AUDITING ACCURACY CHANGE IN HUMAN–AI DECISION-MAKING: PARTIAL IDENTIFICATION AND BEHAVIORAL CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Research on human–AI decision-making aims to help people make more accurate decisions, but how much improvement after AI advice can existing studies establish? The data reported in these studies may include AI recommendations and final human decisions but may not determine this change when people's independent judgments are missing. We thus provide an analytical lens for evaluating how information and assumptions about human behavior constrain the possible accuracy changes. When people either retain their initial judgment or adopt AI advice, these records leave a range of possible accuracy changes whose width equals the final human–AI agreement rate and cannot be narrowed by collecting more data of the same kind. We examine two ways to narrow these bounds: sparse measurement of initial judgments and behavioral restrictions based on patterns or models. Evaluations on six human-AI decision tasks show that, behavioral patterns narrow the bounds, while the additional contribution of behavioral models varies across tasks. The relative precision of direct measurement and behavioral constraints depends on the task and measurement budget. Using this lens, we audit 94 claim–workflow instances from 56 studies and find sufficient evidence for the conclusions in 80 instances; 14 require additional information or assumptions for stronger interpretations. We provide implications for evaluation and interpretation in human–AI decision-making.

1 INTRODUCTION

AI advice can help people correct mistaken judgments, but it can also lead them to abandon correct ones (Schemmer et al., 2023b). Understanding these changes is central to the growing body of research on human–AI decision-making (Raees et al., 2026; Guo et al., 2026; Wilder et al., 2020): researchers need to know how people's judgments develop when they receive advice, alongside how accurately they decide in the end. Prior studies assess the net change in decision quality resulting from these corrections and errors by comparing the accuracy of people's pre-advice judgments with that of their final decisions (Vodrahalli et al., 2022a; Schemmer et al., 2023b). This target complements evaluations of final performance (Bansal et al., 2021; Kerrigan et al., 2021; Cresswell et al., 2024) and randomized comparisons of offering assistance (Ben-Michael et al., 2025), which answer different questions about human–AI systems.

However, evaluating this change requires information that many decision workflows do not record. In practice, researchers may observe the correct answer, the AI recommendation, and the person's final decision without observing their initial judgment (Tejeda et al., 2022; Fogliato et al., 2022). Consider a final decision that agrees with incorrect AI advice. The person may have held the same incorrect answer before seeing the advice, or may have changed a correct answer to match it. These histories produce identical records but imply different accuracy changes. Agreement with correct advice is similarly ambiguous: it may reflect an initially correct judgment or a correction after advice (Schemmer et al., 2023b). Even when final performance is measured precisely, the change from initial to final accuracy can therefore remain uncertain.

Missing initial judgments do not necessarily leave researchers without further evidence. A study may record initial answers on some trials, measure an aggregate tendency to adopt advice (Zhang

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

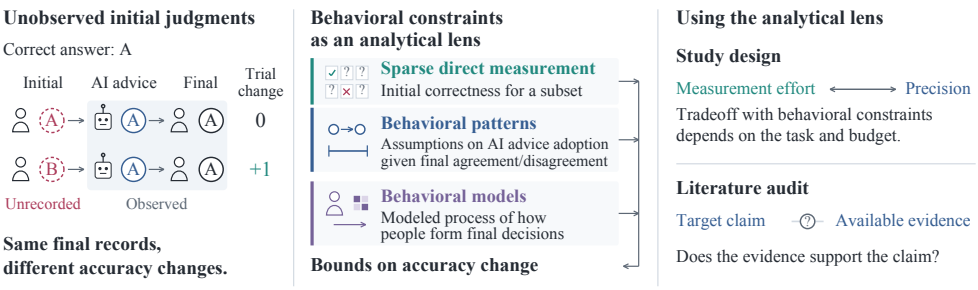


Figure 1: **Behavioral constraints as an analytical lens.** Final-only data can leave accuracy change uncertain. Sparse direct measurement, behavioral patterns, and behavioral models constrain the possible accuracy changes. This lens informs study design and the literature audit.

et al., 2020; Guo et al., 2024), or use a behavioral model to infer how people arrived at their final decisions (Tejeda et al., 2022). Each source can restrict the initial judgments compatible with the observed outcomes, but the strength of the resulting conclusion depends on what those restrictions establish. To assess such evidence, we need to determine what final records already imply, how much additional behavioral information resolves, and what further measurements would support a more precise conclusion. This motivates our central question: how does what we know or assume about people determine what we can conclude about their accuracy change?

We address this question by characterizing the full range of accuracy changes compatible with observed records and specified behavioral constraints (Manski, 1989). To establish what final records alone reveal, we first consider decisions in which people either retain their initial answer or adopt the AI recommendation: disagreement then reveals the initial answer, while agreement leaves it unresolved. We show that the sharp range of possible accuracy changes has width exactly equal to the final human–AI agreement rate, an ambiguity that more final-only records cannot eliminate. We then examine two ways to narrow this range. One is to measure initial judgments directly on a sparse subset of trials. The other is to restrict the possible initial judgments using behavioral patterns, including adoption rates and the human ability gap between agreement and disagreement cases, or explicit behavioral models (Tejeda et al., 2022; Vodrahalli et al., 2022b) of how people form judgments and respond to advice.

We empirically evaluate these two approaches across six tasks from two public studies containing paired initial and final judgments (Vodrahalli et al., 2022a; Tejeda et al., 2022). Behavioral constraints narrow these ranges, but the additional contribution of fitted mechanisms varies across tasks. By varying how many participants have their initial judgments measured, we examine the tradeoff between sparse direct measurement and behavioral constraints: which yields tighter bounds depends on the task and measurement sparsity. These comparisons evaluate how much uncertainty each source removes while checking whether the resulting ranges contain the revealed gain.

Finally, using this analytical lens, we audit published studies in human–AI decision-making to determine whether their recorded data, study designs, and stated behavioral assumptions suffice to establish their conclusions. Across 94 in-scope claim–workflow instances from 56 studies (Raees et al., 2026), the available evidence supports the conclusions in 80 instances; the remaining 14 require additional information or assumptions for their stronger interpretations. Our contributions are:

- We quantify the uncertainty in accuracy change left by unrecorded initial judgments, showing that under the stated behavioral constraint its range has width equal to final human–AI agreement and cannot be narrowed by collecting more final-only records.
- We provide an analytical lens for evaluating human–AI decision-making by characterizing how sparse direct measurement and behavioral restrictions based on patterns or models restrict the possible accuracy changes.
- We apply this analytical lens to evaluate these information sources across six tasks from two public studies and audit 94 in-scope claim–workflow instances from 56 studies. We provide implications for research in human–AI decision-making.

2 ACCURACY-CHANGE BOUNDS FROM FINAL DECISIONS

Consider a scenario of *human–AI decision-making* (Schemmer et al., 2023b; Vodrahalli et al., 2022a; Tejeda et al., 2022) in which, given a task input $x \in \mathbf{X}$ with ground-truth label $y \in \mathbf{D}$, a human receives AI assistance in making a decision $\hat{y}$. Here, $\mathbf{D}$ is a finite decision space containing at least two labels. This decision process is conceptualized as three stages: the human first forms a pre-advice judgment $\hat{y}_{\mathrm{h}}$, then reviews the AI recommendation $\hat{y}_{\mathrm{a}}$, and finally makes the decision $\hat{y}$ by either accepting the AI recommendation or retaining the pre-advice judgment (i.e., $\Pr(\hat{y} \in \{\hat{y}_{\mathrm{h}}, \hat{y}_{\mathrm{a}}\}) = 1$) (Soll & Larrick, 2009; Tejeda et al., 2022).

Specifically, the effectiveness of human–AI decision-making is measured by the extent to which AI helps people make better decisions, that is, by the accuracy change between their initial judgments and final decisions:

$$\Delta := \Pr(\hat{y} = y) - \Pr(\hat{y}_{\mathrm{h}} = y). \tag{1}$$

In practice, however, this accuracy change is not observable: although people form an initial judgment internally, it is usually not explicitly recorded, and eliciting it is costly (Fogliato et al., 2022). Evaluations of AI-based systems usually record only the AI recommendation $\hat{y}_{\mathrm{a}}$ and the final decision $\hat{y}$, with ground truth y available for evaluation, giving $o = (y, \hat{y}_{\mathrm{a}}, \hat{y})$, the *final-only data*. This leads to the evaluation question: Can these records tell us whether AI advice improved human accuracy?

In some cases they cannot, as we observe only whether people's final decisions agree with the AI recommendation. Depending on whether $\hat{y}_{\mathrm{a}}$ is correct, we can further categorize agreement and have probabilities of correct agreement, incorrect agreement, and any agreement:

$$p_+ := \Pr(\hat{y} = \hat{y}_{\mathrm{a}} = y), \quad p_- := \Pr(\hat{y} = \hat{y}_{\mathrm{a}}, \hat{y}_{\mathrm{a}} \neq y), \quad g := \Pr(\hat{y} = \hat{y}_{\mathrm{a}}) = p_+ + p_-.$$

We assume $g < 1$ throughout the main text. However, we do not observe how these agreements arise, as the records do not reveal the process through which people rely on the AI recommendation. Intuitively, $\hat{y} \neq \hat{y}_{\mathrm{a}}$ implies that $\hat{y}_{\mathrm{h}} = \hat{y}$. When $\hat{y} = \hat{y}_{\mathrm{a}}$, however, the record cannot distinguish between pre-advice agreement (i.e., $\hat{y}_{\mathrm{h}} = \hat{y}_{\mathrm{a}}$) and reliance (i.e., $\hat{y}_{\mathrm{h}} \neq \hat{y}_{\mathrm{a}}$). For example, an incorrect agreement may reflect either (a) an already incorrect pre-advice judgment, with no change in accuracy, or (b) a correct judgment revised to match the AI recommendation, with a decrease in accuracy. Both cases produce the same decisions. Therefore, from final-only data, we can identify only a set of possible accuracy changes, the *identified set* $\mathbf{I}_\Delta$.

Proposition 1. *The sharp identified set is* $\mathbf{I}_\Delta = [-p_-, p_+]$*, with width* g*. For every sample size, any estimator of* Δ *based on these records has worst-case expected absolute error at least* $g/2$*.*

In particular, collecting more final decisions cannot eliminate this uncertainty (proof in Appendix D). As the uncertainty arises because the records do not reveal which reliance process produced the observed agreement, Section 3 therefore examines how additional information or assumptions about human behavior can rule out some otherwise compatible processes and progressively narrow the bounds on accuracy change.

3 NARROWING THE BOUNDS WITH BEHAVIORAL INFORMATION

Section 2 shows that uncertainty about accuracy change comes entirely from trials where the final decision agrees with the AI recommendation. On these trials, final-only records do not reveal whether the human already agreed with the AI before receiving advice or instead changed their judgment after seeing the recommendation. As a result, evaluating the accuracy change requires additional information about the human's pre-advice judgment.

We consider two ways to obtain such information. One is to measure pre-advice correctness directly, possibly on only a subset of decisions. The other is to use information about how people respond to AI advice to restrict the pre-advice judgments that are consistent with the observed final decisions. We first study direct measurement of pre-advice correctness, then consider behavioral constraints based on switching and differences in pre-advice accuracy across final-agreement and final-disagreement trials, and finally examine behavioral models of how people form judgments and respond to AI advice.

