Supplementary Information

A Cognitive Account of Advocacy Bias

Moshe Hoffman[1,*], Evan Russek[2], and Tom Griffiths[3]

[1]Independent Researcher, Boston, MA, USA

[2]Department of Psychology, Hunter College, City University of New York, New York, NY, USA

[3]Departments of Psychology and Computer Science, Princeton University, Princeton, NJ, USA

[*]Corresponding author: hoffman.moshe@gmail.com

Contents

General information and reporting conventions **5**

S1 Study 1: Stick-censoring task **6**

S1.1 Preregistration, sample, and design 6

S1.2 Normative benchmarks 7

S1.3 Preregistered analyses 8

S1.4 Preregistered results 9

S1.5 Non-preregistered proportional-adjustment analysis 10

S2 Study 2: Threshold-censoring task **11**

S2.1 Experimental design and procedure 11

S2.2 Preregistration and planned sample size 12

S2.3 Participants and exclusions 12

S2.4 Naive and Bayesian benchmarks 13

S2.4.1 Naive benchmark 13

S2.4.2 Bayesian benchmark . . . 13
S2.4.3 Block 1 . . . 13
S2.4.4 Block 2 . . . 14
S2.4.5 Block 3 . . . 15
S2.5 Primary benchmark-gap analysis . . . 15
S2.6 Primary results . . . 16
S2.7 Mixed-effects robustness . . . 17
S2.8 Secondary preregistered analyses . . . 18
S2.8.1 Ordering across censoring conditions . . . 18
S2.8.2 Deviation from the Bayesian benchmark . . . 18

S3 Study 3: Updating from selectively presented legal evidence 19

S3.1 Participants, preregistration, and exclusions . . . 19
S3.2 Legal cases and evidence . . . 20
S3.3 Procedure . . . 20
S3.4 Randomization . . . 21
S3.5 Outcomes . . . 22
S3.6 Preregistered hypotheses . . . 22
S3.7 Primary regression specifications . . . 23
S3.8 Primary results . . . 23
S3.9 Robustness analyses . . . 24
S3.10 Results by legal case . . . 24
S3.11 Interpretation . . . 26

S4 Study 4: Policy-dossier reviewers 26

S4.1 Participants, preregistration, and exclusions . . . 26
S4.2 Design and procedure . . . 27
S4.3 Outcome construction . . . 29

S4.4 Preregistered H1: advocacy effects 29
S4.5 Preregistered H2: claim-congenial asymmetric updating 30
S4.6 Preregistered H3: evidence-congenial asymmetric updating 31
S4.7 Preregistered H4: advocate processing and attenuation 32
S4.8 Preregistered robustness analyses 33

S5 Study 5: Legal advocates' selective processing and belief updating 33

S5.1 Participants, preregistration, and exclusions 33
S5.2 Design and procedure 34
S5.3 Process measures 35
S5.4 Outcome construction 35
S5.5 H1: advocacy effects 36
S5.6 H2: asymmetric updating 36
S5.7 Robustness to outcome scale 37
S5.8 H3: selective processing of congenial evidence 38
S5.9 H4: attenuation after controlling for process measures 38
S5.10 Exploratory selective-processing analyses 39
S5.11 Summary of preregistered tests 40

S6 Study 6: Policy advocacy 40

S6.1 Participants, preregistration, and data-quality procedures 40
S6.2 Policy materials 41
S6.3 Procedure 42
S6.4 Outcome orientation and standardization 43
S6.5 Claim alignment and evidence alignment 43
S6.6 H1: advocacy effects 44
S6.7 H2: greater responsiveness to all evidence about advocacy-aligned claims . . 44
S6.8 H3: asymmetric responsiveness to advocacy-aligned evidence 45

S6.8.1 H3: moderation by claim alignment 45
S6.9 H4: selective processing of advocacy-aligned evidence 46
S6.10 H5: selective processing of advocacy-aligned claims 47
S6.11 Process-to-judgment associations 48
S6.12 H6: attenuation after controlling for process measures 49
S6.13 Summary of preregistered tests 50

S7 Supplementary experimental materials **51**
S7.1 Stimulus inventory 51
S7.2 Interface documentation 51
S7.3 Archiving of complete experimental materials 51

S8 Deviations and post-preregistration data-quality procedures **52**
S8.1 Incomplete runs and repeated participation 53
S8.2 Preview and researcher-test sessions 53
S8.3 External-assistance/template screen 53
S8.4 Cross-study exclusion accounting 54
S8.5 Legal-advocate study: correction to the asymmetric-updating contrast 55
S8.6 Legal-advocate study: constituent terms in the process interaction 56
S8.7 Sensitivity analyses and reporting principle 56

General information and reporting conventions

All participants were recruited online through Prolific and provided informed consent before beginning the study. Eligibility for all six studies required a current country of residence of the United States, a Prolific approval rate of at least 95%, and at least 20 previous submissions. We also attempted to exclude anyone who had participated in any earlier study in this manuscript or an associated pilot. In two collections in which different conditions were posted simultaneously, this restriction did not fully prevent participation in more than one condition; repeat participation was identified from Prolific IDs and handled using the procedures described in Section S8. The studies were approved by Princeton's Institutional Review Board (Protocol #10857, Computational Cognitive Science). Study-specific compensation, preregistration information, exclusions, and deviations are reported below. Preregistration links are reported in the corresponding study sections. De-identified data, experimental materials, and analysis code are available at `https://github.com/Moshe-Hoffman/02_REPRODUCIBILITY_REPOSITORY`.

All predictions were directional, and the corresponding hypothesis tests use one-sided p-values. Two-sided p-values are used only for the post hoc completion and balance diagnostics, for which no directional hypothesis was specified. Reported sample sizes refer to unique participants after the study-specific completion and exclusion procedures described below. Repeated-measures studies additionally report the number of participant-trial or participant-claim observations where relevant.

Table S1: Overview of the six experiments.

Study	Setting	Final N	Primary purpose
1	Stick censoring	50	Adjustment to a transparent rank-based selection process; 27 repeated trials per participant.
2	Threshold censoring	116	Adjustment to value-dependent censoring under progressively greater uncertainty about the sampling process.
3	Legal evaluation	399	Updating from mechanically imposed one-sided versus two-sided legal evidence, with and without disclosure.
4	Policy reviewers	288	Downstream effects of dossiers produced by advocates randomly assigned to support or oppose a policy.
5	Legal advocacy	204	Effects of randomized advocacy assignment on advocates' evidence processing and beliefs.
6	Policy advocacy	304	Effects of randomized advocacy assignment on evidence processing, quantitative beliefs, claim beliefs, and policy support.

Preregistration identifiers are reported in the corresponding study sections below. All experi-

ments were programmed in JavaScript using jsPsych 8.0 and administered on Cognition.run. Analyses were conducted in Python 3.13.5 on macOS. The analysis scripts specify the required Python packages (including NumPy, pandas, SciPy, statsmodels, and, where applicable, patsy, Matplotlib, and PyArrow), but exact package versions were not retained.

S1 Study 1: Stick-censoring task

S1.1 Preregistration, sample, and design

The study was preregistered on AsPredicted before data collection (AsPredicted #236,681; July 3, 2025; `https://aspredicted.org/tk5a4e.pdf`). The preregistered hypothesis was that participants would recognize and adjust for the censoring process, but would adjust insufficiently, leaving their estimates influenced by which observations were selected for display. The preregistration fixed the sample size at 50 participants; no formal power analysis or other sample-size rationale was preregistered.

All 50 participants completed the study and none met the preregistered exclusion criterion of failing the comprehension test three or more times. The final sample therefore contained 50 participants and 1,350 participant-trials (27 trials per participant).

On each trial, participants estimated the mean height of 10 sticks. They were told that each stick height was independently drawn from a uniform distribution over the integers 1–100. The sticks were displayed in ascending order, with between one and nine sticks covered. In the Tallest-visible condition, the visible sticks were the tallest in the set; in the Shortest-visible condition, they were the shortest; and in the Random condition, the visible sticks were a randomly selected subset. Participants were told which selection rule applied and could see the ordered positions of both revealed and covered sticks. Each participant completed one trial for every combination of the three censoring rules and nine possible numbers of revealed sticks, for 27 trials total. The 27 trials were presented in random order, and a new set of 10 sticks was generated on every trial.

Participants received a $2 base payment. In addition, they were incentivized for accuracy: they were told that one of the 27 rounds would be selected at random for bonus payment, that they would begin that selected round with a $2 bonus, and that the bonus would be reduced by 4 cents for each point (rounded down) by which their estimate differed from the true mean. The mean bonus paid was $1.56.

Before beginning the task, participants completed a five-item comprehension check covering (i) the number of sticks drawn, (ii) whether all sticks would always be visible, (iii) how the covered sticks could be selected, (iv) the quantity they were asked to estimate, and (v) the accuracy-bonus rule. Participants had to answer all five items correctly before proceeding. After an incorrect attempt, the task identified which questions had been answered incorrectly and directed the participant back to the instructions before another attempt.

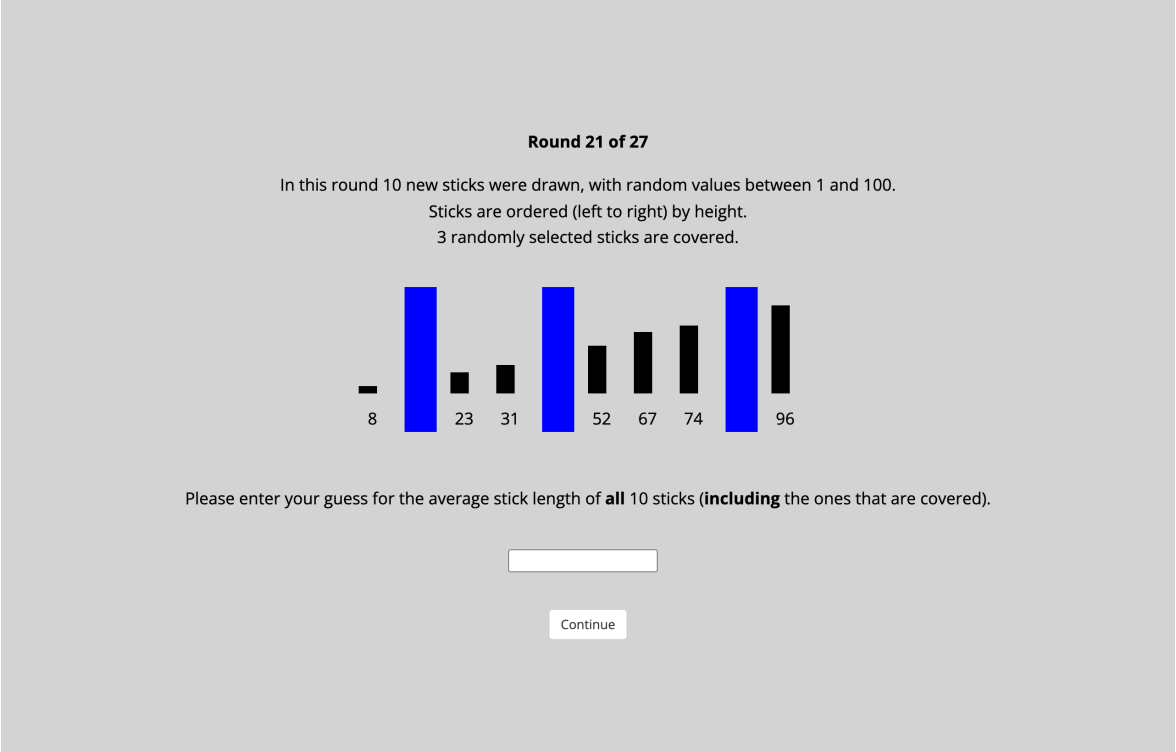


Figure S1: **Study 1 Random-condition judgment screen.** The main-text figure shows corresponding Tallest-visible and Shortest-visible trials. In the Random condition shown here, covered sticks could occupy any of the ten ordered positions. In every condition, participants saw the censoring rule, the revealed values, and the ordered locations of the covered sticks, and estimated the mean of all ten sticks.

S1.2 Normative benchmarks

Let m denote the number of visible sticks, $h = 10 - m$ the number hidden, $\bar{V}$ the mean of the visible sticks, $V_{\min}$ the smallest visible stick, and $V_{\max}$ the largest visible stick.

The preregistration referred to a Bayesian baseline for each censoring condition. For the two selective-censoring conditions, it approximated the expected full-sample mean by assigning the hidden sticks the midpoint of the feasible interval implied by the observed boundary. In Tallest-visible trials,

$$B^{pre} = \frac{m\bar{V} + h\left(\frac{1+V_{\min}}{2}\right)}{10}. \tag{S1}$$

In Shortest-visible trials,

$$B^{pre} = \frac{m\bar{V} + h\left(\frac{V_{\max}+100}{2}\right)}{10}. \tag{S2}$$

For Random trials, the preregistered baseline was

$$B^{pre} = \bar{V}, \tag{S3}$$

which the preregistration explicitly described as an unbiased rather than exact conditional estimate. The preregistration stated that stick heights were uniformly distributed over the integers 1–100 but inadvertently used 0 as the lower endpoint in the Tallest-visible midpoint expression. We replace 0 with 1, consistent with the distribution described to participants. In the two selective-censoring conditions, this correction leaves the reconstructed preregistered baseline very close to the exact discrete posterior benchmark defined below.

For the non-preregistered proportional-adjustment analysis, we calculated the exact posterior expectation implied by the information available to participants. Let $X_{(1)} \leq \cdots \leq X_{(10)}$ denote the ordered values of 10 independent draws from $U\{1, \ldots, 100\}$, let $\mathcal{O}$ denote the set of revealed rank positions, and let y_r denote the value revealed at position r. The exact benchmark is

$$B^{exact} = E\left[\frac{1}{10}\sum_{r=1}^{10} X_{(r)} \;\middle|\; X_{(r)} = y_r \text{ for every } r \in \mathcal{O}\right]. \tag{S4}$$

We evaluated this expectation by dynamic programming over all discrete count vectors consistent with the observed order statistics. For a count vector $c = (c_1, \ldots, c_{100})$, where $\sum_v c_v = 10$, the probability weight of a compatible sorted sample is proportional to

$$\frac{1}{\prod_{v=1}^{100} c_v!}, \tag{S5}$$

because the remaining multinomial and discrete-uniform factors are common across compatible samples and cancel when the conditional expectation is formed. This calculation handles ties exactly and conditions on which ordered positions were revealed or covered. The positional information matters even in the Random condition because, after the sticks are ordered, the ranks of the covered observations provide information about their likely values.

The deployed random-number generator inadvertently used an exclusive upper endpoint and therefore realized values from 1 to 99 rather than the stated range of 1 to 100. We retain the stated distribution for the normative benchmarks because it was the distribution participants were instructed to use. This implementation discrepancy was common to all conditions.

S1.3 Preregistered analyses

The preregistration specified three dependent measures: participants' estimates; estimates minus the mean of the visible sticks; and estimates minus the preregistered Bayesian baseline.

