual\ posterior)^2$

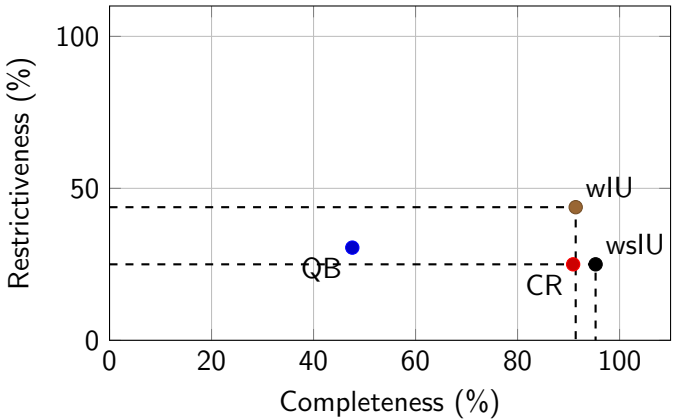
Model	Restrictiveness
QB	30.5%
CR	25.0%
wIU	43.8%
wsIU	25.0%

wIU has a higher level of restrictiveness than CR and wsIU

Completeness vs. Restrictiveness



Completeness vs. Restrictiveness



Model Fit: Out-of-Sample Performance

- So far just looked at aggregate level – now look at individual level
- Smaller number of updates (n=160) likely to produce more extreme behavior
- What is the average mean squared error of individuals across test splits?

Approach	Params	Avg. MSE	Median MSE
QB	3	0.0168	0.0104
CR	4	0.0159	0.0103
wIU	1	0.0222	0.0138
wsIU	3	**0.0158**	**0.0099**

wsIU has the best fit at the individual level also

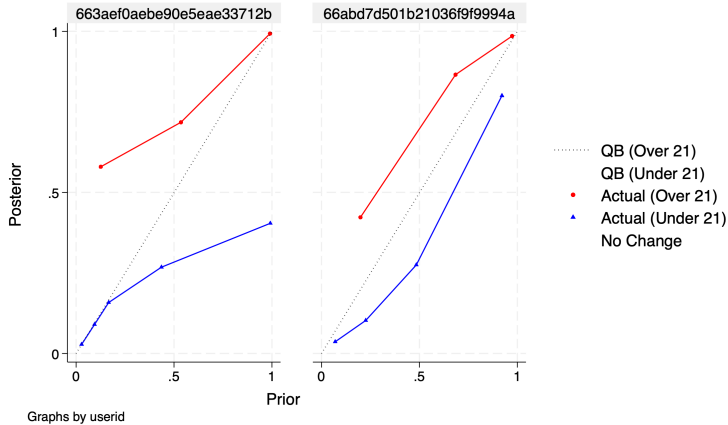
Model Fit: Out-of-Sample Performance

- What is the fraction of individuals with the lowest mean squared error?

Model	Lowest MSE	Lowest MSE	Lowest MSE
QB	32.9%	26.1%	34.7%
CR	32.6%	12.2%	
wIU	34.5%	24.9%	
wsIU		**35.5%**	**65.3%**

wsIU best fits more individuals than QB

QB Predictions + Actual



Contribution

- **Our questions:**
  - How do people revise their beliefs when the DGP is unknown?
  - How do people revise their beliefs in response to AI recommendations?
- **Result:** The non-Bayesian approaches (Contraction Rule and weighted Inertial Updating) have a better fit than the quasi-Bayesian approach
  - The first empirical application of Contraction Rule and weighted Inertial Updating
  - Best fitting model includes elements of both
- Suggests we need new approaches to modeling belief updates in the case of an unknown DGP and also in a world of AI

Thank you!

Heterogeneity | Confidence | Beliefs | Accuracy | Response Times | References

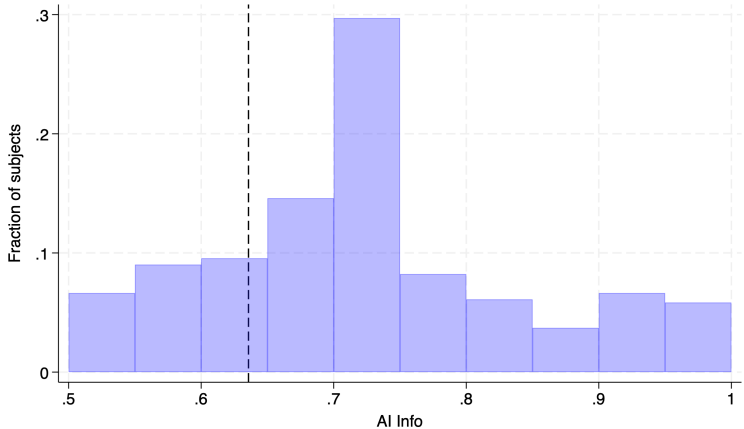
Interpreting Parameters

- How can we interpret the parameters of wsIU?