To see how these sources of information affect the accuracy change, consider the trials on which the final decision agrees with the AI recommendation. For $g > 0$, let

$$q_{\text{agree}} := \Pr(\hat{y}_{\text{h}} = y \mid \hat{y} = \hat{y}_{\text{a}})$$

denote pre-advice accuracy on these trials. Since accuracy is unchanged on final-disagreement trials, the overall accuracy change can be written as

$$\Delta = p_{+} - g q_{\text{agree}}. \tag{2}$$

The final records determine p_{+} and g, so the remaining uncertainty comes from q_{agree}.The sections below first examine direct measurement and then behavioral restrictions, ranging from behavioral constraints to explicit behavioral models.

3.1 Direct Measurement of Pre-Advice Correctness

The most direct way to determine accuracy change is to observe pre-advice correctness. If pre-advice accuracy in the target workflow is known, Eq. 1 identifies Δ as final accuracy minus pre-advice accuracy. More generally, bounds on pre-advice accuracy give corresponding bounds on Δ; Appendix Theorem 2 gives the sharp result.

Full measurement is not required to learn about average accuracy change. Suppose we record whether the pre-advice judgment was correct on a subset of $n \geq 2$ trials. Let m trials be selected uniformly without replacement, with measurement fraction $\rho = m/n$, and let $r_i = 1$ indicate that trial i is measured. Write Δ_n for the average accuracy change over the n trials and g_n for their agreement fraction. Recording pre-advice correctness determines the contribution of each measured trial to Δ_n, while final disagreement already reveals zero accuracy change. The sharp range over the unmeasured judgments therefore has width

$$w(r) = \frac{1}{n}\sum_{i=1}^{n}(1 - r_i)\mathbf{1}\{\hat{y}_i = \hat{y}_{\text{a},i}\}, \qquad \mathbb{E}_r[w(r)] = (1-\rho)g_n. \tag{3}$$

Only measuring an agreement trial reduces this width. With no measurements, the width is g_n; measuring every agreement trial determines Δ_n, as does full measurement at $\rho = 1$. Appendix B.3 gives the endpoints and proves sharpness.

Random measurement also supports statistical inference. For $m \geq 1$, subtracting mean measured pre-advice correctness from the fully observed final accuracy gives an estimator of Δ_n with variance $(1-\rho)s_b^2/m$, where s_b^2 is the finite-sample variance of pre-advice correctness. At a fixed confidence level, the resulting confidence interval has width bounded by $\mathrm{O}(\sqrt{(1-\rho)/(n\rho)})$ and can be combined with the bounds above (Appendix B.3). For the population target Δ, independent random measurement permits consistent estimation when $n\rho \to \infty$, even if $\rho \to 0$ (Appendix Corollary 4).

Measurements collected under a different workflow require an additional link to the target workflow. In particular, eliciting a pre-advice judgment may change subsequent decision making (Fogliato et al., 2022; Himmelstein & Budescu, 2026). Appendix B.3 states the corresponding condition, and Corollary 3 allows the pre-advice accuracies of the two workflows to differ within a bounded range.

3.2 Behavioral Constraints

When pre-advice correctness is not measured directly on the target decisions, information about advice-taking behavior can still narrow the possible accuracy change. We first consider how often people switch to the AI recommendation, and then add information about how pre-advice accuracy differs between trials that end in agreement and disagreement.

Switch fraction. Following the literature on reliance and advice adherence (Guo et al., 2024; Chen et al., 2024), we use θ to denote the switch fraction: the fraction of trials in which the final decision agrees with the AI recommendation among trials in which the pre-advice judgment differs from the AI recommendation:

$$\theta := \Pr(\hat{y} = \hat{y}_{\text{a}} \mid \hat{y}_{\text{h}} \neq \hat{y}_{\text{a}}). \tag{4}$$

We first treat θ as known. Accuracy can change only when the human switches to the AI recommendation, and each switch changes correctness by at most one. Thus,

$$|\Delta| \leq s_\theta := \Pr(\hat{y} = \hat{y}_{\text{a}}, \hat{y}_{\text{h}} \neq \hat{y}_{\text{a}}) = \frac{\theta(1-g)}{1-\theta}. \tag{5}$$

For a fixed agreement fraction, a low switch fraction therefore rules out large accuracy gains and losses. The sharp bounds also account for whether the AI recommendation is correct and, in multiclass tasks, for switches between two incorrect labels.

Proposition 2 (Sharp accuracy bounds given a switch fraction)**.** *A candidate switch fraction is compatible with the final-only distribution if and only if* $0 \le \theta \le g$*. For any compatible* θ*,*

$$\Delta \in [L_\theta, U_\theta], \tag{6}$$

where

$$L_\theta = \max\{-s_\theta,\ s_\theta - 2p_-\},$$

and

$$U_\theta = \begin{cases} \min\{s_\theta,\ 2p_+ - s_\theta\}, & |\mathbf{D}| = 2, \\ \min\{s_\theta,\ p_+\}, & |\mathbf{D}| \ge 3. \end{cases}$$

In multiclass tasks, the upper bound allows accuracy-neutral switches between two incorrect labels. Thus, knowing how often switching occurs can narrow the range without determining whether those switches improve or reduce accuracy. Appendix A.2 proves sharpness and allows the switch fraction itself to vary over a supplied range.

The agreement–disagreement gap. The switch fraction restricts how often an initial disagreement ends with a switch to the AI recommendation, but it does not determine how pre-advice accuracy differs between trials that ultimately agree and disagree with the AI. This distinction is behaviorally relevant because confidence is related both to correctness and to whether people follow AI advice (Tejeda et al., 2022; Vodrahalli et al., 2022a; He et al., 2023b).

When $\hat{y} \neq \hat{y}_\mathrm{a}$, the final decision reveals the pre-advice judgment in our setting. Hence the observed accuracy on these trials is also their pre-advice accuracy:

$$p_\mathrm{disagree} := \Pr(\hat{y} = y \mid \hat{y} \neq \hat{y}_\mathrm{a}) = \Pr(\hat{y}_\mathrm{h} = y \mid \hat{y} \neq \hat{y}_\mathrm{a}) =: q_\mathrm{disagree}.$$

A bound on the agreement–disagreement gap links this observed quantity to pre-advice accuracy on agreement trials:

$$|q_\mathrm{agree} - q_\mathrm{disagree}| \le \kappa, \qquad 0 \le \kappa \le 1,$$

where κ controls how much pre-advice accuracy may differ between the two groups. This gives a one-dimensional sensitivity analysis (Masten & Poirier, 2018; 2020).

The accuracy change when agreement and disagreement trials have equal pre-advice accuracy is denoted by

$$\Delta_\mathrm{equal} := p_+ - g p_\mathrm{disagree}.$$

Equation 2 then gives

$$\Delta \in [\Delta_\mathrm{equal} - g\kappa,\ \Delta_\mathrm{equal} + g\kappa].$$

Combining this restriction with the switch-fraction bounds gives the following result.

Proposition 3 (Sharp accuracy bounds given a switch fraction and the agreement–disagreement gap)**.** *For any compatible* θ *and any* $0 \le \kappa \le 1$*, the sharp bounds are*

$$\Delta \in [L_\theta, U_\theta] \cap [\Delta_\mathrm{equal} - g\kappa,\ \Delta_\mathrm{equal} + g\kappa]. \tag{7}$$

The agreement–disagreement gap narrows the switch-fraction bounds whenever it excludes values that remain possible under the switch fraction alone. At $\kappa = 1$, the switch-fraction bounds are recovered; as κ decreases, the range can only narrow. At $\kappa = 0$, compatibility requires $\Delta_\mathrm{equal} \in [L_\theta, U_\theta]$. Considered on its own, the agreement–disagreement gap determines the sign whenever

$$\kappa < \frac{|\Delta_\mathrm{equal}|}{g}, \qquad \Delta_\mathrm{equal} \neq 0.$$

Retaining the switch-fraction information can determine the sign in additional cases. Appendix A.2 gives the proof and Appendix A.2 gives an example.

Together, the switch fraction and the agreement–disagreement gap describe how often people switch to the AI recommendation and how pre-advice accuracy differs between final-agreement and final-disagreement trials. The next subsection considers behavioral models that place additional structure on pre-advice judgments and responses to AI advice.

3.3 BEHAVIORAL MODELS

Behavioral models go beyond the switch fraction and the agreement–disagreement gap by specifying how pre-advice judgments and responses to AI advice vary with context. In the cumulative comparison below, the restrictions on the switch fraction and agreement–disagreement gap are retained when a behavioral model is added, so the model can narrow the preceding bounds further. We consider two modeling approaches. The Bayesian cognitive approach specifies pre-advice judgments and an advice-taking policy. The activation–integration model, introduced by Vodrahalli et al. (2022a) and instantiated with learned activation and integration models by Vodrahalli et al. (2022b), specifies how responses change after advice. Throughout this subsection, z denotes recorded context, including the ground truth and AI recommendation.

Cognitive model. Bayesian cognitive modeling provides an established framework for explaining human judgment and decision-making under uncertainty (Griffiths et al., 2010; Vul et al., 2014). A related line of work on human–AI decision-making models judgments, confidence, and the combination of human and AI predictions (Steyvers et al., 2022; Tejeda et al., 2022; Li & Steyvers, 2026). We use a representative cognitive specification from this line of work to study how assumptions about pre-advice judgments and advice taking constrain accuracy change. A Bayesian item-response component models pre-advice correctness from person ability, item difficulty, and item discrimination, while the advice-taking policy determines the probability of switching to the AI recommendation as a function of confidence (Tejeda et al., 2022). At context z and parameters ψ, let $\pi_\psi(\cdot \mid z)$ denote the pre-advice label distribution and $q_\psi(z) = \pi_\psi(y \mid z)$ its accuracy. The model distributes incorrect pre-advice judgments uniformly across the wrong labels. Averaging the advice-taking policy over confidence gives the switching probability $\bar{\alpha}_\psi(z)$. The resulting final-label distribution is

$$p_\psi(f \mid z) = [1 - \bar{\alpha}_\psi(z)]\pi_\psi(f \mid z) + \bar{\alpha}_\psi(z)\mathbf{1}\{f = \hat{y}_\mathrm{a}\}. \tag{8}$$

This structure can identify accuracy change when the AI recommendation is incorrect in a multiclass task. For $|\mathbf{D}| \geq 3$ and $\hat{y}_\mathrm{a} \neq y$, let

$$u = \Pr(\hat{y} = y \mid z), \qquad v = \Pr(\hat{y} \notin \{y, \hat{y}_\mathrm{a}\} \mid z).$$

Writing $t = 1 - \bar{\alpha}_\psi(z)$ for the probability of keeping the pre-advice label, the model implies

$$t = u + \frac{|\mathbf{D}| - 1}{|\mathbf{D}| - 2}v, \qquad q_\psi(z) = \frac{u}{t}, \qquad \Delta(z) = u - \frac{u}{t}. \tag{9}$$

Thus the final responses identify pre-advice accuracy and accuracy change in these incorrect-advice contexts. With correct AI advice, however, the same model leaves $\Delta(z) \in [0, g(z)]$, where $g(z) = \Pr(\hat{y} = \hat{y}_\mathrm{a} \mid z)$, the conditional baseline interval from Section 2. Without restrictions shared across contexts, the model therefore obtains its identifying power only from contexts with incorrect AI advice. Because $q_\psi(z)$ is parameterized by person ability, item difficulty, and item discrimination, auxiliary no-advice judgments can also inform pre-advice accuracy, including in contexts with correct AI advice. Relaxing either the uniform allocation of incorrect judgments across wrong labels or the advice-taking policy can only widen the resulting bounds. Appendix A.4 gives the corresponding identification and sensitivity results.