The analyses tested three corresponding predictions. First, estimates should remain higher when the tallest sticks were visible and lower when the shortest sticks were visible. Second, participants should adjust away from the visible-stick mean in the direction implied by the censoring rule. Third, this adjustment should be insufficient relative to the preregistered baseline, producing positive estimate-minus-baseline residuals when the tallest sticks were visible and negative residuals when the shortest sticks were visible. The preregistration further predicted that these effects would diminish as more sticks were revealed.

The preregistration parameterized censoring using m, the number of sticks revealed. For reporting, we use the algebraically equivalent parameterization $h = 10 - m$, the number hidden, and center it at five, $h^c = h - 5$. Let $Tall_{it}$ and $Short_{it}$ indicate Tallest-visible and Shortest-visible trials, with Random omitted. For each preregistered outcome Y_{it}, the fixed-effects portion of the reported model was

$$Y_{it} = \beta_0 + \beta_T Tall_{it} + \beta_S Short_{it} + \beta_h h^c_{it} + \beta_{Th}(Tall_{it} \times h^c_{it}) + \beta_{Sh}(Short_{it} \times h^c_{it}) + u_{it}. \quad \text{(S6)}$$

Because $h^c = 0$ corresponds to five hidden sticks, β_T and β_S capture Tallest-visible and Shortest-visible differences from Random at the midpoint of the design, while the corresponding interactions capture how those differences change as more sticks are hidden.

The preregistration specified participant random intercepts and slopes but did not specify the covariance structure among those random effects or a convergence fallback. In the final implementation, we first estimated a correlated participant random intercept and slopes for the condition indicators, censoring severity, and their interactions. If that model was singular or failed to converge, the same random slopes were retained with their correlations constrained to zero. In the final 1–100 analysis used for the manuscript, the correlated specification converged without a singular covariance for the reported models; participant-fixed-effects regressions with participant-clustered standard errors were additionally used as robustness checks.

S1.4 Preregistered results

Participants strongly adjusted for the censoring process. At five hidden sticks, estimates were 21.54 points below the visible-stick mean in the Tallest-visible condition ($SE = 1.70$, $p < .001$) and 18.67 points above it in the Shortest-visible condition ($SE = 1.70$, $p < .001$). Each additional hidden stick increased the magnitude of these adjustments by 3.76 points in Tallest-visible ($SE = .29$, $p < .001$) and 4.05 points in Shortest-visible ($SE = .27$, $p < .001$).

Nevertheless, the censoring rule continued to affect estimates. At five hidden sticks, estimates were 2.17 points higher in Tallest-visible than Random ($SE = 1.08$, $p = .022$) and 2.43 points lower in Shortest-visible than Random ($SE = .92$, $p = .004$). The Tallest-visible difference increased as more sticks were hidden ($b = .87$, $SE = .31$, $p = .002$); the corresponding Shortest-visible interaction was not significant ($b = -.18$, $SE = .32$, $p = .283$).

The preregistered Bayesian-deviation analysis likewise provided evidence of incomplete adjustment. In the Shortest-visible condition, participants' estimates were 4.04 points below the preregistered baseline at five hidden sticks ($SE = 1.70$, $p = .009$), and this residual became 0.50 points more negative with each additional hidden stick ($SE = .27$, $p = .033$). In the Tallest-visible condition, the residual at five hidden sticks was small and not significant (0.63 points, $SE = 1.76$, $p = .359$), but became 0.65 points more positive per additional hidden stick ($SE = .28$, $p = .009$).

S1.5 Non-preregistered proportional-adjustment analysis

The preregistered Bayesian-deviation analysis expresses incomplete adjustment as an absolute-point residual from the benchmark and tests how that residual varies with censoring. Because the magnitude of the normatively required correction itself varies considerably across trials, however, this specification does not directly estimate the fraction of the required correction that participants make. We therefore conducted an additional, non-preregistered proportional-adjustment analysis using the exact discrete posterior benchmark.

For each participant i and trial t, let E_{it} denote the participant's estimate, V_{it} the mean of the visible sticks, and $B_{it} = B_{it}^{exact}$ the exact posterior expectation of the mean of all 10 sticks. Define

$$A_{it} = E_{it} - V_{it}$$

as the participant's adjustment away from the visible-stick mean and

$$G_{it} = B_{it} - V_{it}$$

as the exact Bayesian adjustment required on that trial. We estimate

$$A_{it} = \alpha_i + \lambda G_{it} + \varepsilon_{it},$$

using participant fixed effects and participant-clustered standard errors. A coefficient of $\lambda = 0$ corresponds to no systematic adjustment in response to the Bayesian correction, whereas $\lambda = 1$ corresponds to one-for-one tracking of the exact Bayesian correction. Values between zero and one therefore indicate adjustment in the normatively appropriate direction but of insufficient magnitude.

Pooling the Tallest-visible and Shortest-visible conditions, we estimate

$$\lambda = .891, \qquad SE = .026.$$

The estimate was strongly greater than zero ($p < .001$) and significantly below one ($p < .001$), indicating substantial but incomplete adjustment. Analyzing the two selective-censoring conditions separately produced the same pattern. In Tallest-visible, $\lambda = .872$ ($SE = .056$, $p_{\lambda<1} = .011$); in Shortest-visible, $\lambda = .897$ ($SE = .057$, $p_{\lambda<1} = .035$).

Thus, across trials requiring different magnitudes of Bayesian correction, participants' judgments moved strongly in the normatively appropriate direction but moved less than one-for-one with the correction required by the selection process.

S2 Study 2: Threshold-censoring task

S2.1 Experimental design and procedure

The truncation study examined inference from selectively recorded observations. Participants completed 30 rounds divided into three blocks of 10 rounds each.

On every round, the latent state s was independently selected from two equally likely machines, $s \in \{H, L\}$. The Low Machine generated observations from a normal distribution with mean 230 and standard deviation 50, and the High Machine generated observations from a normal distribution with mean 270 and standard deviation 50. Generated values were rounded to the nearest integer and bounded between 0 and 500 for presentation. Following the preregistration, the benchmark calculations described below ignore this rounding and bounding and evaluate the observations under the corresponding continuous normal densities.

After the latent machine was selected, N values were sampled and a censoring rule $R \in \{\text{Only High}, \text{Only Low}, \text{Random Drop}\}$ was selected at random. Participants observed only the values surviving the censoring rule and reported, from 0% to 100%, the probability that the record had been generated by the High Machine.

Under **Only High**, only sampled values greater than a cutoff c were recorded. Under **Only Low**, only sampled values less than c were recorded. Under **Random Drop**, each sampled observation was independently dropped with probability 1/3, irrespective of its value.

The blocks differed in the amount of information available about the sampling and censoring processes.

- **Block 1.** Exactly $N = 5$ values were sampled on each round and the cutoff was fixed at $c = 250$. Participants were told both that exactly five values were initially sampled and that the relevant cutoff was 250.
- **Block 2.** The number of sampled values followed the geometric distribution on the positive integers, $P(N = n) = .2(.8)^{n-1}$ for $n = 1, 2, \ldots$, and was not displayed during the judgment round. The cutoff remained fixed at $c = 250$.
- **Block 3.** The number of sampled values again followed $P(N = n) = .2(.8)^{n-1}$ for $n = 1, 2, \ldots$ and remained unobserved. In addition, the cutoff varied independently across rounds according to $c \sim U(230, 270)$ and was not revealed during the judgment.

At the beginning of every block, participants viewed block-specific instructions and could generate an unlimited number of example samples and records before beginning the judgment rounds.

The participant-facing instructions did not explicitly state the geometric parameter governing N in Blocks 2–3 or that the Block-3 cutoff was uniformly distributed within its stated range.

Participants could obtain experience with these processes through the example generators. Consequently, Block 1 supplies the cleanest test of correction for a fully specified censoring process, whereas Blocks 2 and 3 examine whether the same pattern persists as uncertainty about the selection process increases.

Representative Random Drop and Only Low judgment screens appear in Figure 2a,b of the main text. On every judgment round, participants saw the censoring rule and the recorded values but not the selected machine or the complete original sample, and they reported the probability that the record came from the High Machine. In the selective-censoring conditions, the display shaded the region in which values would not have been recorded. The Only High screen used the same layout as the Only Low screen, with the unrecorded region instead appearing below the cutoff.

S2.2 Preregistration and planned sample size

The design, hypotheses, analyses, and sample-size plan were preregistered before data collection on May 4, 2026, on OSF (`https://doi.org/10.17605/OSF.IO/YB92F`). A pilot sample of 10 participants produced a pooled benchmark-gap estimate of approximately $\lambda = .37$ with a participant-clustered standard error of approximately .17. Scaling this standard error implied that 120 completed participants would yield a standard error of approximately .05 and high power to reject both $\lambda = 0$ and $\lambda = 1$. The preregistration therefore targeted $N = 120$ completed participants. It identified the benchmark-gap regression as the primary analysis and specified the mixed-effects, ordering, and deviation-from-Bayesian-benchmark analyses as secondary tests. The power analysis emphasized the pooled partial-adjustment pattern and precision of the block-by-rule estimates; it did not require every individual cell-level coefficient to be statistically significant.

S2.3 Participants and exclusions

Participants received a \$2.75 base payment and could earn up to \$0.60 in performance-based bonuses across the 30 judgment rounds. Each round was worth up to \$0.02. If r denotes the participant's reported probability and b the Bayesian benchmark probability, the round score was

$$\max\left\{0,\, 1 - 4(r-b)^2\right\},$$

and the round bonus equaled \$0.02 multiplied by this score. Round bonuses were summed and the total was rounded to the nearest cent. In the final sample, the mean bonus paid was \$0.41 (unrounded mean = \$0.4069).

The three data batches contained 119 completed sessions on Cognition.run, where completion was defined as providing a response on all 30 judgment rounds. Inspection of Prolific submission identifiers revealed that three `SESSION_ID`s each generated two complete runs of the experiment. In each case, the two runs occurred sequentially within the same Prolific submission and shared the same IP address and browser/operating-system information,

indicating repeat completion by the same participant. We therefore retained the first complete run associated with each Prolific `SESSION_ID` and excluded the three later repeat completions. This yielded a final sample of $N = 116$ participants and 3,480 participant-round observations.

S2.4 Naive and Bayesian benchmarks

Let $Y = \{y_1, \ldots, y_m\}$ denote the m recorded values observed on a round. Let $f_H(y)$ and $f_L(y)$ denote the normal densities under the High and Low Machines, respectively. The prior probability of each machine is 1/2.

S2.4.1 Naive benchmark

The Naive benchmark ignores the censoring process and treats the recorded observations as though they were an uncensored sample directly from the latent machine:

$$\pi^N(H \mid Y) = \frac{\prod_{y \in Y} f_H(y)}{\prod_{y \in Y} f_H(y) + \prod_{y \in Y} f_L(y)}. \tag{S7}$$

This benchmark therefore uses the informational content of the observed values themselves but ignores information contained in the process that determined which values were recorded.

S2.4.2 Bayesian benchmark

The Bayesian benchmark conditions on the censoring process:

$$\pi^B(H \mid Y, R) = \frac{P(Y \mid H, R)}{P(Y \mid H, R) + P(Y \mid L, R)}. \tag{S8}$$

The relevant likelihood differs by block and censoring rule.

S2.4.3 Block 1

In Block 1, both the number of original draws $N = 5$ and the cutoff $c = 250$ are known.

For **Only High**, all unrecorded observations must have fallen at or below the cutoff. Thus, for machine s,

$$P(Y \mid s, R = \text{Only High}) \propto \left[\prod_{y \in Y} f_s(y)\right] F_s(c)^{N-m}, \tag{S9}$$

where $F_s(c) = P_s(X \leq c)$.

For **Only Low**, all unrecorded observations must instead have fallen at or above the cutoff:

$$P(Y \mid s, R = \text{Only Low}) \propto \left[\prod_{y \in Y} f_s(y)\right] [1 - F_s(c)]^{N-m}. \tag{S10}$$

Combinatorial factors that are identical under H and L cancel when the posterior odds are formed.

For **Random Drop**, whether an observation is recorded is independent of its value and of the latent machine. The probabilities associated with recording and dropping values therefore cancel from the posterior likelihood ratio, leaving

$$P(Y \mid s, R = \text{Random Drop}) \propto \prod_{y \in Y} f_s(y). \tag{S11}$$

Consequently,

$$\pi^B(H \mid Y, \text{Random Drop}) = \pi^N(H \mid Y). \tag{S12}$$

S2.4.4 Block 2

In Block 2, N is unobserved and follows the data-generating distribution on the positive integers

$$P(N = n) = .2(.8)^{n-1}, \qquad n = 1, 2, \ldots. \tag{S13}$$

The likelihood therefore integrates over possible original sample sizes. For a selective censoring rule,

$$P(Y \mid s, R) = \sum_{n=m}^{\infty} P(N = n) P(Y \mid s, R, N = n), \tag{S14}$$

where

$$P(Y \mid s, R, N = n) = \binom{n}{m} \left[\prod_{y \in Y} f_s(y) \right] q_s(R, c)^{n-m}, \tag{S15}$$

with

$$q_s(\text{Only High}, c) = F_s(c) \tag{S16}$$

and

$$q_s(\text{Only Low}, c) = 1 - F_s(c). \tag{S17}$$

The Bayesian posterior is obtained by evaluating this likelihood separately under the High and Low Machines and applying Bayes' rule.

S2.4.5 Block 3

In Block 3, both the original number of draws and the cutoff are unobserved. The cutoff follows

$$c \sim U(230, 270). \tag{S18}$$

The likelihood therefore integrates over both sources of uncertainty:

$$P(Y \mid s, R) = \sum_{n=m}^{\infty} P(N = n) \int_{230}^{270} P(Y \mid s, R, N = n, c) \frac{1}{40} \, dc. \tag{S19}$$

The conditional likelihood $P(Y \mid s, R, N = n, c)$ is defined as in Block 2, with likelihood equal to zero for cutoff values inconsistent with the observations actually recorded.

In the analysis implementation, the infinite sum over N was truncated at $N = 200$. The Block-3 cutoff integral was evaluated numerically over a 401-point grid using trapezoidal integration. Calculations were performed in log space to avoid numerical underflow.

S2.5 Primary benchmark-gap analysis

The primary preregistered prediction was that participants would adjust away from the Naive benchmark in the direction of the Bayesian benchmark, but would not adjust fully.

For participant-round i, define

$$G_i = \pi_i^B - \pi_i^N \tag{S20}$$

as the distance between the Bayesian and Naive benchmarks, and

$$A_i = \hat{p}_i - \pi_i^N \tag{S21}$$

as the participant's actual adjustment away from the Naive benchmark.

The preregistered Block × Censoring Rule specification is

$$A_i = \sum_{b,r} \lambda_{br} G_i \, 1\{B_i = b, R_i = r\} + \alpha_j + \epsilon_i, \tag{S22}$$

where j indexes participants, $b \in \{1, 2, 3\}$, and $r \in \{\text{Only High}, \text{Only Low}\}$. Participant fixed effects α_j absorb stable participant-level differences, and standard errors are clustered by participant.