$$\mu_R = w\,\mu + (1 - w)\phi_R$$

- ϕ is the minimum information in an AI recommendation of "more likely" ("AI Info")
  - Summarize with average parameter strength: $(\phi_{Over}+(1\text{-}\phi_{Under}))/2$
- $1 - w$ is the weight put on this information ("AI Trust")
- At the aggregate level, AI Info is 63.5% and AI Trust is 44%
- What about at the subject level?

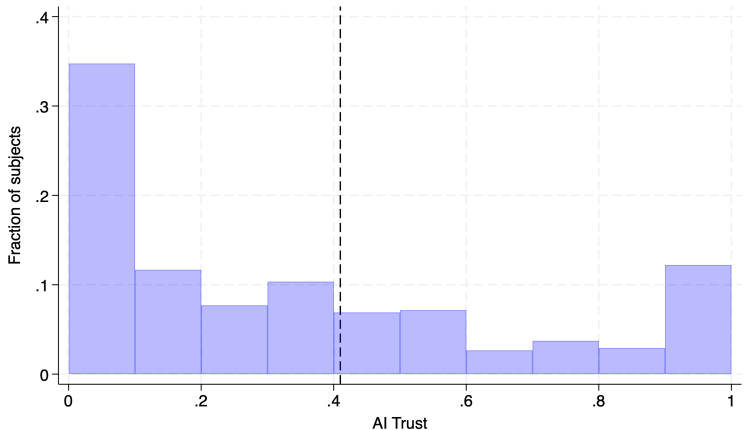
Individual-Level Parameters



What Drives AI Info?

	Coefficient
NOINFO treatment	−0.00809 (0.0124)
Woman	0.00523 (0.0127)
Older (above median)	0.0261 (0.0126)
Confidence in prior (0–100)	−0.000352 (0.000303)
Belief of AI accuracy (0–100%)	0.000327 (0.000303)

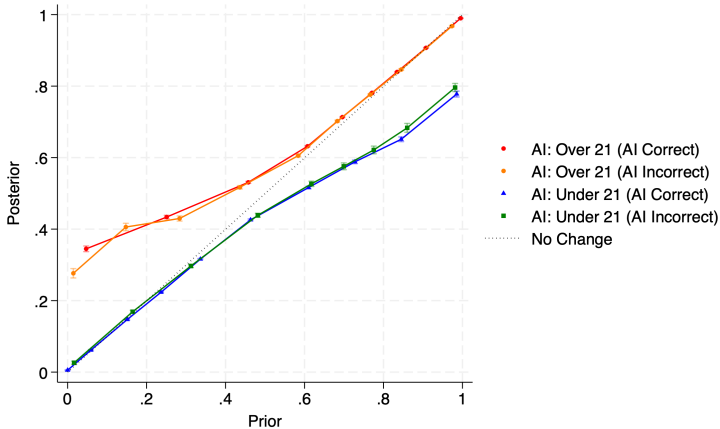
Individual-Level Parameters



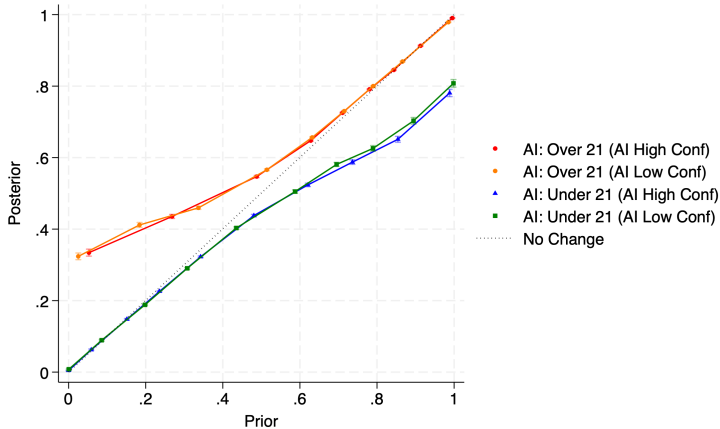
What Drives AI Trust?