Activation–integration model. Unlike the cognitive model, the activation–integration model leaves the pre-advice response distribution unrestricted and instead specifies how responses change after advice. In the two-stage formulation of Vodrahalli et al. (2022a), activation determines whether a person uses the advice, and integration determines how the response is updated conditional on activation. On a finite response scale $\{r_1, \ldots, r_d\}$, let a_j denote the activation probability for pre-advice response r_j, and let B_{kj} denote the probability of final response r_k conditional on activation. If h and p are the pre-advice and final response distributions, then

$$K_{kj} = (1 - a_j)\mathbf{1}\{k = j\} + a_j B_{kj}, \qquad p = Kh. \tag{10}$$

Let c_j indicate whether response r_j predicts the true label. If $c \in \mathrm{row}(K)$, there exists b such that $K^\top b = c$, giving $\Delta(z) = c^\top p - b^\top p$. Thus, accuracy change can be identified even when the full pre-advice response distribution is not. Enlarging the admissible activation probabilities or integration distributions can only widen the bounds; fixing both recovers the known-kernel case. Appendix A.5 gives the sharp bounds and the corresponding sensitivity results.

4 Empirical Evaluation and Audit through the Lens

Section 2 shows that final-only records leave uncertainty about pre-advice correctness on agreement trials. Section 3 develops two ways to reduce this uncertainty: measuring pre-advice correctness and using information about how people respond to AI advice. We compare these approaches across six tasks with paired pre-advice and final judgments, and then examine what information is available in published human–AI decision-making studies.

4.1 Identification Precision Across Information Sources

We evaluate five binary classification tasks from HAIID (Vodrahalli et al., 2022a) and a 16-class object-recognition task from Tejeda et al. (2022). All six record paired pre-advice and final judgments, allowing us to mask the pre-advice judgments for evaluation and use separate participants for calibration. We vary the fraction of participants measured directly and compare the resulting widths with behavioral constraints calibrated on half the participants (Figure 2). The remaining participants form the evaluation sample. Appendices E.2 and E.4 specify the two calibration methods.

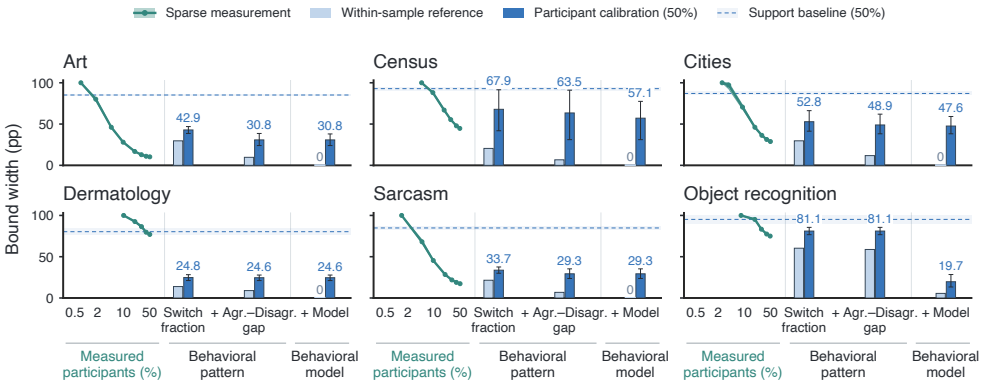


Figure 2: **Widths of the bounds on accuracy change in six tasks.** Direct measurement targets 1%–50% of participants, with at least two per stratum. Measured percentages use a log scale. Behavioral patterns and models use 50% participant calibration. Calibrated means use 100 participant splits per setting. Evaluation uses the remaining participants. Within-sample references and support-only baselines are included. Shading and error bars show the 10th–90th percentiles of widths. Direct intervals use finite-population calibration; behavioral intervals retain bootstrap calibration. Smaller widths indicate tighter bounds, illustrating the tradeoff between direct measurement effort and reliance on behavioral constraints.

Final-only records leave the sign of accuracy change unresolved in all six tasks. Although the recorded pre-advice judgments reveal positive accuracy changes in every task, masking them leaves both improvement and harm compatible with the final-only data. Increasing the final-only sample reduces sampling error without eliminating this ambiguity (Appendix E.1).

Measuring a few participants does not by itself resolve uncertainty about accuracy change. At the smallest measurement budgets (i.e., a requested 1% of participants), the direct intervals remain 100 percentage points wide and include both gains and losses. Across the tested budgets, every direct interval contains the evaluation-sample accuracy change (Appendix E.4).

The switch fraction limits how many decisions advice can change; their effect on accuracy depends on the correctness of the initial judgments. These behavioral patterns narrow the bounds without specifying a full model of individual decisions. For example, in Art, the switch fraction and agreement–disagreement gap together reduce the mean width from 85.2 to 30.8 percentage points. Appendix E.2 examines how this precision changes when the two restrictions are weakened.

Behavioral models tighten bounds, while predictive improvements do not directly translate into tighter bounds. To narrow the uncertainty left by aggregate behavioral patterns, fitted models constrain both initial judgments and their transformation into final decisions. With 50% participant calibration, adding model constraints reduces mean interval width from 81.07 to 19.70 percentage points in object recognition, from 63.46 to 57.13 in Census, and from 48.91 to 47.60 in Cities (Figure 2). However, predictive improvement need not translate into comparable bound contraction. Across the five binary tasks, mean Brier loss falls by 31.3%–52.8% relative to the context-only baseline, whereas adding the fitted models to behavioral patterns reduces mean gain-interval width by only 0%–10.0% (Figure 3; e.g., Dermatology: loss $\downarrow$ 52.8%, width unchanged). The within-task association is weak (partial $R^2 = 0.017$; see Appendix E.3.1). Appendix E.3 reports component ablations and constraint-validity checks.

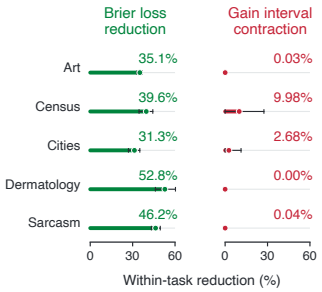


Figure 3: Comparison of prediction and identification gains (see Appendix E.3.1).

The tradeoff between sparse direct measurement and behavioral constraints depends on the measurement budget and the task. Behavioral constraints give tighter bounds at very small direct-measurement budgets. In four of the six tasks, the tested budgets reveal a measurement threshold beyond which sparse direct measurement yields narrower mean intervals than behavioral models. For example, in Art, direct measurement first becomes more precise at a measured fraction of 9.7%. In Dermatology and object recognition, behavioral models remain more precise even at the largest tested budget (a requested 50% of participants). These results motivate planning how much pre-advice information to collect to achieve a desired interval width, given elicitation costs and the behavioral restrictions that can be justified.

4.2 Auditing the Literature through the Analytical Lens

We audit 96 claim–workflow instances from the 56 empirical studies reviewed by Raees et al. (2026). We treat each distinct combination of a reported claim, workflow or condition, and the information used to support it as one instance. The coding protocol, formal identification rule, and inter-coder agreement analysis are reported in Appendix F.

Table 1: **Information requirements by target claim.**

Target	Information needed	Identification and limits
Final accuracy	Ground truth and final decision	Identified directly.
Effect of offering advice	Valid assignment and a matching baseline	Identifies the assistance effect, not pre-to-final accuracy change.
Average accuracy change	Pre-advice correctness in the target workflow	Known accuracy identifies Δ; behavioral constraints and models can narrow the bounds.
Individual switching	Pre-advice judgment or direct change indicator	The aggregate switch fraction does not identify individual switches.
Observed agreement / adherence	AI recommendation and final decision	Identifies observed agreement; switching requires pre-advice information.
Latent reliance / mechanism	Behavioral model and any required auxiliary data	Identified under the specified model.

Recorded data or study design identify most in-scope targets. Among the 94 in-scope instances, recorded data identify 49 targets and study design identifies 31. In 12 further cases, an observed statistic is identified while its stronger behavioral interpretation remains unresolved; two depend on a behavioral model for identification. Figure 4 shows how these cases are distributed across claim families. Among these 12 cases, nine lack a pre-advice judgment; their records leave pre-advice agreement and switching observationally indistinguishable. In the remaining three, switching is observed while multiple behavioral mechanisms remain compatible with the record. These cases include interpretations in terms of anchoring (Rastogi et al., 2022) and automation bias (Wysocki et al., 2023; Vered et al., 2023).

Reconstructed agreement rates imply wide bounds on accuracy change. Across 13 papers for which final agreement can be reconstructed exactly, the median agreement fraction is 0.75 (Appendix Table 20). When pre-advice judgments are not recorded, Proposition 1 makes the agreement

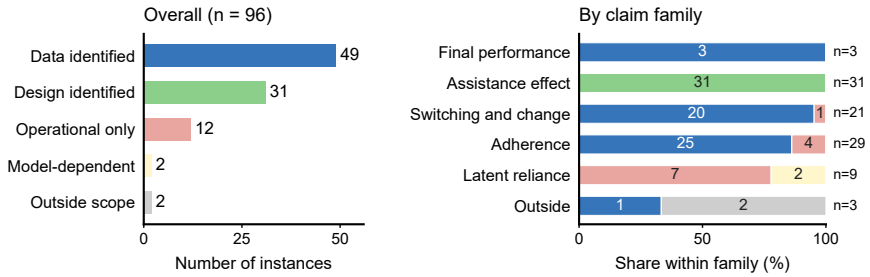


Figure 4: **Identification status of 96 claim–workflow instances.** Left: overall counts, with colors denoting status. Right: within-family shares; labels give counts and n gives family size. "Outside" is a claim family, distinct from the out-of-scope status. Two instances fall outside the audit scope.

fraction the baseline width of the accuracy-change range. The median therefore corresponds to a 75-percentage-point range.

5 Related work.

Human–AI decision-making is an important research area aimed at improving decision quality by combining human judgment with AI capabilities (Raees et al., 2026; Shi et al., 2025). Work on human–AI teams examines complementary team performance (Bansal et al., 2021; Wilder et al., 2020; Kerrigan et al., 2021; Straitouri et al., 2023; Straitouri & Gomez Rodriguez, 2024), while learning-to-defer methods study how to allocate decisions between people and AI (Madras et al., 2018; Mozannar & Sontag, 2020; Charusaie et al., 2022; Mozannar et al., 2023). Realizing these benefits also requires understanding how people respond to advice. Research on advice taking examines how people revise their judgments after receiving human or AI recommendations (Lu & Yin, 2021; Vodrahalli et al., 2022a; Schemmer et al., 2023b; Bonaccio & Dalal, 2006; Soll & Larrick, 2009; Bailey et al., 2023), and reliance studies distinguish observed agreement from stronger claims about how people use advice (Guo et al., 2024; Schöffer et al., 2025). These questions call for different forms of evaluation: paired initial and final judgments reveal revisions, while randomized comparisons can estimate the effect of offering advice on final outcomes (Ben-Michael et al., 2025).