The coefficient λ_{br} has a direct interpretation. $\lambda_{br} = 0$ corresponds to no adjustment away from the Naive benchmark, whereas $\lambda_{br} = 1$ corresponds to one-for-one tracking of the adjustment required to reach the Bayesian benchmark, allowing for participant-specific offsets. The preregistered hypothesis therefore predicts

$$0 < \lambda_{br} < 1. \tag{S23}$$

Both inequalities were evaluated separately.

For ease of summarizing the overall effect, we additionally estimate the pooled specification

$$A_i = \lambda G_i + \alpha_j + \epsilon_i \tag{S24}$$

across all Only High and Only Low observations.

S2.6 Primary results

All six Block × Censoring Rule estimates lie between the Naive and Bayesian benchmarks.

For every cell, the tests of $\lambda > 0$ and $\lambda < 1$ were significant at $p < .001$.

Pooling across blocks and the two selective-censoring conditions produced

$$\hat{\lambda} = .369, \qquad SE = .044. \tag{S25}$$

Table S2: Benchmark-gap estimates by block and censoring rule.

Block	Censoring rule	$\hat{\lambda}$	Clustered SE
1	Only High	.286	.065
1	Only Low	.358	.065
2	Only High	.428	.093
2	Only Low	.396	.098
3	Only High	.407	.077
3	Only Low	.415	.088

The pooled coefficient was strongly greater than zero ($p < .001$) and strongly less than one ($p < .001$). Participants therefore made approximately 37% of the adjustment from the Naive benchmark to the fully informed Bayesian benchmark on average.

The cell-level results also establish that the aggregate pattern was not driven by the more difficult later blocks. In Block 1, where N, the cutoff, and the censoring rule were all specified, the estimates were $\hat{\lambda} = .286$ under Only High and $\hat{\lambda} = .358$ under Only Low. Thus, participants substantially under-corrected even when the relevant censoring process was fully specified.

S2.7 Mixed-effects robustness

As preregistered, we also estimated a linear mixed-effects model allowing participants to differ in their intercepts and in their responsiveness to the Bayes-minus-Naive benchmark gap.

The pooled mixed-effects estimate was

$$\hat{\lambda} = .364, \qquad SE = .045, \tag{S26}$$

closely matching the participant-fixed-effects estimate. The substantive inference was therefore unchanged.

Participant-specific pooled λ estimates were also calculated descriptively. These estimates are noisy because each participant contributed only a limited number of selective-censoring trials. The median participant-specific estimate was .344. Approximately 68% of individual estimates fell between 0 and 1. We therefore treat their distribution as descriptive rather than as an additional hypothesis test.

S2.8 Secondary preregistered analyses

S2.8.1 Ordering across censoring conditions

Incomplete adjustment predicts that average judgments should be shifted upward under Only High and downward under Only Low relative to Random Drop. We estimated OLS regressions of participants' reported probability of the High Machine on indicators for Only High and Only Low, with Random Drop as the omitted condition and participant-clustered standard errors.

The pooled estimates were

$$\text{Only High} - \text{Random Drop} = .117, \qquad SE = .017, \tag{S27}$$

and

$$\text{Only Low} - \text{Random Drop} = -.144, \qquad SE = .017. \tag{S28}$$

Both directional predictions were supported ($p < .001$).

The same ordering was observed separately in every block:

Table S3: Ordering contrasts by block.

Block	Only High − Random Drop	SE	Only Low − Random Drop	SE
1	.130	.025	-.150	.025
2	.136	.024	-.119	.026
3	.085	.024	-.163	.025

All six preregistered directional contrasts were significant at $p < .001$.

S2.8.2 Deviation from the Bayesian benchmark

The preregistration additionally predicted that participants would deviate more from the Bayesian benchmark under selective censoring than under Random Drop, where the Naive and Bayesian benchmarks coincide.

For each participant-round we calculated

$$D_i = \left| \hat{p}_i - \pi_i^B \right|. \tag{S29}$$

Regressing D_i on censoring-condition indicators and block fixed effects produced pooled selective-censoring effects of

$$\text{Only High} - \text{Random Drop} = .065, \qquad SE = .010, \tag{S30}$$

and

$$\text{Only Low} - \text{Random Drop} = .076, \qquad SE = .011. \tag{S31}$$

Both were positive as predicted ($p < .001$).

The effects were also positive within every block:

Table S4: Excess absolute deviation from the Bayesian benchmark by block and censoring rule.

Block	Censoring rule	Excess absolute deviation	SE
1	Only High	.085	.017
1	Only Low	.090	.018
2	Only High	.054	.016
2	Only Low	.068	.017
3	Only High	.055	.014
3	Only Low	.071	.014

All six block-specific effects were significant in the preregistered direction at $p < .001$.

Together, these analyses reinforce the interpretation of the primary benchmark-gap estimates. Participants did not simply ignore censoring: the censoring rule systematically shifted their judgments in the appropriate direction. But their judgments remained farther from the fully informed Bayesian benchmark precisely when observation depended on the values being sampled.

S3 Study 3: Updating from selectively presented legal evidence

S3.1 Participants, preregistration, and exclusions

Participants were recruited through Prolific under the paper-wide eligibility criteria described above. Participants received a $1.60 base payment and no bonus. The study was approved by Princeton's Institutional Review Board (Protocol #10857, Computational Cognitive Science), and all participants provided informed consent before beginning the task.

The design and primary analyses were preregistered before collection of the confirmatory sample on June 19, 2026, on OSF (`https://doi.org/10.17605/OSF.IO/ARZ8N`). The pre-

registered target sample was approximately 400 participants, with approximately equal assignment to the five experimental conditions.

We defined complete runs as sessions containing both pre-evidence and post-evidence judgments. Incomplete runs were excluded first. We then identified repeated participation among completed runs using Prolific participant IDs and retained each participant's earliest completed run. The final sample contained $N = 399$ participants. The final treatment-cell sizes were 94 in Defendant-Only Undisclosed, 68 in Defendant-Only Disclosed, 93 in Both-Sides, 72 in Plaintiff-Only Disclosed, and 72 in Plaintiff-Only Undisclosed.

S3.2 Legal cases and evidence

Each participant was randomly assigned to evaluate one of three anonymized legal disputes: a product-liability dispute involving silicone implants, an employment-discrimination dispute involving promotion and termination decisions, or a consumer-fraud dispute involving product-performance marketing. The task presented each dispute in the format of a legal case, with identifying details removed or replaced by generic descriptors and the materials simplified and modularized for experimental use.

The product-liability stimulus was based on *In re Silicone Gel Breast Implants Products Liability Litigation* (MDL 926). The employment-discrimination and consumer-fraud materials were constructed as representative, anonymized scenarios modeled on real employment-discrimination and consumer-protection litigation patterns, respectively, rather than as reproductions of single identified cases.

The judge experiment used a pruned evidence set containing one plaintiff-favoring and one defendant-favoring fact for each disputed claim. Thus, participants in a one-sided condition saw one fact total per disputed claim, whereas participants in the Both-Sides condition saw two facts per disputed claim—one favoring each side. This differs from the separate advocate experiment, which used a different evidence-presentation structure.

S3.3 Procedure

After consent and a task overview, participants read the summary of their randomly assigned case and the disputed factual propositions. Before reviewing any evidence, they reported an initial probability judgment for every disputed claim and for the overall case outcome.

For each disputed claim, participants reported on a 0–100 scale how likely it was that the judge or jury accepted the claim as more likely true than not. They separately estimated, from 0 to 100, the probability that the plaintiff prevailed in the case.

Participants were then randomly assigned to one of five evidence conditions:

In the disclosed one-sided conditions, participants were told that they would review "one piece

Table S5: Evidence conditions in the legal-judge experiment.

Condition	Evidence shown	One-sided nature disclosed?
T1: Both-Sides	One plaintiff and one defendant fact per claim	Not applicable
T2: Plaintiff-Only Disclosed	One plaintiff-favoring fact per claim	Yes
T3: Defendant-Only Disclosed	One defendant-favoring fact per claim	Yes
T4: Plaintiff-Only Undisclosed	One plaintiff-favoring fact per claim	No
T5: Defendant-Only Undisclosed	One defendant-favoring fact per claim	No

of evidence for each disputed point, selected from the evidence presented by" the plaintiff or defendant in the case. The evidence-review screen reminded them that they were not seeing the opposing side's evidence and therefore that relevant counterevidence might be missing. Before the post-evidence judgments, participants were again reminded that they had reviewed only one side's evidence and were instructed to take into account the evidence that the opposing side might have presented.

In the undisclosed one-sided conditions, participants likewise saw one fact per disputed claim, all favoring the same side, but the interface did not identify the evidence as coming from that side or remind participants that opposing evidence was missing. In the Both-Sides condition, participants saw one piece of evidence from each side for each disputed claim.

Participants had to open each displayed evidence card for at least two seconds before continuing. After reviewing the evidence, they repeated the disputed-claim and overall-verdict judgments using the same 0–100 scales.

Figure 3a of the main text illustrates the claim-by-claim evidence-review interface in a one-sided condition. Evidence was organized under the disputed proposition to which it pertained, and participants opened each evidence item before proceeding. The additional disclosure and reminder language described above was presented only in the disclosed one-sided conditions.

S3.4 Randomization

Participants were randomly assigned to one of the three cases. Experimental assignment produced five approximately equally likely evidence conditions. The study first randomized among the one-sided disclosed, one-sided undisclosed, and Both-Sides structures with probabilities 2:2:1; for one-sided assignments, the displayed side was independently randomized between plaintiff and defendant with probability .5. This procedure gives each of the five final treatment cells equal ex ante probability.

S3.5 Outcomes

For participant i and disputed claim c, the primary claim-level updating outcome was

$$\Delta_{ic} = Y_{ic}^{post} - Y_{ic}^{pre}, \tag{S32}$$

where larger values indicate movement toward the plaintiff on the 0–100 judgment scale.

For the overall case judgment,

$$\Delta_{i} = Y_{i}^{post} - Y_{i}^{pre}, \tag{S33}$$

where larger values indicate movement toward believing that the plaintiff prevailed.

The main-text figure reports participant-level descriptive means. For the claim series, claim-level changes were averaged within participant before treatment means and confidence intervals were computed. The preregistered inferential analysis retained the individual claim observations and clustered standard errors by participant.

S3.6 Preregistered hypotheses

Let T1 denote Both-Sides, T2 Plaintiff-Only Disclosed, T3 Defendant-Only Disclosed, T4 Plaintiff-Only Undisclosed, and T5 Defendant-Only Undisclosed.

The preregistration specified three related predictions. First, updating toward the plaintiff was predicted to follow the ordering

$$T5 < T3 < T1 < T2 < T4. \tag{S34}$$

Second, disclosure of the one-sided nature of the evidence was predicted to attenuate the influence of one-sided evidence:

$$T4 > T2 \qquad \text{and} \qquad T3 > T5. \tag{S35}$$

Third, correction was predicted to remain incomplete even after disclosure:

$$T2 > T1 \qquad \text{and} \qquad T1 > T3. \tag{S36}$$

The second prediction tests whether participants respond at all to information about evidence selection; the third tests whether this response is sufficient to bring their judgments back to the benchmark produced by observing evidence from both sides.

S3.7 Primary regression specifications

For claim-level updating, we estimated

$$\Delta_{ic} = \beta_2 T2_i + \beta_3 T3_i + \beta_4 T4_i + \beta_5 T5_i + \gamma_c + \epsilon_{ic}, \tag{S37}$$

where T1 (Both-Sides) is the omitted category and γ_c denotes disputed-claim fixed effects. Standard errors were clustered at the participant level.

For overall-verdict updating, we estimated

$$\Delta_i = \beta_2 T2_i + \beta_3 T3_i + \beta_4 T4_i + \beta_5 T5_i + \gamma_{\text{case}} + \epsilon_i, \tag{S38}$$

where γ_{case} denotes case fixed effects. Heteroskedasticity-robust standard errors were used, with $\alpha = .05$.

S3.8 Primary results

Table S6 reports the raw participant-level treatment means. The broad pattern followed the predicted progression from defendant-only evidence through Both-Sides evidence to plaintiff-only evidence. Three of the four preregistered adjacent comparisons were statistically significant; the exception was the plaintiff-side disclosure contrast (T4 versus T2).

Table S6: Mean pre-to-post updating by randomized evidence condition. Positive values indicate movement toward the plaintiff. Claim updating is first averaged across disputed claims within participant.

Condition	N	Mean claim Δ	Mean verdict Δ
T5: Defendant-Only Undisclosed	94	-29.87	-32.10
T3: Defendant-Only Disclosed	68	-19.11	-19.66
T1: Both-Sides	93	-3.30	-4.44
T2: Plaintiff-Only Disclosed	72	18.98	20.39
T4: Plaintiff-Only Undisclosed	72	20.34	18.78

For claim updating, disclosure significantly reduced the effect of defendant-only evidence: T3 was 11.76 percentage points more plaintiff-favoring than T5 (SE = 3.29, $p < .001$). The corresponding plaintiff-side disclosure contrast was small and nonsignificant: T4 − T2 = 1.51 points (SE = 2.76, $p = .292$).

The preregistered incomplete-correction contrasts were both strongly supported. Relative to Both-Sides, disclosed plaintiff-only evidence shifted claim judgments an additional 22.69 points toward the plaintiff (SE = 2.73, $p < .001$). Disclosed defendant-only evidence shifted

claim judgments 15.26 points farther toward the defendant than Both-Sides evidence (SE = 3.03, $p < .001$).

The same pattern appeared for verdict updating. Disclosure reduced the effect of defendant-only evidence by 13.24 points (SE = 4.38, $p = .001$), whereas the plaintiff-side disclosure contrast was nonsignificant (T4 − T2 = -2.06, SE = 3.41, $p = .726$). Despite explicit disclosure, plaintiff-only evidence shifted verdict judgments 25.03 points farther toward the plaintiff than Both-Sides evidence (SE = 3.06, $p < .001$), and defendant-only disclosed evidence shifted verdict judgments 14.97 points farther toward the defendant than Both-Sides evidence (SE = 3.87, $p < .001$).

The absence of a detectable plaintiff-side disclosure effect was not predicted. One possible post hoc explanation is that, in at least some cases, participants could infer that the displayed evidence was one-sided from the consistent direction of its content even without explicit disclosure, reducing the incremental effect of the disclosure manipulation. The study did not elicit participants' beliefs about the evidence-selection process, so this explanation cannot be directly tested.

S3.9 Robustness analyses

The preregistration specified two alternative outcome specifications. First, probability judgments were rescaled to the unit interval, winsorized away from 0 and 1, transformed to log odds, and analyzed using pre-to-post changes in logit probability. Second, post-evidence judgments were regressed directly on treatment assignment while controlling for the corresponding pre-evidence judgment (ANCOVA).