	Coefficient
NOINFO treatment	0.0367 (0.0338)
Woman	−0.0468 (0.0345)
Older (above median)	−0.0256 (0.0343)
Confidence in prior (0–100)	−0.00283 (0.00082)
Belief of AI accuracy (0–100%)	0.00342 (0.00083)

AI Correctness (Correct/Not)

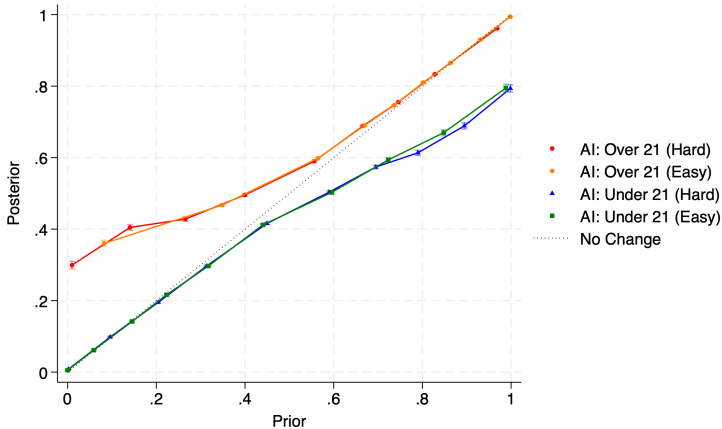


AI Confidence (Higher/Lower)

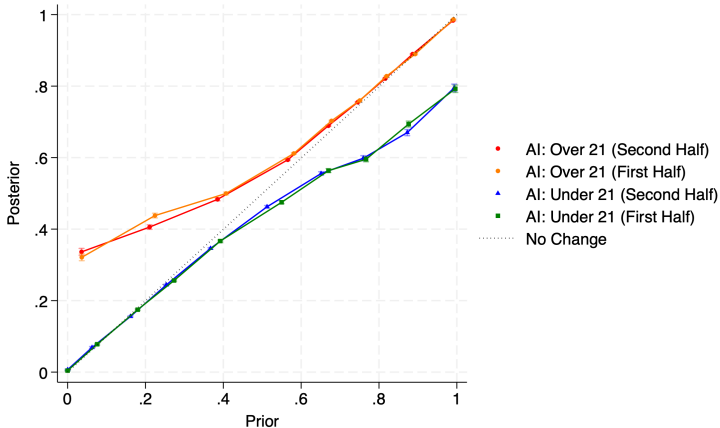


Back End

Image Difficulty (Hard/Easy)

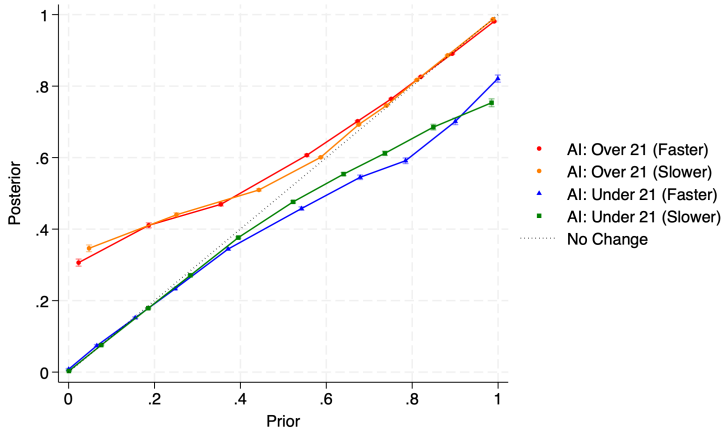


Round (First/Second Half)

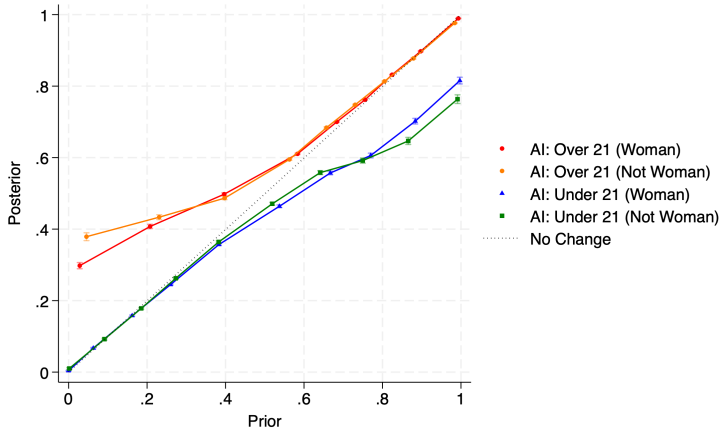


Back End

Response Time (Faster/Slower)