When initial judgments or advice use are unobserved, behavioral models provide a route to inferring these latent quantities from auxiliary data and assumptions (Tejeda et al., 2022; Corvelo Benz & Gomez Rodriguez, 2022). Bayesian models similarly describe how beliefs change with advice (Steyvers et al., 2022). The resulting interpretations depend on the assumptions linking unobserved states to recorded decisions, motivating our question of what the evidence and behavioral restrictions jointly establish. We connect this question to partial identification by characterizing all accuracy changes consistent with the records and maintained constraints (Manski, 1989).

6 Conclusion and Discussion

We developed an identification framework for assessing accuracy change in human–AI decision-making when pre-advice judgments are unobserved. Final-only records leave accuracy change uncertain, and collecting more of them cannot eliminate this ambiguity. Sparse direct measurements, behavioral patterns, and behavioral models narrow the range by restricting which initial judgments remain possible. Across six tasks, their relative precision varies with the task and measurement budget, making the choice between collecting initial judgments and supporting behavioral constraints a study-design tradeoff. The literature audit further shows why evidence should be matched to the target claim: most targets are identified by recorded data or study design, while stronger behavioral interpretations can require additional information. These findings motivate evaluations that make their targets, measurements, and assumptions explicit. Extending the framework to open-ended decisions, characterizing cost-effective measurement budgets, and accounting for elicitation-induced changes in behavior are important directions for future work.

AI USE STATEMENT

We used ChatGPT to assist with identifying potentially relevant related work and with language editing. We used Codex for limited assistance in implementing and running author-designed analyses. The authors specified the research questions, methods, estimands, and analytical choices, and manually reviewed and verified all AI-assisted code, calculations, citations, analyses, and text. All scientific interpretations and final decisions were made by the authors, who take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

Appendices A–F provide the theory extensions, sensitivity results, workflow implications, proofs, public-data analyses, and claim-level audit records supporting the main text. Reported empirical values are derived from the cited public releases and the audit sources documented in Appendices E–F. The direct-measurement refresh preserves the original result files and participant partitions; `run_direct_refresh.py` produces the new intervals and verification records, and `build_direct_refresh_outputs.py` generates the updated manuscript tables.

REFERENCES

Nikhil Agarwal, Alex Moehring, and Alexander Wolitzky. Designing human–ai collaboration: A sufficient-statistic approach. Working Paper 33949, National Bureau of Economic Research, 2025. URL https://www.nber.org/papers/w33949.

Phoebe E. Bailey, Tarren Leon, Natalie C. Ebner, Ahmed A. Moustafa, and Gabrielle Weidemann. A meta-analysis of the weight of advice in decision-making. *Current Psychology*, 42(28):24516–24541, 2023. doi: 10.1007/s12144-022-03573-2.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445717.

Eli Ben-Michael, D. James Greiner, Melody Huang, Kosuke Imai, Zhichao Jiang, and Sooahn Shin. Does AI help humans make better decisions? a statistical evaluation framework for experimental and observational studies. *Proceedings of the National Academy of Sciences*, 122(38):e2505106122, 2025. doi: 10.1073/pnas.2505106122.

Silvia Bonaccio and Reeshad S. Dalal. Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes*, 101(2):127–151, 2006. doi: 10.1016/j.obhdp.2006.07.001.

Michelle Brachman, Zahra Ashktorab, Michael Desmond, Evelyn Duesterwald, Casey Dugan, Narendra Nath Joshi, Qian Pan, and Aabhas Sharma. Reliance and automation for human-AI collaborative data labeling conflict resolution. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27, 2022. doi: 10.1145/3555212.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 454–464. Association for Computing Machinery, 2020. doi: 10.1145/3377325.3377498.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021. doi: 10.1145/3449287.

Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. Contrastive explanations that anticipate human misconceptions can improve human decision-making skills. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–25. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713229.

Federico Cabitza, Andrea Campagner, Riccardo Angius, Chiara Natali, and Carlo Reverberi. AI shall have no dominion: on how to measure technology dominance in AI-supported human decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20. Association for Computing Machinery, 2023. doi: 10.1145/3544548.3581095.

Federico Cabitza, Chiara Natali, Lorenzo Famiglini, Andrea Campagner, Valerio Caccavella, and Enrico Gallazzi. Never tell me the odds: Investigating pro-hoc explanations in medical decision making. *Artificial intelligence in medicine*, 150:102819, 2024.

Federico Cabitza, Lorenzo Famiglini, Caterina Fregosi, Samuele Pe, Enea Parimbelli, Giovanni Andrea La Maida, and Enrico Gallazzi. From oracular to judicial: Enhancing clinical decision making through contrasting explanations and a novel interaction protocol. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 745–754. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712157.

Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. Improving human-AI collaboration with descriptions of AI behavior. *Proceedings of the ACM on Human-Computer Interaction*, 7 (CSCW1):1–21, 2023. doi: 10.1145/3579612.

Shiye Cao and Chien-Ming Huang. Understanding user reliance on AI in assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), 2022. doi: 10.1145/3555572.

Shiye Cao, Catalina Gomez, and Chien-Ming Huang. How time pressure in different phases of decision-making influences human-AI collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, 2023. doi: 10.1145/3610068.

Shiye Cao, Anqi Liu, and Chien-Ming Huang. Designing for appropriate reliance: The roles of AI uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–32, 2024. doi: 10.1145/3637318.

Federico Maria Cau and Lucio Davide Spano. The influence of curiosity traits and on-demand explanations in AI-assisted decision-making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 1440–1457. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712165.

Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Effects of AI and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–42, 2023a. doi: 10.1145/3588320.

Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Supporting high-uncertainty decisions through AI and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 251–263. Association for Computing Machinery, 2023b. doi: 10.1145/3581641.3584080.

Mohammad-Amin Charusaie, Hussein Mozannar, David Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2972–3005, 2022. URL https://proceedings.mlr.press/v162/charusaie22a.html.

Guanting Chen, Xiaocheng Li, Chunlin Sun, and Hanzhao Wang. Learning to make adherence-aware advice. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6cda6dae05ae5e42ea78be85d5a26f77-Abstract-Conference.html.

Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction*, 7, 2023. doi: 10.1145/3610219.

Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces*, pp. 148–161, 2022. doi: 10.1145/3490099.3511121.

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. Are two heads better than one in AI-assisted decision making? comparing the behavior and performance of groups and individuals in human-AI collaborative recidivism risk assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023. doi: 10.1145/3544548.3581015.

Nina L. Corvelo Benz and Manuel Gomez Rodriguez. Counterfactual inference of second opinions. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 453–463. PMLR, 2022. URL https://proceedings.mlr.press/v180/corvelo-benz22a.html.

Jesse C. Cresswell, Yi Sui, Bhargava Kumar, and Noël Vouitsis. Conformal prediction sets improve human decision making. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9439–9457, 2024. URL https://proceedings.mlr.press/v235/cresswell24a.html.

Sander de Jong, Ville Paananen, Benjamin Tag, and Niels van Berkel. Cognitive forcing for better decision-making: Reducing overreliance on AI systems through partial explanations. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–30, 2025. doi: 10.1145/3710946.

Murat Dikmen and Catherine Burns. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies*, 162:102792, 2022. doi: 10.1016/j.ijhcs.2022.102792.

Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Michael Fitzke, Mark Parkinson, Diane Wilson, Paul Fisher, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? influences of human–ai workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1362–1374, 2022. doi: 10.1145/3531146.3533193.

Raffaella Giacomini and Toru Kitagawa. Robust bayesian inference for set-identified models. *Econometrica*, 89(4):1519–1556, 2021. doi: 10.3982/ECTA16773.

Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019. doi: 10.1145/3359280.

Thomas L. Griffiths, Nick Chater, Charles Kemp, Amy Perfors, and Joshua B. Tenenbaum. Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364, 2010. doi: 10.1016/j.tics.2010.05.004.

Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica Hullman. A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 221–236, 2024. doi: 10.1145/3630106.3658901.

Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. The value of information in human–ai decision-making. In *International Conference on Learning Representations*, 2026.

Gaole He, Stefan Buijsman, and Ujwal Gadiraju. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction*, 7, 2023a. doi: 10.1145/3610067.

Gaole He, Lucie Kuiper, and Ujwal Gadiraju. Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023b. doi: 10.1145/3544548.3581025. arXiv:2301.11333.

Gaole He, Abri Bharos, and Ujwal Gadiraju. To err is AI! debugging as an intervention to facilitate appropriate reliance on AI systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media*, pp. 98–105, 2024. doi: 10.1145/3648188.3675130. URL https://arxiv.org/abs/2409.14377.

Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. Is conversational XAI all you need? human-AI decision making with a conversational XAI assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 907–924. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712133.

Mark Himmelstein and David V. Budescu. Causal inference in the absence of direct measurement using planned missing data, 2026. PsyArXiv preprint.

Joel L. Horowitz and Charles F. Manski. Identification and robustness with contaminated and corrupted data. *Econometrica*, 63(2):281–302, 1995. doi: 10.2307/2951627.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004. doi: 10.1111/j.1468-0262.2004.00555.x.

Youngseung Jeon, Christopher Hwang, and Xiang Anthony Chen. Empowering medical data labeling for non-experts with DANNY: Enhancing accuracy and mitigating over-reliance on AI. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 624–640. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712161.

Gavin Kerrigan, Padhraic Smyth, and Mark Steyvers. Combining human predictions with model probabilities via confusion matrices and calibration. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4421–4434, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/234b941e88b755b7a72a1c1dd5022f30-Abstract.html.

Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. "I'm Not Sure, But...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, 2024. doi: 10.1145/3630106.3658941.

Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 29–38. Association for Computing Machinery, 2019. doi: 10.1145/3287560.3287590.

Min Hun Lee and Chong Jun Chew. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative clinical decision making. *Proceedings of the ACM on human-computer interaction*, 7(CSCW2):1–22, 2023.

ZhaoBin Li and Mark Steyvers. Modeling the joint impact of human and AI metacognitive sensitivity on human–ai collaboration. *Journal of Mathematical Psychology*, 129:102988, 2026. doi: 10.1016/j.jmp.2026.102988.

Zhuoran Lu and Ming Yin. Human reliance on machine learning models when performance feedback is limited: Heuristics and risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 78:1–78:16. Association for Computing Machinery, 2021. doi: 10.1145/3411764.3445562.

Zhuoran Lu, Dakuo Wang, and Ming Yin. Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction*, 8 (CSCW1):1–31, 2024. doi: 10.1145/3653708.

Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. Who should I trust: AI or myself? leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2023. doi: 10.1145/3544548.3581058.

Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. "Are You Really Sure?" understanding the effects of human self-confidence calibration in AI-assisted decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024. doi: 10.1145/3613904.3642671.

Shuai Ma, Qiaoyi Chen, Xinru Wang, Chengbo Zheng, Zhenhui Peng, Ming Yin, and Xiaojuan Ma. Towards human-AI deliberation: Design and evaluation of LLM-empowered deliberative AI for AI-assisted decision-making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713423.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Charles F. Manski. Anatomy of the selection problem. *Journal of Human Resources*, 24(3):343–360, 1989. doi: 10.2307/145818.

Charles F. Manski. Nonparametric bounds on treatment effects. *American Economic Review*, 80(2): 319–323, 1990.

Charles F. Manski. Statistical treatment rules for heterogeneous populations. *Econometrica*, 72(4): 1221–1246, 2004. doi: 10.1111/j.1468-0262.2004.00530.x.