The primary conclusions were unchanged. In both the logit-change and ANCOVA specifications, disclosure continued to significantly attenuate defendant-only evidence for both claim and verdict judgments, whereas the plaintiff-side disclosure contrast remained nonsignificant. Most importantly, both incomplete-correction contrasts—Plaintiff-Only Disclosed versus Both-Sides and Both-Sides versus Defendant-Only Disclosed—remained strongly significant for both claim and verdict judgments.

S3.10 Results by legal case

The incomplete-correction pattern was not attributable to a single legal domain. Figure S2 reports claim and verdict updating separately for the product-liability, employment-discrimination, and consumer-fraud cases. In each case, disclosed plaintiff-only evidence produced more plaintiff-favoring updating than Both-Sides evidence, while disclosed defendant-only evidence produced more defendant-favoring updating than Both-Sides evidence. Because treatment-cell sizes become relatively small after disaggregating by case, these estimates are presented descriptively rather than as separately powered confirmatory tests.

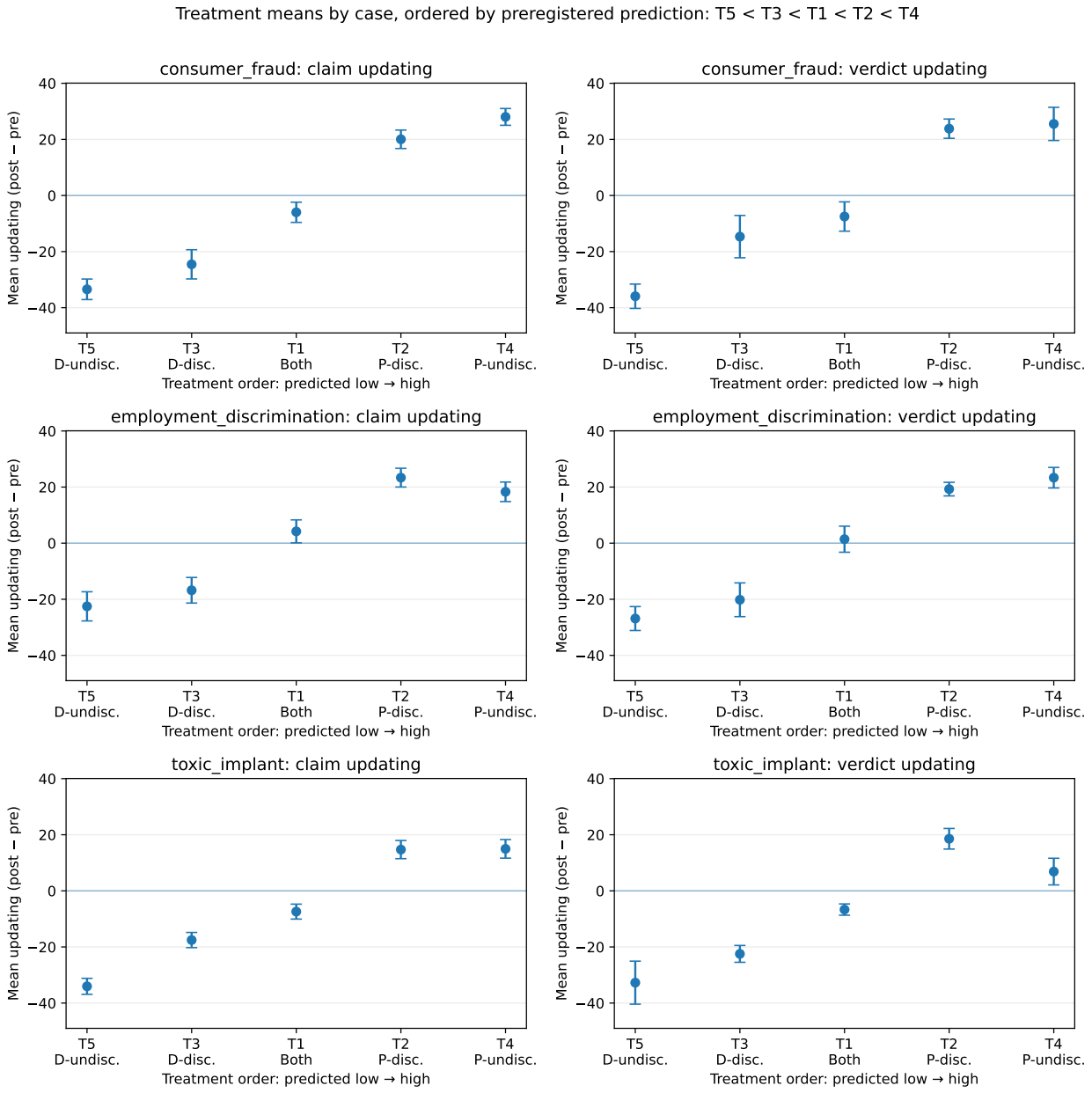


Figure S2: **Evidence-condition effects separately by legal case.** Mean pre-to-post updating is shown for each of the five randomized evidence conditions, separately for disputed-claim and overall-verdict judgments and for each of the three legal disputes. Positive values indicate movement toward the plaintiff and negative values movement toward the defendant. Error bars are normal-approximation 95% confidence intervals calculated from participant-level means. The preregistered analyses pool across cases and include claim or case fixed effects as specified above.

S3.11 Interpretation

The study provides evidence for both correction and persistent influence after disclosure. Participants did not invariably ignore information about selection: explicit disclosure substantially reduced the influence of defendant-favoring evidence. Even under unusually explicit disclosure—participants were told that they had received evidence from only one side, reminded that relevant opposing evidence might be missing, and asked to account for that omission—their judgments nevertheless remained substantially shifted toward the side whose evidence they observed.

Unlike Studies 1 and 2, Study 3 was not designed to derive a formal normative benchmark for participants' judgments. Disclosure identified why the displayed facts had been selected, but participants were not told the complete evidentiary pool from which they came. For example, a participant who expected little credible evidence for the plaintiff could rationally update toward the plaintiff after encountering strong plaintiff-favoring facts, even while accounting for their one-sided selection. The study therefore cannot by itself distinguish incomplete adjustment from rational updating based on the existence or strength of the particular facts presented. We interpret the persistence of the one-sided-evidence effect after disclosure as convergent, rather than standalone, evidence for incomplete adjustment.

The experiment does not manipulate evidence selection by an actual advocate. Rather, it recreates an informational environment that advocacy can naturally generate: a selectively one-sided evidentiary record. The results therefore isolate the audience-side consequence of such selection and show that explicit awareness of one-sided presentation need not be sufficient to eliminate its influence.

S4 Study 4: Policy-dossier reviewers

S4.1 Participants, preregistration, and exclusions

Reviewers were recruited through Prolific under the paper-wide eligibility criteria described above. Participants received a $2.40 base payment and could earn up to $1.00 in the quantitative-accuracy bonus described below. All participants provided informed consent under Princeton Institutional Review Board Protocol #10857 (Computational Cognitive Science).

The study was preregistered before data collection on August 19, 2026, on OSF (`https://doi.org/10.17605/OSF.IO/JA4FY`). The preregistration specified that each advocate dossier would be evaluated by one independent reviewer who was fully informed of the advocate's assigned role.

The raw reviewer exports contained 323 distinct study sessions. Thirty-two sessions did not reach the preregistered primary completion criterion of providing the post-review quantitative judgments. Among the remaining 291 completed sessions, three were later completed

participations by individuals who had already contributed an earlier eligible completed run. Following the cross-study repeat-participation rule described in Section S8, we retained each participant's earliest eligible completed run and excluded the three later completions. The final sample therefore contained $N = 288$ unique reviewers matched to 288 advocate dossiers.

We conducted post hoc balance checks of completion using linear-probability models with policy fixed effects and the numbers of pro- and anti-policy evidence bundles randomly shown to the advocate. All p-values in these diagnostic checks are two-sided. Across all 323 reviewer sessions, completion rates were 89.7% for dossiers produced by support advocates and 90.5% for dossiers produced by oppose advocates; the adjusted difference was -0.009 ($SE = 0.033,\ p = .779$). The two evidence-exposure counts jointly predicted session completion ($p = .009$), driven by higher completion for dossiers whose advocates had been shown more anti-policy bundles ($b = 0.052,\ SE = 0.020,\ p = .009$); the pro-policy bundle count was unrelated to completion ($p = .909$). Because incomplete sessions were generally reassigned, we separately examined final coverage of all 304 advocate dossiers eligible for review. A reviewer was obtained for 93.5% of support dossiers and 96.0% of oppose dossiers; the adjusted difference was -0.026 ($SE = 0.025,\ p = .308$). At this dossier level, the two evidence-exposure counts were not jointly associated with obtaining a reviewer at the conventional 5% threshold ($p = .100$), although the anti-policy bundle count retained a small positive association ($b = 0.035,\ SE = 0.017,\ p = .037$). The primary reviewer analyses condition on the advocate's randomized pro- and anti-policy evidence exposure.

S4.2 Design and procedure

Each reviewer evaluated one dossier produced by an advocate in Study 6. Dossiers concerned either increasing the minimum wage or adopting a carbon-pricing policy in the United States and consisted of the advocate's written argument together with the facts the advocate had chosen to include. Advocates had been randomly assigned to argue either for or against the policy.

Before viewing the dossier, reviewers read a brief policy summary and the five disputed factual claims used in the advocate study. They then reported (i) how warranted they considered each claim on a 0–100 scale (0 = not warranted at all; 100 = completely warranted), (ii) their overall policy support on a 0–100 oppose–support scale, and (iii) numerical estimates of measurable real-world outcomes associated with the claims. Each quantitative item specified its units and allowable response range. Benchmark values were used to score accuracy but were not revealed to reviewers.

For example, one minimum-wage item asked, "Suppose the minimum wage were increased by 10%. Approximately what percentage would restaurant prices increase over the following several years?" Reviewers entered a numerical estimate from 0 to 20%, in units of percentage increase in restaurant prices. The scoring benchmark was 0.36%, based on the reported restaurant-price pass-through associated with a 10% minimum-wage increase in the cited research. Complete question wording, response ranges, benchmark values, and benchmark

sources are included in the archived machine-readable stimulus materials described in the Supplementary experimental materials section.

The final deployed task stated that reviewers could earn up to $1.00 from the quantitative-estimate questions: up to $0.50 from the pre-review block and an additional $0.50 from the post-review block. In the deployed task code, item accuracy was scored as

$$Accuracy_q = \max\left\{0,\, 1 - \frac{|Response_q - Benchmark_q|}{Max_q - Min_q}\right\}, \tag{S39}$$

and the bonus for each block equaled the block maximum multiplied by mean item accuracy. This was also the rule used for final payment: reviewers could receive up to $0.50 for the pre-review quantitative estimates and up to $0.50 for the post-review quantitative estimates.

Before reviewing the dossier, reviewers were explicitly told whether the advocate had been assigned to argue in favor of or against the policy and that the advocate had been asked to choose from the available facts those that would help support the assigned position. A two-item comprehension check then asked reviewers (i) which position the advocate had been assigned to argue and (ii) what the advocate had been asked to do with the available facts. Incorrect responses received corrective feedback, and reviewers had to answer both items correctly before proceeding.

On the dossier-review screen, reviewers were required to open and review the advocate's argument for at least five seconds and each included fact for at least two seconds before they could continue. Facts displayed their substantive text, and reviewers could select "View source" to reveal the corresponding source name and URL. The interface recorded opening and viewing behavior for the argument and included facts.

Reviewers then repeated the claim-warrantedness, policy-support, and quantitative judgments. The post-review instructions reminded them to take into account the facts and argument they had reviewed, as well as the advocate's assigned task. While answering the post-review questions, reviewers could optionally reopen the advocate's argument and included facts. The transition to the final quantitative-estimate block additionally reminded reviewers that they could earn up to $0.50 based on the accuracy of those final estimates.

The reviewer-facing disclosure screen is shown in Figure S3. It explicitly stated both the advocate's assigned position and that the advocate had been asked to choose facts that would help support that position. Figure 4a of the main text shows the subsequent review screen, which presented the facts transmitted by the advocate, organized by claim, together with the advocate's written argument.

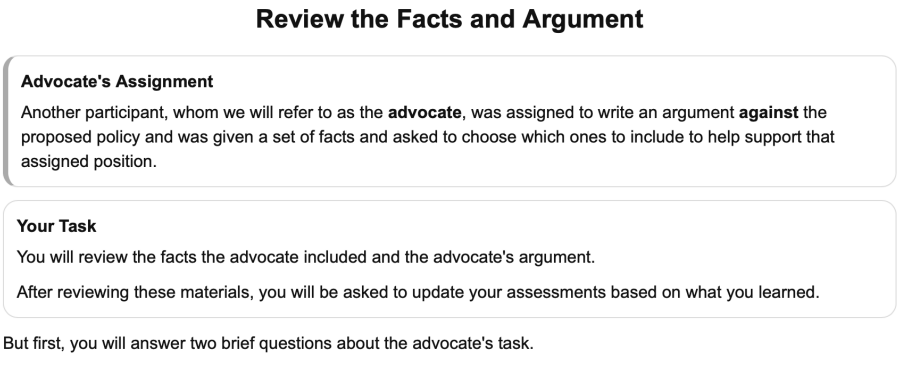


Figure S3: **Study 4 disclosure of the advocate's assignment and selection task.** Before reviewing the dossier, participants were explicitly told the advocate's randomly assigned position and that the advocate had been asked to choose facts useful for supporting it. The example concerns an advocate assigned to oppose the minimum-wage policy.

S4.3 Outcome construction

Reviewer outcomes were constructed analogously to the corresponding Study 6 outcomes. Claim-warrantedness and quantitative responses were first oriented so that larger values indicated greater policy support. For each outcome, we calculated the post-minus-pre change. Following the preregistration, claim-warrantedness changes were standardized within policy-specific claim, quantitative changes within policy-specific quantitative question and then averaged to the corresponding claim when multiple questions measured the same claim, and policy-support changes within policy.

Let $Adv_a = 1$ if advocate dossier a was produced by an advocate assigned to support the policy and $Adv_a = 0$ if the advocate was assigned to oppose it. Let $Pro_{ap\ell}^{shown}$ and $Con_{ap\ell}^{shown}$ indicate whether advocate a had been randomly shown the pro-policy and anti-policy evidence bundles for claim ℓ.

S4.4 Preregistered H1: advocacy effects

For reviewer claim-warrantedness updating, the preregistered model was

$$\Delta W_{arp\ell}^{R} = \alpha + \beta_A Adv_a + \beta_{Pro} Pro_{ap\ell}^{shown} + \beta_{Con} Con_{ap\ell}^{shown} + \eta_{p\ell} + \epsilon_{arp\ell}, \tag{S40}$$

where $\eta_{p\ell}$ are policy-specific claim fixed effects. The corresponding quantitative model replaced ΔW^R with standardized quantitative updating. Policy-support analyses used the number of pro-policy and anti-policy evidence bundles shown to the advocate across the five

claims and included policy fixed effects. Claim-level and quantitative models used standard errors clustered by advocate dossier. H1 predicted $\beta_A > 0$: dossiers generated by advocates assigned to support the policy should shift reviewers more strongly toward policy support than dossiers generated by advocates assigned to oppose it.