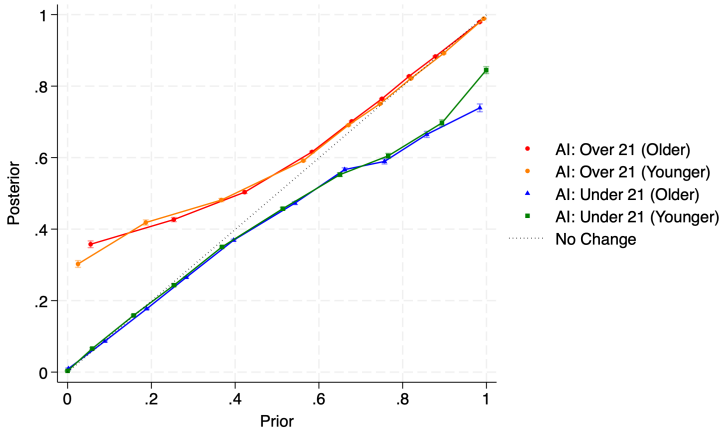


Gender (Woman/Not)



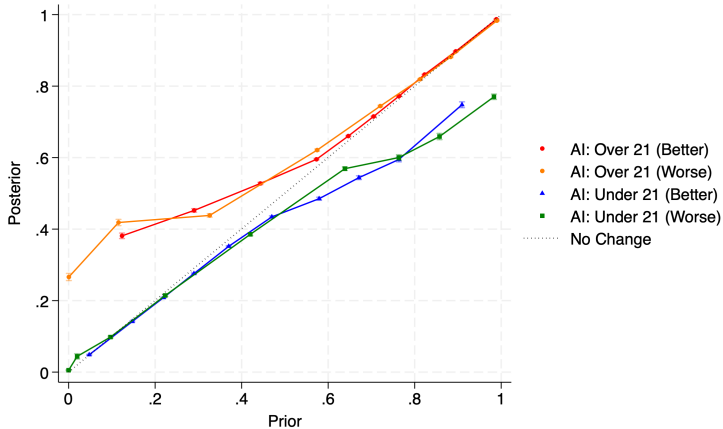
Back End

Age (Younger/Older)



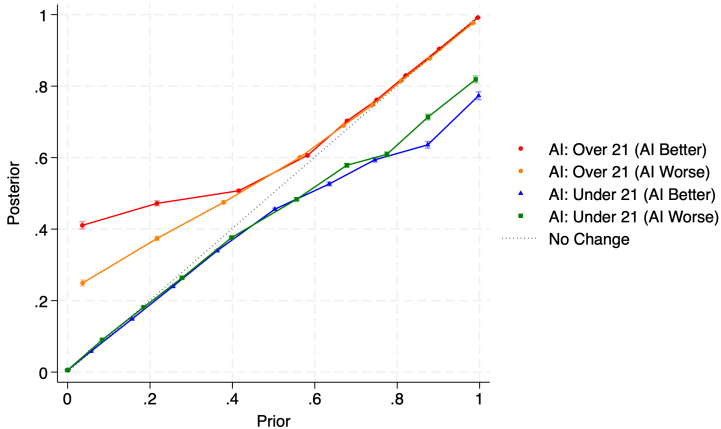
Back End

Subject Ability (Better/Worse)



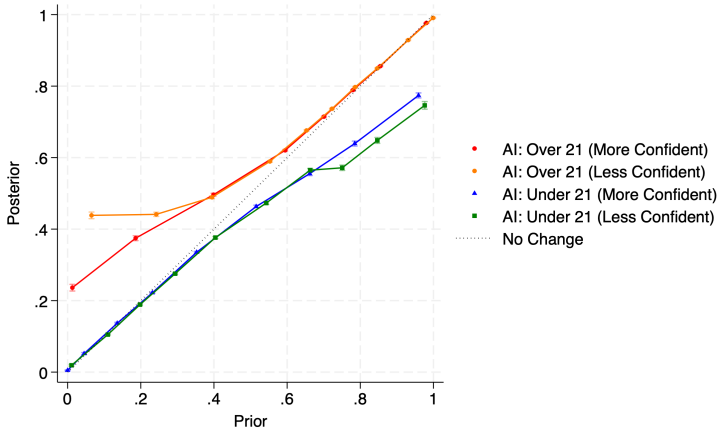
Back End

Belief of AI Accuracy (Higher/Lower)