Matthew A. Masten and Alexandre Poirier. Identification of treatment effects under conditional partial independence. *Econometrica*, 86(1):317–351, 2018. doi: 10.3982/ECTA14481.

Matthew A. Masten and Alexandre Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020. doi: 10.3982/QE1288.

Hyungsik Roger Moon and Frank Schorfheide. Bayesian and frequentist inference in partially identified models. *Econometrica*, 80(2):755–782, 2012. doi: 10.3982/ECTA8360.

Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. Evaluating the impact of human explanation strategies on human-ai visual decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–37, 2023.

Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. The impact of imperfect XAI on human-AI decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–39, 2024. doi: 10.1145/3641022.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020.

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David Sontag. Who should predict? exact algorithms for learning to defer to humans. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10520–10545, 2023. URL `https://proceedings.mlr.press/v206/mozannar23a.html`.

Mahsan Nourani, Chiradeep Roy, Jeremy E. Block, Donald R. Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, 2021. doi: 10.1145/3397481.3450639.

Saumya Pareek, Niels van Berkel, Eduardo Velloso, and Jorge Goncalves. Effect of explanation conceptualisations on trust in AI-assisted credibility assessment. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2):1–31, 2024. doi: 10.1145/3686922.

Lu Peng, Dailin Li, Zhaotong Zhang, Tingru Zhang, Anqi Huang, Shaohui Yang, and Yu Hu. Human-AI collaboration: Unraveling the effects of user proficiency and AI agent capability in intelligent decision support systems. *International Journal of Industrial Ergonomics*, 103:103629, 2024. doi: 10.1016/j.ergon.2024.103629.

Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. doi: 10.1145/3411764.3445315.

Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. Understanding uncertainty: How lay decision-makers perceive and interpret uncertainty in human-AI decision making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 379–396. Association for Computing Machinery, 2023. doi: 10.1145/3581641.3584033.

Muhammad Raees, Vassilis-Javed Khan, Ioanna Lykourentzou, and Konstantinos Papangelis. Do people appropriately rely on AI-advice? an analytical review of HCI research on human-AI decision-making. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 805:1–805:24. Association for Computing Machinery, 2026. doi: 10.1145/3772318.3791467.

Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–22, 2022. doi: 10.1145/3512930.

Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024. doi: 10.1145/3613904.3641905.

Max Schemmer, Andrea Bartos, Philipp Spitzer, Patrick Hemmer, Niklas Kühl, Jonas Liebschner, and Gerhard Satzger. Towards effective human-AI decision-making: The role of human learning in appropriate reliance on AI advice. In *Proceedings of the International Conference on Information Systems*, 2023a. URL https://aisel.aisnet.org/icis2023/hti/hti/14/.

Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pp. 410–422, 2023b. doi: 10.1145/3581641.3584066.

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl. Explanations, fairness, and appropriate reliance in human-AI decision-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024. doi: 10.1145/3613904.3642621.

Jakob Schöffer, Johannes Jakubik, Michael Vössing, Niklas Kühl, and Gerhard Satzger. AI reliance and decision quality: Fundamentals, interdependence, and the effects of interventions. *Journal of Artificial Intelligence Research*, 82:471–501, 2025. doi: 10.1613/JAIR.1.15873.

Quan Shi, Carlos E. Jimenez, Shunyu Yao, Nick Haber, Diyi Yang, and Karthik Narasimhan. When models know more than they can explain: Quantifying knowledge transfer in human-AI collaboration. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.nips.cc/paper_files/paper/2025/hash/975d11c51406cd10be48e47b36fb8698-Abstract-Conference.html.

Jack B. Soll and Richard P. Larrick. Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35 (3):780–805, 2009. doi: 10.1037/a0015145.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–ai complementarity. *Proceedings of the National Academy of Sciences*, 119(11):e2111547119, 2022. doi: 10.1073/pnas.2111547119.

Jörg Stoye. Minimax regret treatment choice with finite samples. *Journal of Econometrics*, 151(1): 70–81, 2009. doi: 10.1016/j.jeconom.2009.02.013.

Eleni Straitouri and Manuel Gomez Rodriguez. Designing decision support systems using counterfactual prediction sets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46722–46744, 2024. URL https://proceedings.mlr.press/v235/straitouri24a.html.

Eleni Straitouri, Lequn Wang, Nastaran Okati, and Manuel Gomez Rodriguez. Improving expert predictions with conformal prediction. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32633–32653, 2023. URL https://proceedings.mlr.press/v202/straitouri23a.html.

Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. Accuracy-time tradeoffs in AI-assisted decision making under time pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 138–154. Association for Computing Machinery, 2024. doi: 10.1145/3640543.3645206.

Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. Personalising AI assistance based on overreliance rate in AI-assisted decision making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pp. 1107–1122. Association for Computing Machinery, 2025. doi: 10.1145/3708359.3712128.

Heliodoro Tejeda, Aakriti Kumar, Padhraic Smyth, and Mark Steyvers. Ai-assisted decision-making: a cognitive modeling approach to infer latent reliance strategies. *Computational Brain & Behavior*, 5:491–508, 2022. doi: 10.1007/s42113-022-00157-y.

Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023. doi: 10.1145/3579605.

Mor Vered, Tali Livni, Piers Douglas Lionel Howe, Tim Miller, and Liz Sonenberg. The effects of explanations on automation bias. *Artificial Intelligence*, 322:103952, 2023. doi: 10.1016/j.artint.2023.103952.

Kailas Vodrahalli, Roxana Daneshjou, Tobias Gerstenberg, and James Zou. Do humans trust advice more if it comes from AI? an analysis of human–ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 763–777, 2022a. doi: 10.1145/3514094.3534150.

Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. Uncalibrated models can improve human-AI collaboration. In *Advances in Neural Information Processing Systems*, volume 35, pp. 4004–4016, 2022b.

Edward Vul, Noah Goodman, Thomas L. Griffiths, and Joshua B. Tenenbaum. One and done? optimal decisions from very few samples. *Cognitive Science*, 38(4):599–637, 2014. doi: 10.1111/cogs.12101.

Xinru Wang and Ming Yin. Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 318–328, 2021. doi: 10.1145/3397481.3450650.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Bryan Wilder, Eric Horvitz, and Ece Kamar. Learning to complement humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 1526–1533, 2020. doi: 10.24963/ijcai.2020/212.

Oskar Wysocki, Jessica Katharine Davies, Markel Vigo, Anne Caroline Armstrong, Dónal Landers, Rebecca Lee, and André Freitas. Assessing the communication gap between AI models and healthcare professionals: Explainability, utility and trust in AI-driven clinical decision-making. *Artificial Intelligence*, 316:103839, 2023. doi: 10.1016/j.artint.2022.103839.

Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 189–201, 2020. doi: 10.1145/3377325.3377480.

Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019. doi: 10.1145/3290605.3300509.

Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305, 2020. doi: 10.1145/3351095.3372852.

Appendix roadmap. Appendix A extends the discrete argument to other targets and response scales. Appendix B records how measurements and behavioral constraints act on pre-advice correctness in agreement trials. Appendix C maps target claims to the variables a study must record. Appendix D collects the shared constructions and statistical arguments; direct consequences are justified beside their statements. Appendix E follows the constraint comparison from baseline ambiguity through calibrated mechanisms to direct measurement. Appendix F documents the claim-level audit and recoverable agreement rates.

Unless stated otherwise, the appendices use the same decision process and notation as the main text: $p_+ = \Pr(\hat{y} = \hat{y}_\mathrm{a} = y)$, $p_- = \Pr(\hat{y} = \hat{y}_\mathrm{a},\, \hat{y}_\mathrm{a} \neq y)$, and $g = \Pr(\hat{y} = \hat{y}_\mathrm{a})$; their context-specific versions condition on z. We retain q_agree and q_disagree for pre-advice accuracy on agreement and disagreement trials. Let p_o denote the distribution of final-only records $o = (y, \hat{y}_\mathrm{a}, \hat{y})$, and let $\mathbf{M}_\mathrm{KA}(p_o)$ contain the joint distributions of $(y, \hat{y}_\mathrm{h}, \hat{y}_\mathrm{a}, \hat{y})$ that satisfy keep-or-adopt and reproduce p_o. The interval logic is standard: an observed-data distribution may be compatible with multiple joint distributions and hence multiple values of the target (Manski, 1989; 1990; Horowitz & Manski, 1995). In advice taking, disagreement reveals $\hat{y}_\mathrm{h}$, so that range sits in the agreement trials.

A Further identification results

This appendix asks how far the baseline argument travels when the target, the response scale, or the loss changes. Final-only samples still pin down the recorded distribution; any remaining range comes from hidden pre-advice judgments that the maintained constraint does not rule out.

A.1 Switching to AI advice

The probability that a person changes a pre-advice judgment to match AI advice has a sharp range. For any distribution of final-only records with $g = \Pr(\hat{y} = \hat{y}_\mathrm{a})$ and at least two possible answers,

$$\Pr(\hat{y} = \hat{y}_\mathrm{a}, \hat{y}_\mathrm{h} \neq \hat{y}_\mathrm{a}) \in [0, g].$$

Only agreement trials can contribute: setting $\hat{y}_\mathrm{h} = \hat{y}_\mathrm{a}$ or $\hat{y}_\mathrm{h} \neq \hat{y}_\mathrm{a}$ on those trials attains the endpoints, and mixtures fill the interval. For $g > 0$, the corresponding share of final agreement has sharp range $[0, 1]$. This describes switching, not a psychological mechanism.

A.2 Switch fractions and relational constraints

Proof of Proposition 2. Let $d = 1 - g > 0$ and $s = \Pr(\hat{y} = \hat{y}_\mathrm{a}, \hat{y}_\mathrm{h} \neq \hat{y}_\mathrm{a})$. The event $\{\hat{y}_\mathrm{h} \neq \hat{y}_\mathrm{a}\}$ is the disjoint union of switching to advice and final disagreement. Hence its probability is $s + d$, so $\theta = s/(s + d)$ and $s = s_\theta = \theta d/(1 - \theta)$. The feasible range $s \in [0, g]$ from Appendix A.1 corresponds exactly to $\theta \in [0, g]$.

Partition switches into beneficial mass b, harmful mass h, and accuracy-neutral mass n. The last is zero in binary tasks; in multiclass tasks it consists of switches between two wrong labels. These masses obey

$$b, h, n \geq 0, \qquad b \leq p_+, \qquad h + n \leq p_-, \qquad b + h + n = s_\theta, \qquad \Delta = b - h. \tag{11}$$

For binary labels, set $n = 0$. All such allocations are attainable while preserving every observed cell: on correct-advice agreement events, assign the pre-advice label to be incorrect on the chosen beneficial mass; on incorrect-advice agreement events, assign it to truth on the harmful mass, to a third label on the neutral mass, and to advice on the remainder. On final disagreement events keep the pre-advice label equal to the final label. Fractional allocation within observed cells preserves their full distribution. A third label exists whenever $|\mathbf{D}| \geq 3$.