The advocate's randomized assignment significantly shifted reviewer judgments in the predicted direction despite explicit disclosure of that assignment. Controlling for the evidence the advocate had been randomly shown, assignment to support rather than oppose the policy shifted reviewer claim warrantedness by 0.383 SD ($SE = 0.070$, $p < .001$), quantitative estimates by 0.207 SD ($SE = 0.050$, $p < .001$), and overall policy support by 0.524 SD ($SE = 0.117$, $p < .001$).

On the original scales, reviewers of anti-policy dossiers changed their claim judgments by an average of 2.5 percentage points in an anti-policy direction, whereas reviewers of pro-policy dossiers moved 6.7 points in a pro-policy direction. Overall policy support changed by -4.6 points among reviewers of anti-policy dossiers and +3.9 points among reviewers of pro-policy dossiers.

S4.5 Preregistered H2: claim-congenial asymmetric updating

A claim was defined as *claim-congenial* for dossier a when accepting that claim favored the advocate's assigned position. Let $CC_{a\ell}$ indicate claim congeniality. The preregistered claim-level model added $CC_{a\ell}$ and its interactions with the advocate's randomized pro- and anti-policy evidence exposure:

$$\begin{aligned}\Delta Y^{R}_{arp\ell} = \alpha + \beta_A Adv_a + \beta_{Pro} Pro^{shown}_{ap\ell} + \beta_{Con} Con^{shown}_{ap\ell} + \beta_C CC_{a\ell} \\ + \gamma_{ProCC}(Pro^{shown}_{ap\ell} \times CC_{a\ell}) + \gamma_{ConCC}(Con^{shown}_{ap\ell} \times CC_{a\ell}) + \eta_{p\ell} + \epsilon_{arp\ell}. \quad \text{(S41)}\end{aligned}$$

With outcomes coded in the policy-supportive direction, the preregistered component predictions were $\gamma_{ProCC} > 0$ and $\gamma_{ConCC} < 0$, with pooled contrast $\gamma_{ProCC} - \gamma_{ConCC} > 0$. An analogous policy-level specification used the numbers of pro- and anti-policy bundles shown on claim-congenial claims.

None of the preregistered H2 tests provided evidence that evidence exposure exerted greater influence when it concerned a claim congenial to the advocate's assigned side. Table S7 reports the component and pooled directional tests.

Table S7: **Study 4 preregistered H2 tests: claim-congenial asymmetric updating.** Entries report coefficient, clustered standard error, and one-sided p-value for the preregistered directional test. The pooled contrast is $\gamma_{ProCC} - \gamma_{ConCC}$.

Outcome	Pro × claim-congenial	Con × claim-congenial	Pooled contrast
Claim warrantedness	0.031 (0.118), $p = .398$	0.053 (0.125), $p = .663$	−0.022 (0.123), $p = .571$
Quantitative estimates	0.049 (0.123), $p = .344$	−0.132 (0.120), $p = .135$	0.182 (0.117), $p = .061$
Policy support	0.028 (0.114), $p = .402$	0.041 (0.123), $p = .631$	−0.013 (0.108), $p = .547$

S4.6 Preregistered H3: evidence-congenial asymmetric updating

H3 tested whether reviewers were more responsive to evidence exposure that favored the advocate's assigned side. At the claim level, the preregistered specification was

$$\begin{aligned}\Delta Y^{R}_{arp\ell} = \alpha + \beta_A Adv_a + \beta_{Pro} Pro^{shown}_{ap\ell} + \beta_{Con} Con^{shown}_{ap\ell} \\ + \gamma_{Pro}(Adv_a \times Pro^{shown}_{ap\ell}) + \gamma_{Con}(Adv_a \times Con^{shown}_{ap\ell}) + \eta_{p\ell} + \epsilon_{arp\ell}.\end{aligned} \tag{S42}$$

The preregistered predictions were $\gamma_{Pro} > 0$, $\gamma_{Con} > 0$, and the pooled contrast $\gamma_{Pro} + \gamma_{Con} > 0$. Policy-support analyses used the corresponding interactions between advocacy assignment and counts of pro- and anti-policy bundles shown to the advocate.

The preregistration further predicted that evidence-congenial asymmetric updating would be especially strong for claim-congenial claims. This was tested both by estimating the H3 model within claim-congenial claims and by adding the corresponding three-way interactions $Adv \times Pro^{shown} \times CC$ and $Adv \times Con^{shown} \times CC$ in the full sample.

The preregistered H3 evidence-congenial interactions were not supported. Table S8 reports the component and pooled directional tests.

Table S8: **Study 4 preregistered H3 tests: evidence-congenial asymmetric updating.** Entries report coefficient, clustered standard error, and one-sided p-value for the preregistered directional test. The pooled contrast is $\gamma_{Pro} + \gamma_{Con}$.

Outcome	$Adv \times Pro$	$Adv \times Con$	Pooled contrast
Claim warrantedness	−0.080 (0.132), $p = .728$	−0.163 (0.128), $p = .899$	−0.243 (0.221), $p = .864$
Quantitative estimates	−0.217 (0.122), $p = .963$	−0.070 (0.131), $p = .702$	−0.286 (0.224), $p = .899$
Policy support	−0.104 (0.135), $p = .779$	0.007 (0.131), $p = .479$	−0.097 (0.232), $p = .662$

The predicted moderation by claim congeniality was also unsupported. Restricting the analysis to claim-congenial claims, the pooled H3 contrast was −0.034 for claim warrantedness ($SE = 0.320$, $p = .542$) and −0.056 for quantitative estimates ($SE = 0.309$, $p = .572$). In the complementary full-sample specification, the preregistered three-way interactions were 0.138 ($SE = 0.167$, $p = .205$) and 0.133 ($SE = 0.183$, $p = .234$) for pro- and anti-policy

evidence, respectively, in the claim-warrantedness model, and 0.170 ($SE = 0.183,\ p = .177$) and -0.059 ($SE = 0.178,\ p = .630$) in the quantitative model.

The absence of support for H2 and H3 should be interpreted with some caution. These hypotheses are identified from interaction terms involving the advocate's randomized evidence exposure and were estimated considerably less precisely than the overall advocacy effect; the H3 pooled contrasts and claim-congenial moderation tests in particular leave limited sensitivity to modest downstream asymmetries. Limited precision is a more plausible explanation for H2 than for the full H3 pattern, however, because the quantitative H2 pooled contrast was positive and close to the preregistered threshold whereas the H3 point estimates were generally opposite-signed. A further possibility is that randomized evidence exposure affected reviewers only indirectly through the dossier the advocate constructed. Advocates who received relatively little congenial evidence could partly compensate through the content and framing of their written argument, selective use and interpretation of the evidence they did receive, or inferences drawn from that evidence, weakening the mapping from randomized exposure to the information ultimately conveyed to reviewers. Studies 5 and 6 therefore provide more direct tests of asymmetric updating, because randomized evidence exposure and subsequent belief updating are measured within advocates themselves rather than across this additional transmission stage.

S4.7 Preregistered H4: advocate processing and attenuation

For each dossier, the reviewer analysis carried forward the advocate's randomized evidence exposure and the Study 6 process measures, separately for pro-policy and anti-policy evidence: fact-opening count, fact-viewing time, source-opening count, source-viewing time, whether any fact was shared, number of facts shared, whether any fact was subsequently activated, and number of facts activated. The preregistered H4 analyses added the pro-policy and anti-policy versions of these process measures to the H1–H3 reviewer-updating models and tested whether the advocacy and asymmetric-updating coefficients attenuated. The analysis considered the complete process set and the preregistered process families of attention, source consultation, sharing, and activation separately. Cross-model attenuation contrasts were estimated with advocate-dossier-clustered standard errors.

The evidence advocates chose to transmit accounted for a substantial portion of the reviewer advocacy effect. Adding the preregistered sharing measures—whether any pro- or anti-policy fact was shared and the number of pro- and anti-policy facts shared—reduced the claim-warrantedness advocacy coefficient from 0.383 ($SE = 0.070$) to 0.243 ($SE = 0.073$), an attenuation of 36.4%. The cross-model difference was 0.139 ($SE = 0.027,\ p < .001$). For quantitative estimates, the advocacy coefficient fell from 0.207 ($SE = 0.050$) to 0.155 ($SE = 0.052$), an attenuation of 25.2%; the cross-model difference was 0.052 ($SE = 0.019,\ p = .003$). For policy support, the coefficient fell from 0.524 ($SE = 0.117$) to 0.288 ($SE = 0.138$), an attenuation of 45.0%; the cross-model difference was 0.236 ($SE = 0.085,\ p = .003$).

Using the complete set of preregistered process measures produced a similar pattern. The

advocacy coefficient was reduced by 35.0% for claim warrantedness, 22.3% for quantitative estimates, and 47.0% for policy support. The corresponding cross-model attenuation tests were significant for claim warrantedness ($b = 0.134,\ SE = 0.031,\ p < .001$), quantitative estimates ($b = 0.046,\ SE = 0.021,\ p = .016$), and policy support ($b = 0.247,\ SE = 0.098,\ p = .006$).

S4.8 Preregistered robustness analyses

The preregistration specified mixed-effects robustness analyses for the repeated claim-level and quantitative outcomes, using a reviewer random intercept. Policy support contains one observation per reviewer and therefore has no corresponding reviewer-level mixed-effects specification.

The mixed-effects estimates reproduced the primary H1 results closely: the advocacy coefficient was 0.383 for claim warrantedness ($SE = 0.070,\ p < .001$) and 0.206 for quantitative estimates ($SE = 0.050,\ p < .001$). The secondary H2 and H3 conclusions were also unchanged. For H2, the pooled claim-congenial contrast was 0.077 for claim warrantedness ($SE = 0.119,\ p = .261$) and 0.180 for quantitative estimates ($SE = 0.121,\ p = .069$). For H3, the pooled evidence-congenial contrast was -0.140 for claim warrantedness ($SE = 0.206,\ p = .752$) and -0.283 for quantitative estimates ($SE = 0.212,\ p = .909$). The mixed-effects claim-congenial moderator tests were likewise nonsignificant in the preregistered directions.

S5 Study 5: Legal advocates' selective processing and belief updating

S5.1 Participants, preregistration, and exclusions

Participants were recruited through Prolific under the paper-wide eligibility criteria described above. Participants received a $3.00 base payment and no bonus. All participants provided informed consent under Princeton Institutional Review Board Protocol #10857 (Computational Cognitive Science).

The study was preregistered before collection of the confirmatory sample on May 13, 2026, on OSF (`https://doi.org/10.17605/OSF.IO/6HNEF`), with a planned sample of approximately 200 participants.

Data-quality exclusions followed the paper-wide procedures described in Section S8. Briefly, incomplete runs were removed before identifying repeat participation. Among otherwise eligible completed runs, only the earliest completed run for each Prolific participant ID was retained. The raw data contained 208 completed sessions corresponding to 206 unique completed participants. Two additional participants were excluded under the post-preregistration external-assistance/template screen described in Section S8, yielding a final sample of $N = 204$.

The final sample contained 108 participants assigned to advocate for the Plaintiff and 96 assigned to advocate for the Defendant. By case and advocacy assignment, the sample consisted of 34 Plaintiff and 33 Defendant advocates in the consumer-fraud case, 38 Plaintiff and 32 Defendant advocates in the employment-discrimination case, and 36 Plaintiff and 31 Defendant advocates in the toxic-implant case.

S5.2 Design and procedure

Participants evaluated one of the same three legal disputes used in the legal-judge study: the product-liability, employment-discrimination, and consumer-fraud disputes. Case construction and provenance are described in the corresponding Supplementary Information section for the legal-judge study.

Participants first read the case summary and disputed factual propositions. Before reviewing the experimental evidence or receiving their advocacy role, they reported on 0–100 scales how likely they believed each disputed claim was to be accepted and how likely they believed it was that the Plaintiff should prevail.

Participants were then randomly assigned to prepare an argument for either the Plaintiff or the Defendant. They were instructed to prepare a brief dossier on behalf of their assigned side consisting of a written argument and whatever displayed evidence they thought helpful for making the strongest case for that side. The task required a minimum of 80 words and described approximately 100–200 words as the intended length.

Evidence exposure was independently randomized for each disputed claim. The pruned stimulus set contained one Plaintiff-favoring and one Defendant-favoring fact for each claim. For each participant and claim, the interface displayed:

1. the Plaintiff-favoring fact only,
2. the Defendant-favoring fact only, or
3. both facts.

Advocacy assignment and evidence exposure were randomized independently. Consequently, differences between Plaintiff and Defendant advocates in belief updating cannot be attributed to systematic differences in which evidence they were assigned to receive.

During the dossier task, evidence items appeared as cards that participants could open. Participants could insert a numbered evidence reference into their argument, which also attached that evidence to the dossier, or attach an evidence item without inserting an inline reference. Participants were free to use as much or as little of the displayed evidence as they considered appropriate.

Figure 5a of the main text shows the legal dossier-construction interface. Participants could reveal the disputed issue associated with an evidence item, open the evidence, insert a

numbered reference into the argument, or attach the evidence without an inline reference. The interface also displayed the 80-word minimum for the written argument.

After submitting and reviewing their dossier, participants repeated the disputed-claim and overall-verdict judgments. They subsequently completed an activation task indicating which displayed evidence items had come to mind while making their final judgments.

S5.3 Process measures

For every displayed evidence item, the task recorded:

- the number of times the item was opened;
- total viewing time;
- whether the item was referenced in or attached to the submitted dossier; and
- whether the item was selected in the subsequent activation task.

Following the preregistration, opening count and viewing time were treated as separate attention measures. Viewing time was transformed as $\log(1 + \text{time})$. The reference indicator equaled one when an evidence item was incorporated into the submitted dossier through the interface, and activation equaled one when the participant selected the item in the subsequent activation task.

Evidence was defined as *congenial* when it favored the participant's assigned advocacy side.

S5.4 Outcome construction

Let $Y_{ic\ell}^{pre}$ and $Y_{ic\ell}^{post}$ denote participant i's pre- and post-review judgments concerning claim ℓ in case c. Claim judgments were oriented so that larger values indicate greater support for the Plaintiff. Responses were rescaled to the unit interval and bounded to $[.005, .995]$ before logit transformation.

The primary claim-level updating outcome was

$$\Delta \operatorname{logit}_{ic\ell} = \operatorname{logit}(Y_{ic\ell}^{post}) - \operatorname{logit}(Y_{ic\ell}^{pre}). \tag{S43}$$

The overall-verdict outcome was constructed analogously from the participant's reported probability that the Plaintiff should prevail.

Let $Adv_i = 1$ for participants assigned to advocate for the Plaintiff and 0 for Defendant advocates. Let $P_{ic\ell}^{shown}$ and $D_{ic\ell}^{shown}$ indicate exposure to the Plaintiff- and Defendant-favoring evidence items for a claim.