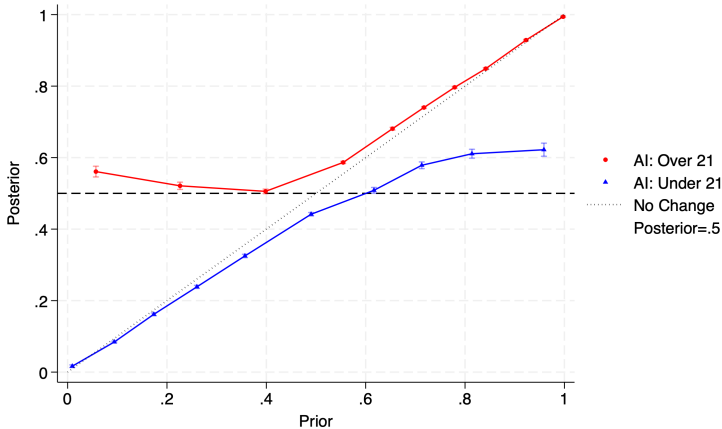
Back End

Decision Confidence (Higher/Lower)



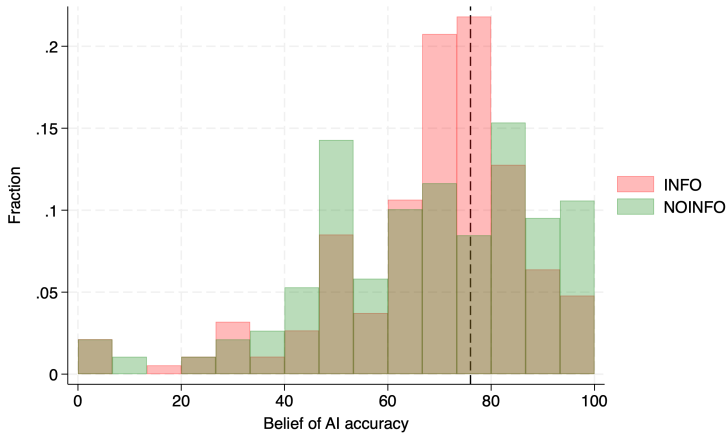
Back End

Higher Belief of AI Accuracy + Lower Decision Confidence



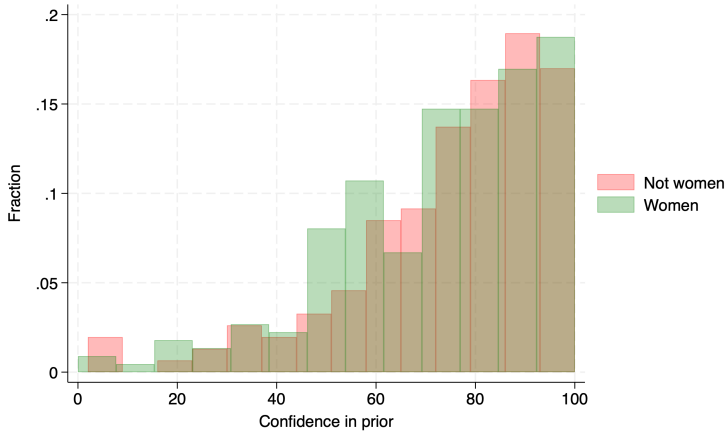
Back End

Belief of AI Accuracy (0-100%)



Back End

Decision Confidence (Prior)



Back End

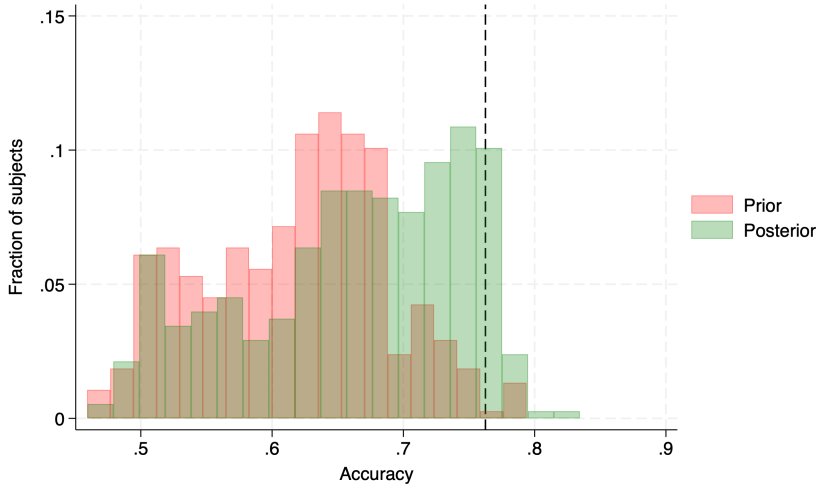
Results: Accuracy

- Measure accuracy differently for humans and AI
  - Humans "correct" if:
    - Report $> 50\%$ and (image) age > 21
    - Report $< 50\%$ and (image) age < 21
  - AI "correct" if recommendation $=$ (image) age
- Overall accuracy:

Images	Subjects	Initial	Final	AI
All	All	62%	66%	76%
All	NOINFO	61%	66%	76%
	INFO	62%	66%	76%

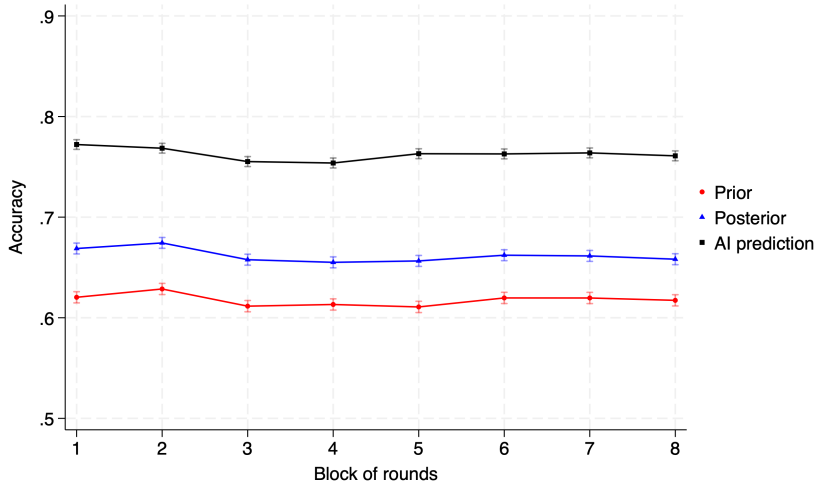
End

Results: Accuracy (Subject Level)



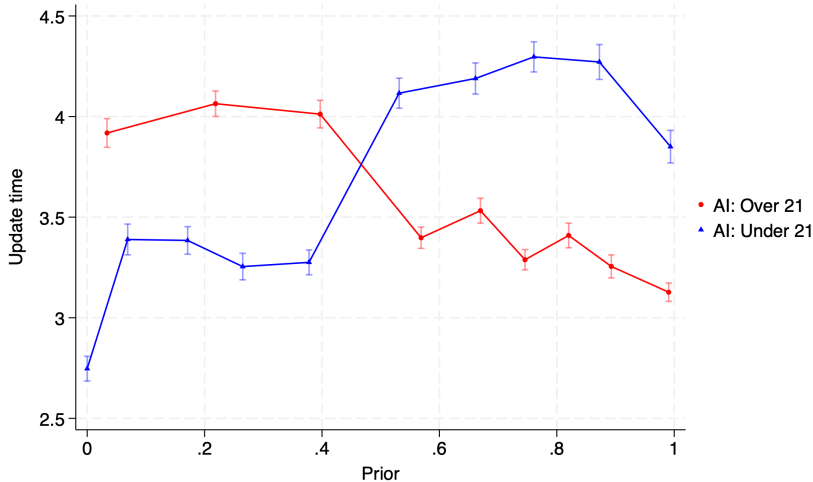
End

Is accuracy consistent over the experiment?



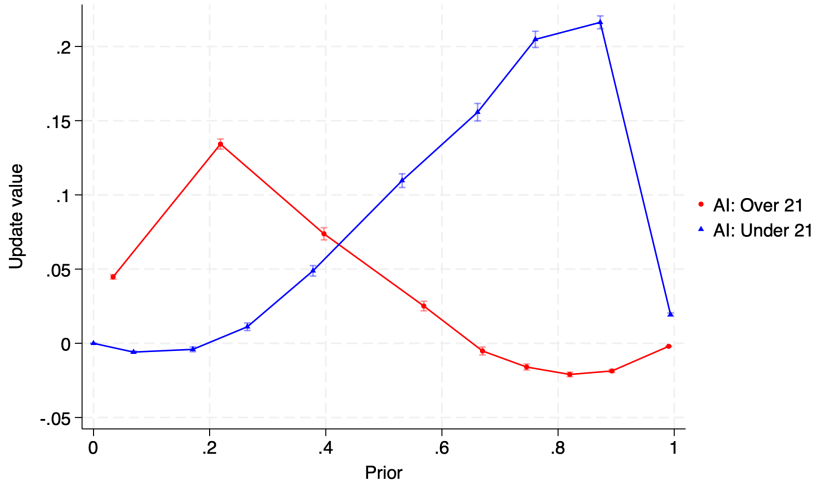
End

What is the update time by prior $\mu(s_1)$?