For either label space, minimizing Δ uses no neutral switches and the smallest possible beneficial mass, $b = \max(0, s_\theta - p_-)$. This gives $L_\theta = 2\max(0, s_\theta - p_-) - s_\theta$. In the binary case, maximizing uses $b = \min(s_\theta, p_+)$ and $h = s_\theta - b$, giving $U_\theta = 2\min(s_\theta, p_+) - s_\theta$. In the multiclass case, maximize by taking $b = \min(s_\theta, p_+)$, $h = 0$, and allocating any remaining switches to n. These expressions give Eq. 6. The feasible polytope is convex and compact, so all intermediate target values are attained. □

The condition $g < 1$ ensures the denominator in the switch-fraction expression is positive in every compatible distribution. If $g = 1$ and initial disagreement has positive probability, its conditional switching probability is always one, regardless of the positive switching mass. Equation 5 cannot be inverted in this case. If initial disagreement has probability zero, the switch fraction is undefined.

An interval for the switch fraction. Suppose $0 \le \theta_L \le \theta_U \le 1$ and $g < 1$. The information is compatible if and only if $\theta_L \le \min(\theta_U, g)$. When compatible, define

$$\theta_U^* = \min(\theta_U, g), \qquad s_L = \frac{\theta_L(1-g)}{1-\theta_L}, \qquad s_U = \frac{\theta_U^*(1-g)}{1-\theta_U^*}.$$

Replace the equality on the total switch mass in Eq. 11 by $s_L \le b + h + n \le s_U$. Its sharp target interval is $[L, U]$, where

$$\begin{aligned} L &= \max\{-p_-,\ -s_U,\ s_L - 2p_-\}, \\ U &= \begin{cases} \min\{p_+,\ s_U,\ 2p_+ - s_L\}, & |\mathbf{D}| = 2, \\ \min\{p_+,\ s_U\}, & |\mathbf{D}| \ge 3. \end{cases} \end{aligned} \tag{12}$$

Optimizing $b-h$ over the modified polytope gives these endpoints; the preceding cellwise construction and convexity establish sharpness. The full feasible range of the switch fraction $[0, g]$ recovers $[-p_-, p_+]$. A finite-data estimate of the switch fraction requires an uncertainty range rather than an exact population restriction.

The following result isolates bounded selection before imposing the propensity restriction. For $0 < g < 1$, let $p_{\text{disagree}} = \Pr(\hat{y} = y \mid \hat{y} \ne \hat{y}_{\text{a}})$ and $\Delta_{\text{equal}} = p_+ - g p_{\text{disagree}}$, and assume $|q_{\text{agree}} - p_{\text{disagree}}| \le \kappa$.

Proposition 4 (Sharp bounds under bounded selection)**.** *If $0 < g < 1$, then*

$$\mathbf{I}_\Delta(\kappa) = [-p_-, p_+] \cap [\Delta_{\text{equal}} - g\kappa,\ \Delta_{\text{equal}} + g\kappa] \tag{13}$$

is sharp. At $\kappa = 0$, $\Delta = \Delta_{\text{equal}}$; at $\kappa = 1$, the worst-case region is recovered. If $\Delta_{\text{equal}} \ne 0$, the sign is identified whenever

$$\kappa < \frac{|\Delta_{\text{equal}}|}{g}.$$

Equation 13 gives a one-dimensional sensitivity analysis: as κ increases, the constraint weakens and the identified set expands. For nonzero Δ_{equal}, the threshold $|\Delta_{\text{equal}}|/g$ is the supremum of deviations from equal pre-advice accuracy that preserve a strict sign conclusion under this selection constraint alone. At equality the interval includes zero. This follows bounded conditional-dependence sensitivity and breakdown analysis (Masten & Poirier, 2018; 2020). The joint propensity-and-selection result in Eq. 7 may identify a sign even when this selection-only criterion does not.

Bounded selection, alone and with propensity. Since $q_{\text{disagree}} = p_{\text{disagree}}$ and $q_{\text{agree}} = (p_+ - \Delta)/g$, selection is equivalent to $|\Delta - \Delta_{\text{equal}}| \le g\kappa$. Intersecting this band with the baseline set proves Proposition 4; intersecting it with the propensity set proves Eq. 7, including interval-valued propensity. Every retained target value has a construction from the corresponding sharp set, and that construction automatically satisfies selection because the constraint depends only on Δ. An empty intersection is incompatible.

A cumulative illustration at a fixed observed distribution. For a constructed binary example, let $(p_+, p_-) = (0.60, 0.20)$ and $\theta = 0.60$. Then $s_\theta = 0.30$ and the baseline interval $[-0.20, 0.60]$ contracts to $[-0.10, 0.30]$. One compatible distribution assigns 0.30 to beneficial switches and none to harmful switches; another assigns 0.10 to beneficial and 0.20 to harmful switches. They have identical final-only distributions and the same average adoption propensity, but gains of 0.30 and -0.10.

For the same observed distribution, take $p_{\text{disagree}} = 0.50$, so $\Delta_{\text{equal}} = 0.20$. Retaining $\theta = 0.60$ and adding $\kappa = 0.125$ narrows $[-0.10, 0.30]$ to $[0.10, 0.30]$, identifying a positive gain. This second contraction comes from specifying the relationship between correctness and agreement while keeping the same marginal propensity. The examples in Appendices A.4 and A.5 separately isolate model components; they are not further numerical steps for this observed distribution.

A marginal propensity versus a homogeneous policy. Knowing Eq. 4 does not assert that every kind of initially disagreeing trial has the same switching probability. The latter assumption is substantially

stronger. If that conditional probability equals a constant $a < 1$ at every initially disagreeing $(\hat{y}_{\mathrm{h}}, \hat{y}_{\mathrm{a}}, y, z)$, then

$$p_{\mathrm{final}} = (1-a)q_{\mathrm{pre}} + ap_{\mathrm{AI}}, \qquad \Delta = \frac{a}{1-a}(p_{\mathrm{AI}} - p_{\mathrm{final}}),$$

where $p_{\mathrm{AI}} = \Pr(\hat{y}_{\mathrm{a}} = y)$. This model can identify the gain when its constant probability is known and the observable law is compatible. It cannot be substituted for marginal propensity information without adding the homogeneity assumption. Similarly, a marginal confidence distribution becomes identifying only through restrictions connecting it to judgments or behavior. For example, if $|\Pr(\hat{y}_{\mathrm{h}} = y \mid C = c) - c| \leq \varepsilon$ almost surely, then

$$q_{\mathrm{pre}} \in [\max(0, \mathbb{E}C - \varepsilon),\ \min(1, \mathbb{E}C + \varepsilon)].$$

These are valid bounds from a calibration constraint; they are not asserted to be sharp jointly with other behavioral information. At $\varepsilon = 0$, exact mean confidence supplies exact pre-advice accuracy. Without any such connection, assigning confidence independently of each compatible joint distribution preserves its marginal without excluding any target value.

A.3 Observational equivalence and Bayesian updating

Richer observed records need not distinguish latent judgments when no constraint links the added variables to the missing pre-advice state.

Corollary 1 (AI-side variables without information about the pre-advice state)**.** *Let v be any additional observed AI-side variable. If no constraint links v to $\hat{y}_{\mathrm{h}}$, then observing v does not shrink $\mathbf{I}_\Delta$.*

The endpoint constructions of Proposition 1 apply within every observed (o, v) cell, preserving its probability; mixtures still fill the baseline interval.

Scores, explanations, or metadata help identify this target only when they restrict the missing human judgment. The same distinction underlies the marginal-confidence and calibration results in Appendix A.2.

Proposition 5 (Observational equivalence of behavioral models)**.** *Let $\lambda \in \mathbf{\Lambda}$ index behavioral models with observed-data distribution p_o^λ and latent target $\theta(\lambda)$. If $p_o^{\lambda_0} = p_o^{\lambda_1}$ but $\theta(\lambda_0) \neq \theta(\lambda_1)$, then no selector based only on $o_{1:n}$ can distinguish the two models with worst-case error below $1/2$ for any n, and every estimator $\widehat{\theta}_n$ has worst-case absolute error at least $|\theta(\lambda_1) - \theta(\lambda_0)|/2$ on the pair.*

Under the common sampling law, the two selection errors sum to one, and the triangle inequality makes the sum of the two expected absolute errors at least the target separation. Taking the larger error gives both claims.

In Bayesian partially identified models, data need not update prior mass among joint distributions that induce the same observed-data distribution (Moon & Schorfheide, 2012; Giacomini & Kitagawa, 2021). The following result specializes Proposition 5 to a fixed observed-data distribution.

Proposition 6 (No posterior updating within an observational equivalence class)**.** *Let $q_{\theta,\lambda}$ index compatible joint distributions, where θ fixes p_o and λ indexes joint distributions sharing that observed-data distribution. For any prior and final-only sample,*

$$\pi(d\lambda \mid \theta, o_{1:n}) = \pi(d\lambda \mid \theta).$$

Thus posterior odds among joint distributions that induce the same observed-data distribution equal their prior odds. If Δ varies with λ, final-only data cannot learn how prior mass is allocated within the identified set.

For fixed θ, the observable likelihood is constant in λ and cancels from the normalization in Bayes' rule.

A.4 Identification under the cognitive model

Let $\mathbf{M}_1$ contain compatible distributions with the specified marginal propensity, $\mathbf{M}_2 \subseteq \mathbf{M}_1$ also satisfy bounded selection, and $\mathbf{M}_3 \subseteq \mathbf{M}_2$ additionally belong to the chosen model family. Their images under the same target obey

$$\mathbf{I}_{\Delta,3} \subseteq \mathbf{I}_{\Delta,2} \subseteq \mathbf{I}_{\Delta,1}$$

whenever the classes are nonempty. A richer model need not give a strict reduction, and incompatible assumptions reject a class. Model-component illustrations that omit earlier restrictions are separate analyses and do not represent successive contractions along this hierarchy.

Structural specification and compatible parameters. Section 3.3 presents a representative cognitive specification as constraints on a compatible joint distribution; the parameterization below follows Tejeda et al. (2022). We condition on a finite or countable context z that fixes person i, item j, classifier k, AI confidence c, truth y, and recommendation $\hat{y}_{\mathrm{a}}$. The distribution of contexts is held at its observed distribution in the target workflow. Write ψ for all item-response, confidence, and policy parameters, and let $m = |\mathbf{D}|$. The pre-advice correctness probability is

$$q_\psi(z) = \mathrm{logit}^{-1}(s_j b_i - d_j),$$

where b_i is ability, d_j difficulty, and $s_j > 0$ discrimination. The pre-advice label distribution is

$$\pi_\psi(h \mid z) = \begin{cases} q_\psi(z), & h = y, \\ [1 - q_\psi(z)]/(m-1), & h \neq y. \end{cases} \tag{14}$$

For three confidence levels, the ordered-probit probabilities can be written

$$\mu_\psi(r \mid z) = \Phi\left(\frac{v_{i,r} - q_\psi(z)}{\sigma_i}\right) - \Phi\left(\frac{v_{i,r-1} - q_\psi(z)}{\sigma_i}\right), \qquad r = 1, 2, 3,$$

where $v_{i,0} = -\infty$, $v_{i,3} = \infty$, $v_{i,1} < v_{i,2}$, and $\sigma_i > 0$. Conditional on z and ψ, the pre-advice label and confidence are sampled independently. The policy $\alpha_\psi(i, r, c, k)$ is a probability indexed by person, human confidence, a bin of AI confidence, and classifier. Population shrinkage priors on these parameters are distinct from structural constraints on their allowed values.