S5.5 H1: advocacy effects

The preregistered claim-level specification was

$$\Delta \operatorname{logit}_{ic\ell} = \alpha + \beta_A Adv_i + \beta_P P_{ic\ell}^{shown} + \beta_D D_{ic\ell}^{shown} + \gamma_\ell + \epsilon_{ic\ell}, \tag{S44}$$

where γ_ℓ are claim fixed effects and standard errors are clustered by participant.

At the verdict level, the corresponding model replaces the claim-level exposure indicators with the number of Plaintiff- and Defendant-favoring evidence items shown across the case and includes case fixed effects. H1 predicted $\beta_A > 0$.

The prediction was strongly supported. At the claim level,

$$\hat{\beta}_A = 0.882, \qquad SE = 0.214, \qquad p < .001. \tag{S45}$$

At the verdict level,

$$\hat{\beta}_A = 1.368, \qquad SE = 0.284, \qquad p < .001. \tag{S46}$$

For descriptive interpretation on the original 0–100 scale, Defendant advocates changed their disputed-claim judgments by an average of -4.96 percentage points, whereas Plaintiff advocates changed by $+9.30$ points. Verdict judgments changed by -8.63 and $+14.50$ points, respectively.

The preregistered secondary mixed-effects analysis of claim updating likewise yielded a positive advocacy effect ($b = 0.784$, SE $= 0.200$, $p < .001$).

S5.6 H2: asymmetric updating

To test whether advocates responded more strongly to evidence congenial to their assigned side, we estimated

$$\begin{aligned}\Delta \operatorname{logit}_{ic\ell} = \alpha &+ \beta_A Adv_i + \beta_P P_{ic\ell}^{shown} + \beta_D D_{ic\ell}^{shown} \\ &+ \gamma_P (Adv_i \times P_{ic\ell}^{shown}) + \gamma_D (Adv_i \times D_{ic\ell}^{shown}) + \gamma_\ell + \epsilon_{ic\ell}.\end{aligned} \tag{S47}$$

As explained in Section S8, the preregistration contained a sign error in writing the pooled directional contrast as $\gamma_P - \gamma_D > 0$. Under the preregistered coding, the verbal asymmetric-updating prediction corresponds to $\gamma_P > 0$, $\gamma_D > 0$, and therefore the pooled contrast $\gamma_P + \gamma_D > 0$.

At the claim level, both component interactions were independently positive:

$$\gamma_P = 1.068, \quad SE = 0.445, \quad p = .008, \tag{S48}$$

and

$$\gamma_D = 1.116, \quad SE = 0.404, \quad p = .003. \tag{S49}$$

The pooled asymmetry was

$$\gamma_P + \gamma_D = 2.184, \quad SE = 0.718, \quad p = .001. \tag{S50}$$

The raw probability-point pattern is shown in the main-text figure. Within each evidence-exposure condition, claim changes are first averaged within participant before computing the plotted mean and confidence interval.

At the verdict level, the pooled interaction was directionally positive but did not reach the preregistered significance threshold:

$$\gamma_P + \gamma_D = 0.707, \quad SE = 0.469, \quad p = .066. \tag{S51}$$

The secondary mixed-effects claim model yielded the same substantive pattern: the Plaintiff-evidence interaction was $b = 1.023$ (SE = 0.431, $p = .009$) and the Defendant-evidence interaction was $b = 0.988$ (SE = 0.433, $p = .011$).

S5.7 Robustness to outcome scale

Studies 3 and 5 preregistered different primary scales for otherwise similar probability-updating outcomes: Study 3 used raw post-minus-pre percentage-point changes, whereas Study 5 used changes in log odds. To place the two legal studies on a common scale, we therefore conducted a non-preregistered robustness analysis that repeated the Study 5 H1 and H2 regressions using raw percentage-point changes, with the same covariates, fixed effects, and participant-clustered standard errors as in the preregistered models.

The raw-change models reproduced the principal results. Advocacy shifted claim judgments by 13.397 percentage points ($SE = 2.746$, $p < .001$) and verdict judgments by 22.343 points ($SE = 3.781$, $p < .001$). The pooled claim-level asymmetric-updating contrast was 27.191 points ($SE = 9.255$, $p = .002$). The pooled verdict-level contrast was also positive and significant on the raw scale (13.532 points, $SE = 6.752$, $p = .023$), whereas it was directionally positive but not significant in the preregistered log-odds specification ($p = .066$). Thus, the advocacy and claim-level asymmetric-updating conclusions do not depend on the outcome scale, while the strength of the verdict-level asymmetry does.

We also repeated the matched attenuation analyses on the raw percentage-point scale used in the main text and Figure 5d. Adding all process measures reduced the claim-level advocacy coefficient from 13.397 to 6.755 points (49.6%; cross-model difference = 6.642, $SE = 1.583$, $p < .001$), the verdict-level advocacy coefficient from 22.343 to 10.420 points (53.4%; difference = 11.923, $SE = 3.752$, $p = .001$), and the claim-level asymmetric-updating contrast from 27.191 to 16.192 points (40.5%; difference = 10.999, $SE = 3.523$, $p = .001$). Controlling for sharing alone reduced these three coefficients by 32.1%, 43.7%, and 20.0%, respectively. Thus, the attenuation pattern is also robust to expressing belief change in raw percentage points.

S5.8 H3: selective processing of congenial evidence

The process-as-outcome analyses were restricted to evidence items displayed to the participant. For each process measure, the implemented evidence-item-level regression included advocacy assignment, evidence side, their interaction, the amount of evidence shown for the relevant claim and case, and claim fixed effects. As explained in Section S8, this specification includes the constituent main effects omitted from the corresponding equation in the preregistration.

Advocacy assignment significantly shifted every preregistered process measure toward congenial evidence:

Table S9: Advocacy-by-evidence-side interactions predicting processing of displayed evidence.

Process outcome	Interaction estimate	SE	p
Opening count	1.912	0.306	$< .001$
log(1 + viewing time)	1.180	0.360	.001
Sharing/incorporation	0.903	0.072	$< .001$
Activation	0.855	0.079	$< .001$

The descriptive differences were large. Across displayed evidence items, congenial evidence was opened 3.21 times on average compared with 2.28 times for uncongenial evidence and was viewed for approximately 31.0 versus 20.2 seconds. Participants shared or incorporated 83.9% of congenial evidence versus 37.8% of uncongenial evidence and activated 82.7% versus 38.9%.

These results support the preregistered prediction that advocacy changes which information participants attend to, deploy in their arguments, and subsequently report having in mind.

S5.9 H4: attenuation after controlling for process measures

The preregistered attenuation analyses added process measures to the belief-updating regressions after aggregating them separately for Plaintiff- and Defendant-favoring evidence. As in Study 6, each baseline model and its process-adjusted counterpart were estimated jointly

using a stacked-equation specification, and the directional cross-model attenuation contrasts used participant-clustered standard errors.

Adding all process measures reduced the estimated claim-level advocacy coefficient from 0.882 to 0.449, an attenuation of 49.1% (cross-model difference = 0.433, $SE = 0.124$, $p < .001$). The verdict-level advocacy coefficient fell from 1.368 to 0.660, an attenuation of 51.8% (difference = 0.709, $SE = 0.278$, $p = .006$). The claim-level asymmetric-updating contrast fell from 2.184 to 1.494, an attenuation of 31.6% (difference = 0.690, $SE = 0.266$, $p = .005$). All reported attenuation p-values are from the preregistered one-sided directional tests.

For completeness, we applied the same matched-specification comparison to the verdict-level asymmetric-updating contrast. Although that contrast was only directionally positive in the preregistered baseline model, it declined from 0.707 to -0.076 after adding all process measures (cross-model difference = 0.783, $SE = 0.318$, one-sided $p = .007$). The verdict-level pattern therefore attenuated as well.

The attenuation varied by process family:

Table S10: Attenuation of legal-advocacy belief effects after adding process measures.

Process controls	Claim advocacy	Claim asymmetry	Verdict advocacy	Verdict asymmetry
Attention only	11.7%	9.6%	31.3%	64.6%
Sharing only	32.7%	15.1%	38.9%	50.0%
Activation only	41.2%	28.9%	57.4%	98.9%
All process measures	49.1%	31.6%	51.8%	110.7%

These analyses establish statistical attenuation rather than causal mediation. In particular, the activation measure was elicited after participants had already provided the post-review judgments, so it cannot constitute a temporally prior mediator of those judgments. Its association is instead informative about which evidence remained cognitively active following the advocacy task.

S5.10 Exploratory selective-processing analyses

We additionally examined whether within-participant selective processing predicted advocacy-congruent belief updating. These analyses were not part of the preregistered primary tests and are treated as exploratory.

Selective opening, reference, and activation each significantly moderated the evidence-congenial asymmetric-updating pattern. The corresponding pooled interaction contrasts were significant for selective opening ($p = .008$), reference ($p = .004$), activation ($p < .001$), and a combined selective-processing index ($p < .001$); the viewing-time interaction was directionally positive but not significant ($p = .067$).

These exploratory results are consistent with the preregistered attenuation analyses but do not establish that selective processing causally mediates the effect of advocacy assignment.

S5.11 Summary of preregistered tests

Test	Result
H1: advocacy shifts claim beliefs	Supported
H1: advocacy shifts verdict beliefs	Supported
H2: asymmetric claim updating	Supported
H2: asymmetric verdict updating	Directionally positive but not significant
H3: selective processing of advocacy-aligned evidence	Supported for all four measures
H4: attenuation of claim advocacy effect	Supported
H4: attenuation of verdict advocacy effect	Supported
H4: attenuation of claim-level asymmetric updating	Supported

Assigned advocacy changed both belief updating and information processing. Participants assigned to opposite sides received evidence generated by the same randomized exposure process, yet updated in opposite directions, responded asymmetrically to the evidence they happened to receive, and selectively attended to and deployed evidence favorable to their assigned side. Controlling for these selective-processing measures substantially reduced the advocacy effect, consistent with selective information processing constituting an important component of advocacy-induced belief change.

S6 Study 6: Policy advocacy

S6.1 Participants, preregistration, and data-quality procedures

Participants were recruited through Prolific under the paper-wide eligibility criteria described above and received a $5.00 base payment. Participants could earn up to $2.00 in performance-based bonuses, including an advocacy bonus determined by the later reviewer's judgment and accuracy bonuses tied to the quantitative-estimate tasks. All participants provided informed consent under Princeton Institutional Review Board Protocol #10857 (Computational Cognitive Science).

The study and analyses were preregistered before collection of the final confirmatory sample on August 12, 2026, on OSF (`https://doi.org/10.17605/OSF.IO/FAHME`). The preregistered recruitment target was $N = 300$ advocates.

Data-quality procedures followed the paper-wide procedures described in Section S8. These procedures were adopted after preregistration because repeat participation and apparent external/template assistance had not been anticipated in the preregistered exclusion rules. Across the pooled final exports, we identified 384 distinct study exposures. Sixty-five exposures

did not reach the second post-advocacy quantitative-question block and were excluded under the preregistered completion criterion. Among the remaining eligible completed exposures, nine later completed exposures were removed because the same Prolific participant ID had an earlier eligible completed exposure. Six additional participants were excluded under the paper-wide external-assistance/template screen. No preview-mode sessions survived to this stage. The final sample contained $N = 304$ participants.

The final sample contained 149 participants evaluating carbon pricing (80 assigned to support and 69 to oppose) and 155 evaluating the minimum wage (75 assigned to support and 80 to oppose).

We conducted a post hoc completion-balance check using a linear-probability model with policy fixed effects and the numbers of pro- and anti-policy evidence bundles randomly assigned to the participant. All p-values in this diagnostic check are two-sided. Of the 384 identified non-preview exposures, 86.0% of those assigned to support and 80.3% of those assigned to oppose reached the preregistered completion point. After accounting for policy and randomized evidence exposure, the support–oppose difference was 0.060 ($SE = 0.038,\ p = .115$). The pro- and anti-policy exposure counts did not jointly predict completion ($p = .277$). Thus, these checks did not provide statistically significant evidence that completion depended on advocacy assignment or randomized evidence exposure, although the estimated assignment difference was imprecise.

S6.2 Policy materials

Participants evaluated one of two policy proposals:

1. **Minimum wage:** "The minimum wage should be increased."
2. **Carbon pricing:** "The United States should adopt a carbon-pricing policy."

Each policy contained five disputed factual claims. Importantly, the claims themselves differed in their implications for the ultimate policy judgment: accepting some claims supported the policy, whereas accepting others opposed it. The complete deployed policy summaries, claims, evidence items, quantitative questions, source information, and benchmark definitions are available in the project repository; the Supplementary Information reports the features needed to understand the design and analyses.

For each claim, the evidence set contained four facts: two supporting the claim and two opposing it. The two facts on each side were shown together as an evidence bundle. Depending on the policy valence of the underlying claim, a claim-supporting bundle could therefore be either pro-policy or anti-policy. For example, when greater acceptance of a claim supported the policy, evidence supporting the claim was pro-policy; when greater acceptance of a claim opposed the policy, evidence supporting the claim was anti-policy.

Each quantitative question concerned a measurable real-world outcome corresponding to one of the disputed claims. Quantitative estimates were incentivized for accuracy relative to externally specified benchmarks. The benchmark value itself was hidden from participants.

S6.3 Procedure

Participants first read a neutral summary of the policy and its five disputed factual claims. Before receiving an advocacy assignment, they completed three sets of baseline judgments:

- warrantedness judgments for each disputed claim;
- quantitative estimates corresponding to measurable outcomes associated with the claims; and
- overall support for the policy.

Participants were then randomly assigned to advocate either in favor of or against the policy proposal. They were told that a later reviewer would see their argument and the facts they chose to include before reporting judgments about the policy and related claims. Participants could earn up to $1.00 based on how strongly the reviewer ultimately judged the policy in the direction of the participant's assigned advocacy position.

For each claim, evidence exposure was independently randomized to one of three conditions:

pro-policy bundle only, anti-policy bundle only, both bundles.

Because each bundle contained two facts, a participant saw either two or four facts for a given claim. Advocacy assignment was independent of evidence exposure.

Participants then wrote an advocacy argument of at least 100 words. The dossier interface displayed the available evidence grouped by claim. Participants could open a fact and its source, attach a fact so that the reviewer would see it with the argument, or insert an inline numerical reference to the fact at a particular point in the argument. Only facts attached or referenced through the interface were included with the submitted argument.

Figure 6a of the main text illustrates the policy dossier-construction interface. The participant's assigned advocacy direction was displayed at the top of the page; claims were grouped by whether accepting them supported or opposed the proposed policy, and each fact was separately labeled according to whether it supported or opposed the claim. Opening a fact revealed its text and allowed participants to view the source name and underlying link before deciding whether to attach the fact or insert a numbered reference to it.

After submitting the dossier, participants repeated the claim-warrantedness, quantitative-estimate, and policy-support judgments. Finally, they completed an activation task indicating

which of the displayed facts had come to mind while answering the post-advocacy questions, followed by demographic questions.

S6.4 Outcome orientation and standardization

All updating outcomes were oriented so that positive values indicate movement in a more policy-supportive direction.