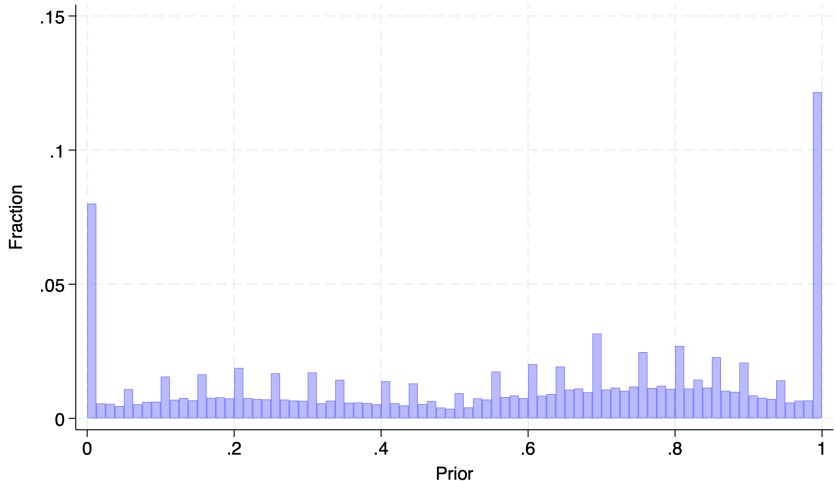


Back

What is the update value of the prior $\mu(s_1)$?

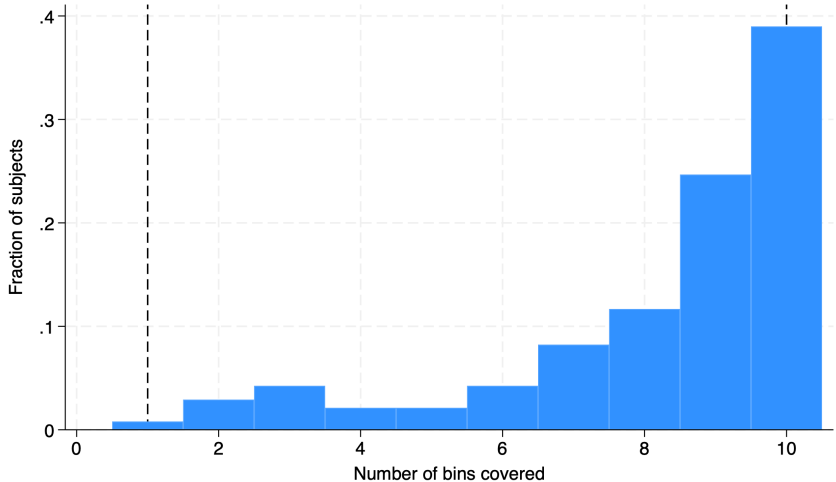


What is the distribution of the prior $\mu(s_1)$?



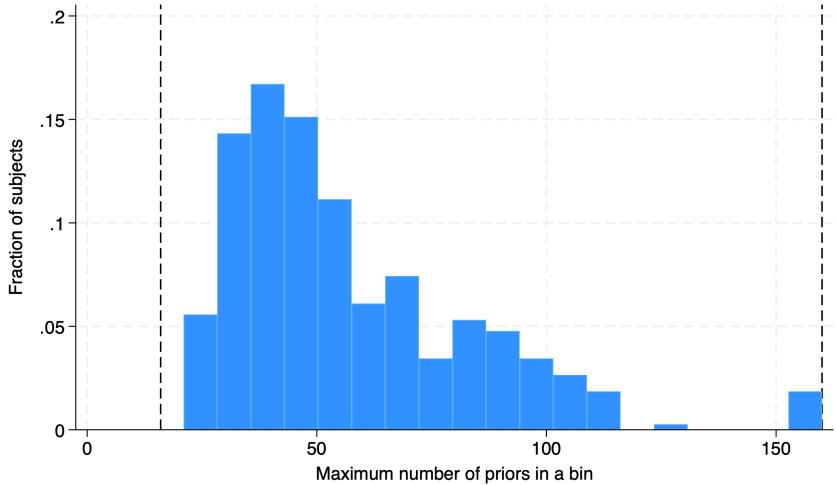
Back

What is the subject-level distribution of the prior $\mu(s_1)$?