Write $\bar{\alpha}_\psi(z) = \sum_r \mu_\psi(r \mid z)\alpha_\psi(i, r, c, k)$ for the switching probability averaged over confidence. For a given pre-advice label h and confidence r, the complete final-label kernel is

$$k_\psi(f \mid h, r, z) = [1 - \alpha_\psi(i, r, c, k)]\mathbf{1}\{f = h\} + \alpha_\psi(i, r, c, k)\mathbf{1}\{f = \hat{y}_{\mathrm{a}}\}.$$

When $h = \hat{y}_{\mathrm{a}}$, both terms select the same label, so their probabilities sum to one. Summing $\pi_\psi(h \mid z)\mu_\psi(r \mid z)k_\psi(f \mid h, r, z)$ over h, r yields Eq. 8. In particular,

$$\begin{aligned} p_\psi(\hat{y} = y \mid z) &= [1 - \bar{\alpha}_\psi(z)]q_\psi(z) + \bar{\alpha}_\psi(z)\mathbf{1}\{\hat{y}_{\mathrm{a}} = y\}, \\ \Delta_\psi(z) &= \bar{\alpha}_\psi(z)[\mathbf{1}\{\hat{y}_{\mathrm{a}} = y\} - q_\psi(z)]. \end{aligned} \tag{15}$$

This simplification uses independence of the pre-advice label and confidence at fixed context and parameters. It need not hold after pooling heterogeneous contexts or averaging parameter uncertainty as if it were known.

The pooled adoption propensity induced by this mechanism is

$$\theta_\psi = \frac{\mathbb{E}_z\{[1 - \pi_\psi(\hat{y}_{\mathrm{a}} \mid z)]\bar{\alpha}_\psi(z)\}}{\mathbb{E}_z[1 - \pi_\psi(\hat{y}_{\mathrm{a}} \mid z)]}. \tag{16}$$

Its denominator is positive for any compatible model when $g < 1$. Equating this quantity to a supplied θ retains the marginal information without imposing $\bar{\alpha}_\psi(z) = \theta$ at every context.

Let $\mathbf{\Psi}_{\mathrm{aux}}$ encode any constraints obtained from auxiliary observations together with explicitly maintained links to the target workflow, including any supplied marginal confidence information. If no auxiliary data are used, impose no such constraint. In particular, using common person–item and confidence parameters across workflows is an assumption, not a consequence of observing a no-assistance arm. For the cumulative comparison with $0 < g < 1$, use the pooled propensity θ_ψ in Eq. 16 and define

$$q_{\mathrm{agree},\psi} = \frac{\mathbb{E}_z\{q_\psi(z)[\bar{\alpha}_\psi(z) + (1 - \bar{\alpha}_\psi(z))\mathbf{1}\{\hat{y}_{\mathrm{a}} = y\}]\}}{g}.$$

The parameter-compatible class is

$$\mathbf{\Psi}_{\mathrm{obs}} = \left\{ \psi \in \mathbf{\Psi}_{\mathrm{aux}} : \begin{array}{l} p_\psi(f \mid z) = p_{\mathrm{obs}}(f \mid z) \text{ for all } f, z, \\ \theta_\psi = \theta, \quad |q_{\mathrm{agree},\psi} - p_{\mathrm{disagree}}| \leq \kappa \end{array} \right\}.$$

Observable equality is required for every label and every context of positive probability; the propensity and selection restrictions above are pooled. The propensity weights each conditional switching probability by the probability of initial disagreement in that context. It is generally not the unweighted mean of $\bar{\alpha}_\psi(z)$, and does not require every context to have the same propensity. Propensity intervals replace equality by interval membership. For a model-only analysis, omit the marginal and selection constraints; the cellwise component results below isolate that case. With unknown parameters, observational matching is imposed after marginalizing the latent pre-advice label and confidence, not after replacing either by a fitted prediction.

Model-specific identified set. Write $p_{\text{final}} = \Pr(\hat{y} = y)$. For a nonempty compatible parameter class, the sharp set is

$$\mathbf{I}_\Delta^{\mathrm{T}} = \{p_{\text{final}} - \mathbb{E}_z q_\psi(z) : \psi \in \mathbf{\Psi}_{\text{obs}}\}. \tag{17}$$

Each compatible ψ generates a joint distribution by sampling z, then $(\hat{y}_{\text{h}}, r)$, then $\hat{y}$ from k_ψ; conversely, every distribution in the specified family has such a parameterization. This establishes Eq. 17. Its width is the range of $\mathbb{E}_z q_\psi(z)$ over compatible parameters. The image need not be an interval or attain its endpoints, since the parametric family need not be closed under mixtures. Point identification requires agreement on mean pre-advice accuracy, not identification of every cognitive parameter; an empty class is incompatible.

In the cumulative comparison, the retained propensity and selection restrictions imply containment in $\mathbf{I}_\Delta(\theta, \kappa)$. Adding constraints can only reduce the set, but need not reduce its width strictly. Direct accuracy information from Appendix B.1 can also be imposed; an exact target-population mean already fixes aggregate Δ. These statements concern population identification. A posterior credible interval or a finite-sample confidence region for parameters requires a separate inferential interpretation.

The following result isolates uniform errors and label-independent adoption before imposing cross-context parameter sharing or pooled marginal information.

Proposition 7 (Identification from final decisions that differ from AI advice)**.** *Fix a context z with $\hat{y}_{\text{a}} \neq y$, $m \geq 3$, and an observed final-label distribution compatible with Eqs. 14 and 8. Write*

$$u = \Pr(\hat{y} = y \mid z), \qquad v = \Pr(\hat{y} \notin \{y, \hat{y}_{\text{a}}\} \mid z).$$

If $u + v > 0$, Eq. 9 identifies retention, pre-advice accuracy, adoption, and gain. If instead $\hat{y}_{\text{a}} = y$, the cellwise model with free correctness and adoption probabilities in $[0, 1]$ leaves the sharp range $\Delta(z) \in [0, g(z)]$ for a compatible final-label distribution.

Proof of Proposition 7. Fix an incorrect-advice context and write $q = q_\psi(z)$ and $t = 1 - \bar{\alpha}_\psi(z)$. The true label and the $m - 2$ labels other than truth and advice can appear finally only when retained. Uniform initial errors and label-independent retention therefore give

$$u = tq, \qquad v = t(1 - q)\frac{m - 2}{m - 1}.$$

Adding the first equation to $(m - 1)/(m - 2)$ times the second identifies t; $u + v > 0$ ensures $t > 0$, identifying q and hence Δ. Compatibility also requires $t \leq 1$ and equal observed probabilities for the $m - 2$ third labels. A binary task has no third label and does not admit this inversion. If $u = v = 0$, $t = 0$ for $m \geq 3$, and the cell alone does not restrict q in the closed probability model.

For a correct-advice context, $g = q + (1 - q)\bar{\alpha}$. For $g < 1$, every $q \in [0, g]$ is attained by $\bar{\alpha} = (g - q)/(1 - q)$, reproducing the full compatible final-label distribution because all wrong labels have probability $(1 - g)/(m - 1)$. Thus $\Delta = g - q$ fills $[0, g]$. For $g = 1$, choose $\bar{\alpha} = 1$ and any $q \in [0, 1]$. This establishes the cellwise sharpness statement with boundary probabilities allowed. Item-response sharing across contexts can remove some of these conditional distributions, so one cannot independently combine local endpoints to assert sharp aggregate bounds for the full cognitive model.

Relaxing the distribution of pre-advice errors. The next results hold in a single incorrect-advice context with $m \geq 3$, $u > 0$, $v > 0$, and $g = 1 - u - v \geq 0$. They leave the initial allocation across third labels unrestricted and omit cross-context item-response and confidence constraints. This isolates two components of the model. Let

$$\rho = \Pr(\hat{y}_{\text{h}} = \hat{y}_{\text{a}} \mid \hat{y}_{\text{h}} \neq y, z)$$

be the share of initial errors that already match the incorrect advice. Uniform errors set $\rho = 1/(m-1)$.

Proposition 8 (Sensitivity to the distribution of pre-advice errors)**.** *Suppose retention is independent of the pre-advice label at this context and* $0 \le \rho_L \le \rho \le \rho_U \le 1$*. Define*

$$\rho_U^* = \min\left\{\rho_U,\, 1 - \frac{v}{1-u}\right\}, \qquad Q(\rho) = \frac{u(1-\rho)}{u(1-\rho)+v}.$$

The constraints are compatible if and only if $\rho_L \le \rho_U^*$*. When compatible, the sharp interval is*

$$\mathbf{I}_\Delta^{\mathrm{err}} = [u - Q(\rho_L),\ u - Q(\rho_U^*)], \tag{18}$$

whose width is $Q(\rho_L) - Q(\rho_U^*)$*. In particular, setting* $\rho_L = \rho_U = 1/(m-1)$ *recovers Eq. 9 whenever compatible.*

Proof. Writing t for the common retention probability gives $u = tq$ and $v = t(1-q)(1-\rho)$. Hence $t = u + v/(1-\rho)$ and $q = Q(\rho)$. The constraint $t \le 1$ is exactly $\rho \le 1 - v/(1-u)$; the remaining probability constraints follow from $u, v > 0$. For any feasible ρ, choose initial masses q on truth, $(1-q)\rho$ on advice, and $(1-q)(1-\rho)$ on the third labels, allocated in proportion to their observed final probabilities. Common retention t reproduces each non-AI final-label probability, and normalization reproduces g. This constructs every feasible joint distribution. Since Q is continuous and decreasing on the feasible interval, Eq. 18 is sharp. □

Relaxing label-independent retention. Keep $\rho \in [0, 1)$ fixed but define separate retention probabilities

$$t_y = \Pr(\hat{y} = \hat{y}_{\mathrm{h}} \mid \hat{y}_{\mathrm{h}} = y, z),$$
$$t_o = \Pr(\hat{y} = \hat{y}_{\mathrm{h}} \mid \hat{y}_{\mathrm{h}} \notin \{y, \hat{y}_{\mathrm{a}}\}, z).$$

Instead of $t_y = t_o$, assume $\lambda^{-1} \le t_y/t_o \le \lambda$ for $\lambda \ge 1$. Positive u, v imply that both retention probabilities are positive. This relaxation allows selection on pre-advice correctness that the factorized cognitive specification excludes at fixed context.