For claim ℓ under policy p, let $V_{p\ell} = 1$ if greater warrantedness of the claim supports the policy and $V_{p\ell} = -1$ if greater warrantedness opposes it. Claim judgments were first reoriented using $V_{p\ell}$.

For quantitative question q, let $S_{p\ell q} = 1$ if a larger numeric response implies greater support for the policy and $S_{p\ell q} = -1$ otherwise. Quantitative judgments were reoriented analogously. Policy-support judgments already had the desired orientation.

For each outcome X, the raw pre-to-post change was

$$\Delta X_{iu}^{raw} = X_{iu}^{post,*} - X_{iu}^{pre,*}.$$

Following the preregistration, change scores were standardized within their relevant policy-specific claim, question, or policy:

$$\Delta X_{iu} = \frac{\Delta X_{iu}^{raw} - \overline{\Delta X_{u}^{raw}}}{\mathrm{sd}(\Delta X_{u}^{raw})}.$$

The resulting outcomes were claim-warrantedness updating ΔW, quantitative-estimate updating ΔB, and policy-support updating $\Delta Support$. When a claim contained multiple quantitative questions, standardized question-level changes were averaged to construct the primary claim-level quantitative outcome.

S6.5 Claim alignment and evidence alignment

The policy design allowed us to distinguish between the alignment of a claim itself and the alignment of the evidence bearing on it.

A claim was *advocacy-aligned*—termed *claim-congenial* in the preregistration—when accepting the claim favored the participant's assigned advocacy position:

$$ClaimAligned_{ip\ell} = \begin{cases} 1 & \text{if accepting claim } \ell \text{ favors participant } i\text{'s assigned position,} \\ 0 & \text{otherwise.} \end{cases}$$

Evidence was *advocacy-aligned*—termed *evidence-congenial* in the preregistration—when the evidence itself favored the participant's assigned position.

The preregistered analyses used these two forms of alignment to distinguish selective processing of claims that could support the assigned position, selective processing and updating from evidence that favored the assigned position, and whether asymmetric responsiveness to advocacy-aligned evidence was especially strong on advocacy-aligned claims. H2 tested the additional possibility that *all* evidence about an advocacy-aligned claim would exert greater influence, regardless of whether that particular evidence helped or hurt the assigned side.

S6.6 H1: advocacy effects

For claim warrantedness, the preregistered model was

$$\Delta W_{ip\ell} = \alpha + \beta_A Adv_i + \beta_{Pro} Pro_{ip\ell}^{shown} + \beta_{Con} Con_{ip\ell}^{shown} + \eta_{p\ell} + \epsilon_{ip\ell},$$

where $Adv_i = 1$ denotes assignment to support the policy and $\eta_{p\ell}$ are policy-specific claim fixed effects. Standard errors were clustered by participant. The analogous quantitative model replaced ΔW with ΔB. The policy-support specification used counts of pro- and anti-policy bundles shown across the five claims and included policy fixed effects.

H1 predicted $\beta_A > 0$. The prediction was supported for all three outcomes:

Outcome	Advocacy coefficient	SE	p
Claim warrantedness	0.4595	0.0735	$< .001$
Quantitative estimates	0.1816	0.0538	$< .001$
Overall policy support	0.8311	0.1053	$< .001$

On the original scales, average oriented claim-warrantedness change was -5.61 percentage points among participants assigned to oppose the policy and +5.90 points among those assigned to support it. Overall policy support changed by -14.28 and +5.46 points, respectively. Because the quantitative questions used heterogeneous units, descriptive quantitative changes are reported in standardized units.

S6.7 H2: greater responsiveness to all evidence about advocacy-aligned claims

H2 tested whether evidence exposure had a larger effect for advocacy-aligned claims, irrespective of whether the particular evidence favored or opposed the assigned position.

For claim warrantedness, the preregistered model added *ClaimAligned* and interactions between claim alignment and pro- and anti-policy evidence exposure. The directional predictions corresponded to stronger evidence effects on advocacy-aligned claims, with the preregistered composite pooling the two directions.

The claim-warrantedness results did not support H2: the composite was 0.0600 (SE = 0.1187, $p = .307$). The analogous policy-support composite was also unsupported (-0.0133, SE = 0.1189, $p = .545$). For quantitative estimates, the composite was positive and reached the preregistered threshold (0.2039, SE = 0.1177, $p = .042$), although neither component interaction was individually significant.

S6.8 H3: asymmetric responsiveness to advocacy-aligned evidence

H3 tested whether participants' beliefs responded more strongly to evidence that favored their assigned advocacy position.

For claim warrantedness,

$$\begin{aligned}\Delta W_{ip\ell} = \alpha + \beta_A Adv_i + \beta_{Pro} Pro_{ip\ell}^{shown} + \beta_{Con} Con_{ip\ell}^{shown} \\ + \gamma_{Pro}(Adv_i \times Pro_{ip\ell}^{shown}) + \gamma_{Con}(Adv_i \times Con_{ip\ell}^{shown}) + \eta_{p\ell} + \epsilon_{ip\ell}.\end{aligned}$$

Because outcomes were coded in policy-support units, both predicted interactions were positive. The preregistered pooled contrast was

$$\gamma_{Pro} + \gamma_{Con} > 0.$$

Across all claims, the pro-policy interaction was 0.1854 (SE = 0.1271, $p = .072$) and the anti-policy interaction was 0.2224 (SE = 0.1152, $p = .027$). Their pooled contrast was 0.4078 (SE = 0.2097, $p = .026$).

S6.8.1 H3: moderation by claim alignment

The preregistration further predicted that asymmetric responsiveness to advocacy-aligned evidence would be especially strong when the underlying claim was itself advocacy-aligned.

Restricting the analysis to advocacy-aligned claims yielded

$$\gamma_{Pro} = 0.2953 \quad (SE = 0.1653,\ p = .037)$$

and

$$\gamma_{Con} = 0.5085 \quad (SE = 0.1543,\ p < .001),$$

with a pooled contrast of

$$\gamma_{Pro} + \gamma_{Con} = 0.8038 \quad (SE = 0.2658,\ p = .001).$$

The complementary full-sample moderator specification directly tested whether the advocacy-by-evidence interactions increased on advocacy-aligned claims. Both preregistered three-way interactions were positive:

$$\delta_{Pro} = 0.2495 \quad (SE = 0.1422,\ p = .040)$$

and

$$\delta_{Con} = 0.3801 \quad (SE = 0.1441,\ p = .004).$$

Thus the evidence-alignment asymmetry was substantially larger when the evidence concerned a claim whose acceptance would itself favor the assigned advocacy position.

The corresponding evidence-alignment asymmetry was not supported for quantitative-estimate updating or overall policy support. Across all claims, the pooled evidence-alignment contrast was 0.2396 for quantitative-estimate updating ($SE = 0.2140,\ p = .132$) and -0.0921 for overall policy support ($SE = 0.2107,\ p = .669$). Restricting the quantitative analysis to advocacy-aligned claims likewise did not support the prediction ($-0.3142,\ SE = 0.3138,\ p = .842$).

S6.9 H4: selective processing of advocacy-aligned evidence

The randomized unit of evidence exposure was the two-fact evidence bundle. For each shown bundle, the preregistration defined eight process outcomes:

1. attention-open count;
2. attention time;
3. source-open count;
4. source-viewing time;
5. whether any fact was shared;
6. number of facts shared;

7. whether any fact was subsequently activated; and

8. number of facts activated.

For each process outcome, the preregistered evidence-bundle regression compared advocacy-aligned with advocacy-opposed evidence while controlling for the number of bundles shown for the claim and policy-specific claim fixed effects. Standard errors were clustered by participant.

H4 was supported for all eight outcomes:

Process outcome	Alignment coefficient	SE	p
Fact openings	0.4746	0.1018	$< .001$
Viewing time	1.5658	0.3959	$< .001$
Source openings	0.0676	0.0310	.015
Source-viewing time	0.6095	0.2452	.007
Any sharing	0.2718	0.0249	$< .001$
Sharing count	0.4418	0.0391	$< .001$
Any activation	0.1860	0.0282	$< .001$
Activation count	0.3460	0.0509	$< .001$

Descriptively, advocacy-aligned versus advocacy-opposed shown bundles were opened 2.15 versus 1.68 times on average. At least one fact was shared from 48.5% versus 21.5% of bundles, and at least one fact was subsequently activated for 69.3% versus 50.9%.

S6.10 H5: selective processing of advocacy-aligned claims

H5 tested whether participants devoted more total processing to claims whose acceptance favored their assigned advocacy position. Process measures were aggregated across all evidence shown for each claim.

All eight preregistered claim-alignment effects were positive and significant:

Process outcome	Claim-alignment coefficient	SE	p
Fact openings	1.4181	0.1548	$< .001$
Viewing time	7.3499	0.6601	$< .001$
Source openings	0.1626	0.0361	$< .001$
Source-viewing time	1.3443	0.2889	$< .001$
Any sharing	0.3440	0.0307	$< .001$
Sharing count	0.5080	0.0462	$< .001$
Any activation	0.1200	0.0324	$< .001$
Activation count	0.2189	0.0555	$< .001$

For illustration, participants opened evidence about advocacy-aligned claims 3.28 times on average compared with 1.80 times for non-aligned claims, and shared at least one fact concerning 59.6% versus 26.9% of the respective claims.

S6.11 Process-to-judgment associations

The preregistered process-to-judgment analyses added pro-policy and anti-policy versions of the process measures to the updating regressions. These specifications ask whether individual differences in what participants actually processed predicted subsequent updating after accounting for advocacy assignment and randomized evidence exposure.

Table S11 reports every process coefficient from the complete-process specifications for all three outcomes. The individual associations were considerably less systematic than the randomized process-as-outcome effects, in part because the measures of attention, source consultation, sharing, and activation are correlated. Five of the 48 directional coefficients were significant at $p < .05$: greater opening and subsequent activation of pro-policy evidence predicted more policy-supportive claim updating; longer viewing of anti-policy evidence and longer consultation of pro-policy sources predicted quantitative updating in the corresponding directions; and sharing any pro-policy information predicted greater policy-support updating. The remaining 43 coefficients were not significant in their preregistered directions. The corresponding process-family specifications are included in the archived analysis output.

Table S11: Study 6 preregistered process-to-judgment associations from the complete-process specifications. Each cell reports the coefficient, clustered standard error in parentheses, and one-sided p-value for the preregistered directional test. Pro-policy coefficients were predicted to be positive; anti-policy coefficients were predicted to be negative. All models also include advocacy assignment, randomized pro- and anti-policy evidence exposure, and the corresponding fixed effects described in the text.

Process coefficient	Claim warrantedness	Quantitative estimates	Policy support
Pro-policy fact openings	0.051 (0.020), .005	−0.016 (0.018), .817	0.018 (0.015), .108
Anti-policy fact openings	−0.024 (0.024), .154	0.019 (0.017), .869	0.000 (0.015), .507
Pro-policy viewing time	0.002 (0.005), .346	0.008 (0.005), .056	−0.007 (0.004), .954
Anti-policy viewing time	0.002 (0.006), .648	−0.009 (0.005), .041	0.004 (0.004), .792
Pro-policy source openings	0.027 (0.086), .378	−0.023 (0.067), .634	0.020 (0.109), .428
Anti-policy source openings	0.000 (0.093), .501	0.003 (0.055), .524	−0.121 (0.106), .126
Pro-policy source-viewing time	−0.009 (0.011), .793	0.017 (0.009), .038	−0.007 (0.014), .689
Anti-policy source-viewing time	0.012 (0.013), .823	0.000 (0.010), .514	0.028 (0.014), .974
Any pro-policy sharing	0.135 (0.179), .225	0.215 (0.152), .079	0.235 (0.111), .017
Any anti-policy sharing	−0.148 (0.194), .223	0.034 (0.180), .574	−0.226 (0.141), .055
Pro-policy sharing count	0.011 (0.130), .465	−0.168 (0.095), .960	−0.103 (0.073), .921
Anti-policy sharing count	0.042 (0.126), .629	−0.088 (0.103), .196	0.096 (0.087), .864
Any pro-policy activation	0.296 (0.152), .026	−0.074 (0.146), .693	0.137 (0.123), .134
Any anti-policy activation	−0.061 (0.182), .370	−0.240 (0.157), .063	0.010 (0.112), .537
Pro-policy activation count	0.010 (0.089), .457	0.078 (0.090), .192	−0.032 (0.064), .693
Anti-policy activation count	−0.127 (0.112), .129	0.173 (0.094), .967	−0.081 (0.066), .111

S6.12 H6: attenuation after controlling for process measures

H6 compared each baseline updating model with a matched model adding separate pro-policy and anti-policy process measures. The preregistered cross-model attenuation contrasts were estimated using participant-clustered standard errors.

Using the complete set of preregistered process measures, the advocacy effect on claim warrantedness declined from 0.4595 to 0.3384, an attenuation of 26.3% ($p < .001$ for the directional attenuation contrast). The policy-support coefficient declined from 0.8311 to 0.6571, an attenuation of 20.9% ($p = .010$). The quantitative-estimate coefficient declined by 11.5%, but this attenuation was not significant ($p = .197$).

To match the theoretically focal estimand in Figure 6d, we first applied the same cross-model comparison to the pooled asymmetric-updating contrast among advocacy-aligned claims. The contrast declined from 0.8038 without process controls to 0.6685 after controlling for sharing and 0.3709 after controlling for all process measures. The full adjustment produced a 53.9% attenuation (matched cross-model difference = 0.4329, $SE = 0.1437$, one-sided $p = .001$); sharing alone produced a 16.8% attenuation that was not significant (difference = 0.1353, $SE = 0.1116$, one-sided $p = .113$).

As secondary analyses across all claims, we also examined the two evidence-alignment interactions separately. The pro-policy interaction declined from 0.1854 to 0.0448, a 75.8% attenuation ($p < .001$); the anti-policy interaction declined from 0.2224 to 0.0763, a 65.7% attenuation ($p = .002$).

The preregistration also specified attenuation analyses by process family. Table S12 reports these analyses for all three advocacy effects and both claim-level evidence-alignment interactions. Sharing measures are especially informative temporally because they were recorded during dossier construction before the post-advocacy judgments.

Table S12: Study 6 attenuation by process family. Entries are percentage reductions in the indicated baseline coefficient after adding the corresponding process controls. The pro- and anti-evidence columns report the separate interaction components estimated across all claims; Figure 6d instead displays their pooled contrast among advocacy-aligned claims. Negative values indicate that the coefficient increased slightly rather than attenuated.

Process controls	Claim advocacy	Quant. advocacy	Policy advocacy	Pro-evidence interaction	Anti-evidence interaction
All process measures	26.3	11.5	20.9	75.8	65.7
Attention only	5.4	5.6	0.6	21.9	26.4
Source consultation only	-0.8	2.6	-5.8	5.2	1.3
Sharing only	19.3	18.4	8.3	55.6	24.0
Activation only	18.9	5.8	16.9	49.2	49.3

These analyses establish statistical attenuation rather than causal mediation. The process measures are post-treatment variables, and activation was elicited after participants had already supplied the post-advocacy judgments. The attenuation results therefore show that the advocacy and asymmetric-updating effects are closely associated with the selective

processing induced by advocacy, but do not by themselves identify a causal mediating pathway.