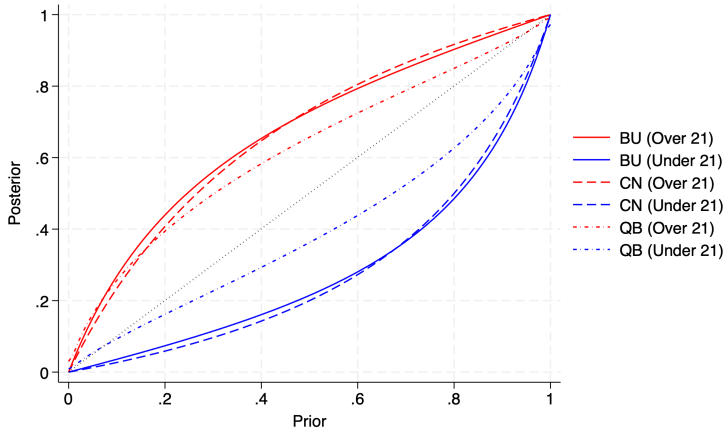
Back

What is the subject-level distribution of the prior $\mu(s_1)$?



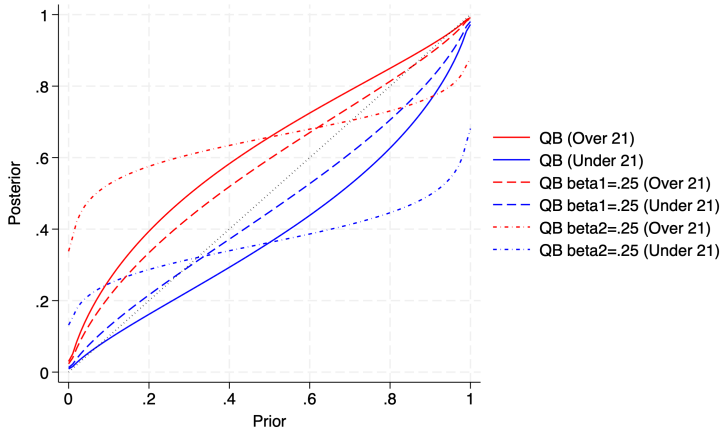
Back

Predictions (BU)



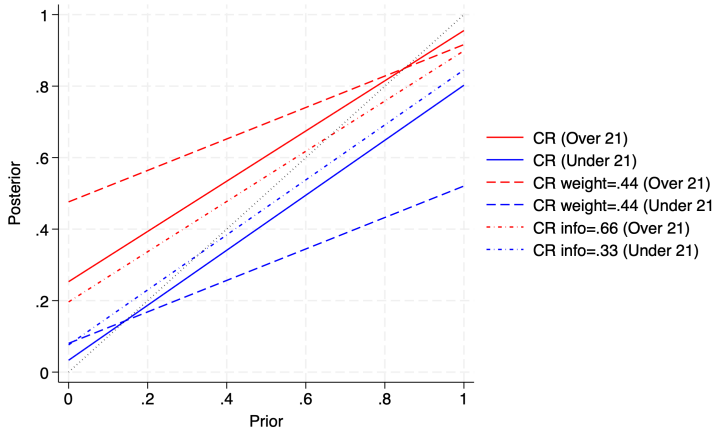
Back

Predictions (QB)



Back

Predictions (CR)



Back

Agarwal, N., Moehring, A., Rajpurkar, P. & Salz, T. (2023), Combining human expertise with artificial intelligence: Experimental evidence from radiology, Technical report, National Bureau of Economic Research.

Caplin, A., Deming, D. J., Li, S., Martin, D. J., Marx, P., Weidmann, B. & Ye, K. J. (2024), The abc's of who benefits from working with ai: Ability, beliefs, and calibration, Technical report, National Bureau of Economic Research.

Dominiak, A., Kovach, M. & Tserenjigmid, G. (2021), Inertial updating with general information, Technical report.

Fudenberg, D., Gao, W. & Liang, A. (2023), 'How flexible is that functional form? quantifying the restrictiveness of theories', *Review of Economics and Statistics* pp. 1–50.

Fudenberg, D., Kleinberg, J., Liang, A. & Mullainathan, S. (2022), 'Measuring the completeness of economic models', *Journal of Political Economy* **130**(4), 956–990.

Grether, D. M. (1980), 'Bayes rule as a descriptive model: The representativeness heuristic', *The Quarterly journal of economics* **95**(3), 537–557.

Hoong, R. & Dreyfuss, B. (2025), 'Improving ai-assisted decision-making through calibrated coarsening', *Available at SSRN 5286198* .

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

Rothe, R., Timofte, R. & Van Gool, L. (2018), 'Deep expectation of real and apparent age from a single image without facial landmarks', *International Journal of Computer Vision* **126**(2), 144–157.

Back