S6.13 Summary of preregistered tests

Test	Result
H1: advocacy shifts claim beliefs	Supported
H1: advocacy shifts quantitative estimates	Supported
H1: advocacy shifts policy support	Supported
H2: all evidence exerts greater influence on advocacy-aligned claims	Generally unsupported for claim beliefs and policy support; quantitative composite supported weakly
H3: stronger responsiveness to advocacy-aligned evidence	Supported for claim beliefs
H3 moderator: evidence-alignment asymmetry is stronger on advocacy-aligned claims	Supported for claim beliefs
H3: quantitative / policy-support evidence asymmetry	Not supported
H4: selective processing of advocacy-aligned evidence	Supported for all eight measures
H5: selective processing of advocacy-aligned claims	Supported for all eight measures
H6: attenuation of claim advocacy effect	Supported
H6: attenuation of policy-support advocacy effect	Supported
H6: attenuation of quantitative advocacy effect	Not significant
H6: attenuation of pooled asymmetric updating on advocacy-aligned claims	Supported with all process measures; sharing-only reduction not significant
H6: attenuation of separate evidence-alignment interactions across all claims	Strongly supported

Overall, the policy study distinguishes selective processing of potentially useful claims from selective processing of favorable evidence and shows that these forms of alignment interact in belief updating. Advocates devoted more processing to claims whose acceptance could support their assigned position and to evidence that directly favored that position. The strongest asymmetric-updating effect arose when both were true: advocacy-aligned evidence had especially large effects when it bore on an advocacy-aligned claim.

S7 Supplementary experimental materials

This section contains participant-facing materials needed to understand the manipulations and measures without running the experimental code. The complete runnable tasks and machine-readable stimulus files will be archived with the repository materials.

S7.1 Stimulus inventory

Table S13: Inventory of the final legal and policy stimulus sets.

Stimulus set	Domains	Claims	Facts	Quantitative questions
Legal (Studies 3 and 5)	3 cases	11	22	—
Policy (Studies 4 and 6)	2 policies	10	40	13

The final legal stimulus file contains three anonymized disputes—product liability, employment discrimination, and consumer fraud—with 11 disputed factual claims in total. The pruned evidence set contains exactly one Plaintiff-favoring and one Defendant-favoring fact for each claim.

The final policy stimulus file contains two policies—minimum wage and carbon pricing—with five disputed claims per policy. Each claim has four sourced facts, two supporting and two opposing the claim, for 40 facts in total. The minimum-wage module contains seven quantitative questions and the carbon-pricing module six. Each final policy fact is paired with a source name and source URL, and the final stimulus file records that source fit was checked for every fact. We therefore describe the final materials as researcher-checked against their underlying sources; we do not characterize the final stimulus set as independently participant-validated.

S7.2 Interface documentation

Representative participant-facing interfaces appear in the main-text figures. The Supplementary Information adds the Study 1 Random-condition screen and the Study 4 disclosure screen because these document methodologically relevant information not already visible in the corresponding main-text panels. Screenshots are illustrative rather than exhaustive; the complete runnable tasks will be archived with the repository materials.

S7.3 Archiving of complete experimental materials

The Supplementary Information provides representative examples and interface screenshots needed to understand the manipulations and measures. The complete deployed participant-facing materials, machine-readable stimulus files, and runnable tasks for all six studies will

be archived in the project repository rather than reproduced in full in this PDF. For Studies 3 and 5, the repository will include the final case summaries, disputed claims, and evidence items. For Studies 4 and 6, it will include the complete policy summaries, claims, sourced facts, quantitative-question wording, response units and allowable ranges, benchmark definitions, and source metadata.

Representative policy fact and source display. For example, one minimum-wage fact stated, "A study of fast-food restaurants in New Jersey and Pennsylvania found little evidence that New Jersey's minimum-wage increase reduced employment." Its source was displayed on demand as "Card & Krueger" together with a link to the original paper. An opposing fact on the same substantive issue stated, "A major review of minimum-wage studies concludes that the evidence generally points to job losses for low-skilled workers after minimum-wage increases," with the source identified as the Neumark & Wascher review.

Representative quantitative item. One minimum-wage item asked, "Suppose the minimum wage were increased by 10%. Approximately what percentage would restaurant prices increase over the following several years?" The allowable response range was 0–20%, and the scoring benchmark was 0.36%, based on the reported restaurant-price pass-through for a 10% minimum-wage increase in the cited research. In the final reviewer task, the benchmark and source used for scoring were not shown to reviewers; they saw the question, units, allowable range, and accuracy incentive.

S8 Deviations and post-preregistration data-quality procedures

Several data-quality issues arose during data collection that were not anticipated when the studies were preregistered and therefore were not covered by the original exclusion criteria. In particular, some participants entered the same study more than once, and a small number of open-ended advocacy responses exhibited objective indications of external assistance or pasted/template-generated text. We therefore adopted additional cross-study data-quality procedures for repeated participation and, where applicable, external assistance. These procedures were based on participant identity, response text, and behavioral/process data rather than on the direction or magnitude of participants' belief-updating outcomes.

We distinguish these post-preregistration data-quality procedures from corrections to statistical specifications. Data-quality procedures address unanticipated features of the collected data; specification corrections address cases in which the written preregistered equation did not algebraically implement the accompanying verbal hypothesis or omitted terms required for the intended interaction specification. Both types of deviation are described below.

S8.1 Incomplete runs and repeated participation

We first applied each study's completion criterion. Runs that did not reach the outcome measures required by the study's preregistered inclusion rule were excluded before repeated participation was assessed. We then identified repeated participation using Prolific participant IDs. When the same participant had more than one otherwise eligible completed run of a study, we retained the participant's earliest completed run and excluded later completed runs.

This ordering is deliberate. An incomplete initial attempt does not by itself prevent a participant from contributing a later completed run, whereas allowing multiple completed runs from the same participant could allow one person to contribute observations under more than one randomized treatment assignment.

Repeated completed participation was not anticipated when the relevant studies were preregistered, so this rule was adopted post-preregistration. It was applied without reference to participants' treatment assignments or belief-updating outcomes.

S8.2 Preview and researcher-test sessions

Where study exports contained sessions explicitly marked as preview or researcher-test sessions, these sessions were excluded from confirmatory participant analyses. Preview status is a technical property of the session rather than a participant response and was not used as an outcome-dependent exclusion criterion. No preview-mode sessions entered the final legal-judge, legal-advocate, or policy-advocacy samples.

S8.3 External-assistance/template screen

The preregistrations for the advocacy studies did not anticipate that participants might use large language models, external writing assistance, or pasted argument templates while completing the open-ended advocacy task. We therefore conducted a post-preregistration screen for objective indications that a submitted argument was not composed in the intended manner during the study.

Because the source of externally generated text generally cannot be determined from the submitted response alone, we refer to this procedure as an *external-assistance/template screen* rather than an "LLM detector." Participants were not excluded merely because their prose appeared polished, formulaic, unusually articulate, or potentially AI-generated. Instead, exclusion required stronger observable evidence, such as explicit prompt or instruction leakage, unfilled template placeholders, or an extreme combination of argument-production speed and process data inconsistent with ordinary completion of the evidence-review and writing task.

The screen did not use participants' advocacy assignments, pre-to-post belief changes, or

other substantive outcomes as exclusion criteria.

In the legal-advocate study, two participants were excluded under this procedure. One completed the dossier stage in approximately 65 seconds, opened no evidence items, and submitted a polished generic legal template containing literal unfilled placeholders such as "[key fact from evidence]" and "[explanation]." A second participant completed the combined evidence-review and argument-composition page in approximately 108 seconds, making this the second-fastest completion among 206 otherwise eligible unique participants; the participant produced a 97-word argument, opened only one evidence item for approximately 2.5 seconds, and manually included an evidence-style citation to an item that had not been opened. The next-fastest participant required approximately 181 seconds. We therefore excluded both participants based on the conjunction of extreme timing and process/text evidence, rather than timing alone.

In the policy-advocacy study, six participants were excluded under the corresponding external-assistance/template screen. Three submitted arguments contained explicit prompt or meta-instruction leakage indicative of externally generated text. Three additional participants met the fixed broad screen used in the final analysis based on unusually rapid production of polished argument text together with timing or paste patterns inconsistent with ordinary in-task composition. As in the legal-advocacy study, these exclusions were determined without reference to advocacy assignment or belief-updating outcomes.

S8.4 Cross-study exclusion accounting

Table S14 summarizes the finalized accounting for repeated participation and post-preregistration data-quality exclusions. Study 1 contained 50 complete participants, none of whom met the preregistered comprehension-exclusion rule; no Prolific ID appeared in more than one session. In Study 2, 119 completed sessions were reduced to (N=116) participants after three later repeat completions were removed. In Study 4, 291 sessions met the preregistered reviewer completion criterion; three were later completed participations by participants who had already contributed an earlier eligible completed run, yielding (N=288) reviewers.

Table S14: Post-preregistration data-quality exclusions in finalized studies. "Eligible runs" refers to sessions remaining after the study-specific completion requirement and before repeat-participation or external-assistance exclusions.

Study	Eligible runs	Repeats removed	Assistance exclusions	Final N
Stick censoring	50	0	N/A	50
Threshold censoring	119	3	N/A	116
Legal judges	402	3	N/A	399
Policy reviewers	291	3	N/A	288
Legal advocates	208	2	2	204
Policy advocacy	319	9	6	304

For the policy-advocacy study, the 319 eligible completed runs in Table S14 were obtained

from 384 distinct study exposures after excluding 65 exposures that did not reach the second post-advocacy quantitative-question block. Nine later completed exposures were then removed because the same Prolific participant ID had an earlier eligible completed exposure, and six additional participants were excluded under the external-assistance/template screen, yielding $N = 304$.

For the legal-judge study, 402 completed sessions yielded 399 unique participants after three later completed repeat sessions were removed. For the legal-advocate study, 208 completed sessions yielded 206 unique participants after two later completed repeat sessions were removed; the two external-assistance/template exclusions described above then yielded $N = 204$.

For the policy-reviewer study, the raw exports contained 323 distinct study sessions. Thirty-two did not reach the preregistered primary completion criterion, leaving 291 eligible completed sessions. Three of those were later completed participations by participants who had already contributed an earlier eligible completed run; excluding the later run in each case yielded $N = 288$ reviewers.

S8.5 Legal-advocate study: correction to the asymmetric-updating contrast

The legal-advocate preregistration correctly stated the substantive hypothesis that participants would respond more strongly to evidence congenial to their assigned advocacy side. However, the written pooled contrast contained a sign error.

The preregistered interaction model was

$$\Delta Y = \ldots + \gamma_P(Adv \times P^{shown}) + \gamma_D(Adv \times D^{shown}) + \epsilon, \tag{S52}$$

where $Adv = 1$ denotes Plaintiff advocacy and larger values of the outcome indicate movement toward the Plaintiff. Under this coding, the stated substantive prediction implies

$$\gamma_P > 0 \tag{S53}$$

because Plaintiff advocates should respond more strongly than Defendant advocates to Plaintiff-favoring evidence, and

$$\gamma_D > 0 \tag{S54}$$

because Defendant advocates should respond more negatively than Plaintiff advocates to Defendant-favoring evidence. The corresponding pooled directional contrast is therefore

$$\gamma_P + \gamma_D > 0, \tag{S55}$$

rather than $\gamma_P - \gamma_D > 0$ as written in the preregistration.

The implemented analysis corrects this algebraic sign error. The correction does not alter the verbal hypothesis or its directional implications; it aligns the statistical contrast with the hypothesis and variable coding specified in the preregistration.

S8.6 Legal-advocate study: constituent terms in the process interaction

The preregistered process-as-outcome equation included the interaction between advocacy assignment and evidence side but omitted the corresponding constituent main effects. The study did not include participant fixed effects: advocacy assignment was between participants, and claim fixed effects do not absorb either advocacy assignment or evidence side. The implemented regression therefore includes both constituent main effects together with their interaction, as required for the intended difference-in-differences interpretation.

This correction leaves unchanged the preregistered substantive question: whether assignment to advocate for the Plaintiff rather than the Defendant differentially changes processing of Plaintiff- versus Defendant-favoring evidence.

S8.7 Sensitivity analyses and reporting principle

Because the repeat-participation and external-assistance procedures were adopted after preregistration, we conducted sensitivity analyses that relaxed the relevant post-preregistration data-quality rules. Table S15 reports compact checks for the principal estimands. Study 1 had no post-preregistration participant exclusions. For Studies 2–4, we restored the later repeat runs or sessions. For Studies 5 and 6, where the more judgment-dependent decision concerned the external-assistance/template screen, we re-estimated the principal models after restoring those participants while continuing to retain only one completed run per person.

The substantive conclusions are unchanged in every sensitivity analysis. In Study 2, retaining the three later repeat runs leaves the pooled adjustment coefficient well above zero and well below one. In Study 3, the two preregistered incomplete-correction contrasts remain large for both claim and verdict updating. In Study 4, the three advocacy effects are essentially unchanged. In Study 5, restoring the two external-assistance/template cases leaves both the advocacy effect and claim-level asymmetric-updating contrast nearly unchanged. In Study 6, restoring all six external-assistance/template cases likewise leaves the three H1 advocacy effects and the primary claim-level asymmetry result substantively unchanged.

Throughout the Supplementary Information, we report post-preregistration data-quality exclusions separately from corrections to statistical specifications. This distinction is intended to make clear which deviations concern participant eligibility and which concern the algebraic implementation of preregistered hypotheses.

Table S15: Sensitivity of principal estimates to post-preregistration data-quality decisions. Values in the final-analysis and sensitivity columns are the key preregistered estimands indicated in the second column.

Study	Sensitivity check / estimand	Final analysis	Sensitivity analysis
2: Threshold	Restore three later repeat runs; pooled benchmark-gap coefficient λ	$N = 116$: .369	119 runs: .361
3: Legal judges	Restore three later repeat sessions; incomplete-correction contrasts: claim $T2 - T1, T1 - T3$; verdict $T2 - T1, T1 - T3$	$22.69, 15.26$; $25.03, 14.97$	$22.27, 15.42$; $24.47, 15.09$
4: Policy reviewers	Restore three later repeat sessions; H1 effects on claim warrantedness, quantitative estimates, policy support	$.383, .207, .524$	$N = 291$: $.385, .209, .526$
5: Legal advocates	Restore two assistance-screen cases while retaining one completed run per person; claim advocacy, claim asymmetry, verdict advocacy	$N = 204$: $.882, 2.184, 1.368$	$N = 206$: $.886, 2.157, 1.323$
6: Policy advocates	Restore six assistance-screen cases while retaining one completed run per person; H1 claim, quantitative, policy effects; asymmetry on aligned claims	$N = 304$: $.460, .182, .831$; .804	$N = 310$: $.468, .179, .836$; .774