Proposition 9 (Sensitivity to label-dependent retention)**.** *For fixed* $\rho \in [0, 1)$ *and* $\lambda \ge 1$*, let*

$$Q_\rho(b) = \frac{u(1-\rho)}{u(1-\rho)+bv},$$
$$q_L = \max\{u, Q_\rho(\lambda)\}, \qquad q_U = \min\{1 - v/(1-\rho), Q_\rho(1/\lambda)\}.$$

The constraints are compatible if and only if $q_L \le q_U$*. When compatible, the sharp interval is*

$$\mathbf{I}_\Delta^{\mathrm{ret}} = [u - q_U,\ u - q_L], \tag{19}$$

with width $q_U - q_L$*. At* $\lambda = 1$ *it reduces to a singleton whenever compatible; as* $\lambda \to \infty$*, the sharp correctness range approaches* $[u, 1 - v/(1-\rho)]$*.*

Proof. The observable probabilities satisfy $u = qt_y$ and $v = (1-q)(1-\rho)t_o$. Consequently $q = Q_\rho(b)$ for $b = t_y/t_o$. This function decreases continuously in b, yielding $Q_\rho(\lambda) \le q \le Q_\rho(1/\lambda)$. The constraints $t_y, t_o \le 1$ give $u \le q \le 1 - v/(1-\rho)$. Conversely, for every q in this intersection choose $t_y = u/q$ and $t_o = v/[(1-q)(1-\rho)]$. Allocate third-label initial mass in proportion to its observed final probabilities and give each such label retention probability t_o. These choices reproduce every final-label probability and satisfy the ratio bound. Mapping q to $u - q$ proves sharpness and the limiting statement. □

For a constructed three-class context with incorrect AI advice and final probabilities $(u, g, v) = (0.30, 0.55, 0.15)$, the final-only interval is $[-0.55, 0]$. Uniform errors and label-independent adoption give $q = 0.50$, $\bar{\alpha} = 0.40$, and $\Delta = -0.20$. Allowing $\rho \in [0.40, 0.60]$ while retaining label-independent adoption expands the sharp interval to approximately $[-0.245, -0.144]$. Alternatively, keeping $\rho = 0.50$ and allowing a retention ratio in $[1/2, 2]$ gives approximately $[-0.367, -0.033]$. These are analytic illustrations, not estimates from the Tejeda dataset. They attribute contraction to specific constraints rather than to model complexity or posterior concentration. The analysis maintains keep-or-adopt; departures such as those observed in the sequential data require a separate support relaxation (Appendix Theorem 3).

These are distinct relaxations, not successive numerical estimates. An error-allocation interval containing $1/(m-1)$ enlarges the uniform-error component; a retention ratio range containing one enlarges the label-independence component. Their explicit widths measure the sensitivity of the cellwise result. With both components unrestricted, take $\rho = 0$ and permit arbitrary positive t_y, t_o. Then every $q \in [u, 1-v]$ is feasible, recovering $\Delta \in [-g, 0]$, exactly the incorrect-advice baseline. Additional auxiliary information or parameter sharing must be imposed on the same compatible joint distributions; intersecting separately projected ranges alone need not be sharp when those constraints also concern other latent features. For the full model, removing the item-response or confidence constraints likewise enlarges the parameter-compatible family, but the resulting global width must be obtained from Eq. 17, not by pooling independently optimized local endpoints.

A.5 Identification under activation and integration

Let $r_{\mathrm{h}}, r_{\mathrm{f}} \in [-1, 1]$ encode a binary judgment and its confidence, and let z fix the observed context, including the advice and truth. Represent activation by $a_\psi(r_{\mathrm{h}}, z)$ and integration by a conditional response distribution $B_\psi(dr_{\mathrm{f}} \mid r_{\mathrm{h}}, z)$. The final-response kernel is

$$K_\psi(dr_{\mathrm{f}} \mid r_{\mathrm{h}}, z) = [1 - a_\psi(r_{\mathrm{h}}, z)]\delta_{r_{\mathrm{h}}}(dr_{\mathrm{f}}) + a_\psi(r_{\mathrm{h}}, z)B_\psi(dr_{\mathrm{f}} \mid r_{\mathrm{h}}, z). \tag{20}$$

Inactivity leaves the response unchanged. Activation restricts which pre-advice states tend to change, while integration restricts where they can move. The distributional extension and identification analysis are developed here.

In Vodrahalli et al. (2022b), activation and integration are learned from paired pre-advice and final responses, with pre-advice confidence, advice confidence, agreement, and participant characteristics as inputs. Their integration network predicts a response change; interpreting that prediction as an exact response is a deterministic special case of our distributional specification. A conditional mean prediction alone does not specify B_ψ or its residual distribution. Likewise, a high predictive score on paired data does not establish invertibility from final-only data. The unchanged-response branch idealizes inactivity; using the original operational threshold for activation requires handling small response changes consistently in the observation model. Learning these components from paired data also requires coverage of the pre-advice states relevant to the target workflow.

Fix a context z that includes the truth and the advice, and a finite response scale $\mathbf{R} = \{r_1, \ldots, r_d\}$. Define $h_j = \Pr(r_{\mathrm{h}} = r_j \mid z)$, $p_k = \Pr(r_{\mathrm{f}} = r_k \mid z)$, $a_j = \Pr(\text{activation} \mid r_{\mathrm{h}} = r_j, z)$, and $B_{kj} = \Pr(r_{\mathrm{f}} = r_k \mid r_{\mathrm{h}} = r_j, z, \text{activation})$. Each column of B sums to one. The resulting transition matrix and marginal relation are given in Eq. 10. The response-to-label rule, including any tie convention, is fixed in advance; c_j indicates whether r_j predicts the true label. These are exact finite-response models. An arbitrary binning of continuous responses does not preserve a fixed K: its columns can depend on the unobserved within-bin distribution of pre-advice responses.

To bound accuracy change, consider all pre-advice distributions compatible with the final responses, $\mathbf{H}_K(p) = \{h \geq 0 : \mathbf{1}^\top h = 1, Kh = p\}$. When this set is nonempty, the sharp interval is

$$\mathbf{I}_\Delta(K, p) = \left[c^\top p - \max_{h \in \mathbf{H}_K(p)} c^\top h,\ c^\top p - \min_{h \in \mathbf{H}_K(p)} c^\top h \right]. \tag{21}$$

Proposition 10 (Identification of accuracy under a known response kernel)**.** *Let K be a known column-stochastic matrix and $\mathbf{H}_K(p)$ be nonempty. Equation 21 is a sharp interval with attained endpoints. It is a singleton if and only if*

$$c^\top(h - h') = 0 \qquad \textit{for all } h, h' \in \mathbf{H}_K(p).$$

In particular, $c \in \mathrm{row}(K)$ is sufficient: if $K^\top b = c$, the gain is $\Delta(z) = c^\top p - b^\top p$. If $\mathbf{H}_K(p)$ contains an h^* whose coordinates are all strictly positive, this row-space condition is also necessary. Full column rank of K identifies h itself and hence every such accuracy target.

Proof. Every feasible h defines a joint distribution with mass $h_j K_{kj}$ on $(r_{\mathrm{h}}, r_{\mathrm{f}}) = (r_j, r_k)$, reproducing p and giving $\Delta = c^\top p - c^\top h$. Conversely, every joint distribution with this conditional kernel and final marginal supplies a feasible h. The feasible set is compact and convex, so the linear target attains both endpoints and every intermediate value. Constancy of $c^\top h$ gives the stated necessary and sufficient condition. If $K^\top b = c$, then $c^\top h = b^\top K h = b^\top p$. For the converse with positive h^*, if

$c \notin \mathrm{row}(K)$ there exists $v \in \ker(K)$ with $c^\top v \neq 0$. Column stochasticity gives $\mathbf{1}^\top v = \mathbf{1}^\top K v = 0$. For sufficiently small $t > 0$, both $h^* + tv$ and $h^* - tv$ are feasible probability vectors, and their accuracies differ. Full column rank makes the kernel trivial. □

For example, a mechanism that maps both incorrect response states to one incorrect state but preserves a correct state has identical first two columns. It cannot identify their separate pre-advice masses, yet it identifies their total and thus pre-advice accuracy. Conversely, mapping all states to the advice erases every distinction between their accuracies. On the boundary of the simplex, a particular observed p may identify accuracy even without the row-space condition; the exact constancy criterion still applies. The same linear argument handles a bounded response loss by replacing correctness with its response-specific loss vector and using the sign convention for Δ_ℓ.

When component parameters are unknown, retain every pair (ψ, h) satisfying $K_\psi h = p$ and any auxiliary constraints. The union of their target values need not be an interval. Across contexts, parameters shared by neural networks or other structures must be shared in the compatible class as well. Averaging separately optimized cellwise bounds is sharp only when those cells' constraints can be satisfied independently. An empty compatible class rejects the maintained model; it does not identify a target. These are population statements. Estimating the observed-data distribution and the components from finite samples requires separate uncertainty assessment, including any assumptions transporting a kernel learned from paired observations to a final-only workflow.

Componentwise uncertainty as a linear program. Fix reference activation probabilities a_j^0 and reference integration distributions $B_{\cdot j}^0$. Allow $a_j \in [a_j^-, a_j^+]$ and

$$\mathrm{TV}(B_{\cdot j}, B_{\cdot j}^0) = \tfrac{1}{2} \sum_k |B_{kj} - B_{kj}^0| \leq \varepsilon_{I,j}.$$

For an activation tolerance ε_A, take $a_j^- = \max(0, a_j^0 - \varepsilon_A)$ and $a_j^+ = \min(1, a_j^0 + \varepsilon_A)$, except for any probabilities explicitly held fixed. Let $\mathbf{S}_j$ be a nonempty set of permitted activated destinations containing the support of $B_{\cdot j}^0$. For convex movement, these destinations lie between r_j and the advice; no such constraint is inherent in the general activation–integration model.

Use nonnegative variables s_j for inactive mass initially at r_j and v_{kj} for activated mass moving from r_j to r_k. Write $t_j = \sum_k v_{kj}$ and $h_j = s_j + t_j$. Let $\mathbf{F}(p)$ be the feasible set specified by

$$\begin{aligned} p_k &= s_k + \sum_j v_{kj}, \qquad \sum_j h_j = 1, \\ a_j^- h_j &\leq t_j \leq a_j^+ h_j, \qquad v_{kj} = 0 \quad (k \notin \mathbf{S}_j), \\ \sum_k |v_{kj} - B_{kj}^0 t_j| &\leq 2\varepsilon_{I,j} t_j \qquad \text{for every } j. \end{aligned} \tag{22}$$

The absolute-value constraints are linearized with nonnegative auxiliary variables bounding each positive and negative residual.

Proposition 11 (Sharp component-relaxation bounds)**.** *Suppose $\mathbf{F}(p)$ is nonempty and no additional cross-column or cross-context parameter constraints are imposed. The sharp interval over the stated componentwise class is*

$$\begin{aligned} L &= \min_{(s,v)\in\mathbf{F}(p)} c^\top (s+t), \qquad U = \max_{(s,v)\in\mathbf{F}(p)} c^\top (s+t), \\ \mathbf{I}_\Delta^{\mathrm{AI}}(z) &= [c^\top p - U,\ c^\top p - L]. \end{aligned} \tag{23}$$

Both endpoints are obtained by linear programming. Enlarging any activation interval or integration tolerance can only enlarge this identified interval. Setting all component tolerances to zero recovers Eq. 21 for the reference kernel.

Proof. Given any compatible model, set $s_j = (1 - a_j)h_j$ and $v_{kj} = a_j h_j B_{kj}$. These masses satisfy Eq. 22. Conversely, for any feasible masses, take $a_j = t_j/h_j$ when $h_j > 0$, and $B_{kj} = v_{kj}/t_j$ when $t_j > 0$. At zero-mass states choose any admissible a_j; when $t_j = 0$, choose $B_{\cdot j} = B_{\cdot j}^0$. These choices generate exactly the feasible masses and satisfy the component bounds. The nonnegative masses are bounded by one, and their feasible set is closed and convex. Linear optimization therefore

yields attained sharp endpoints and mixtures fill the interval. Enlarging tolerances only removes constraints. With zero tolerances, $t_j = a_j^0 h_j$ and $v_{kj} = B_{kj}^0 t_j$, recovering $p = K^0 h$. □

This componentwise class need not preserve the parameter sharing of a fitted neural network; it is a specified relaxation whose bounds are sharp for that class. Any retained linear constraints on pre-advice accuracy can be added to the programs. If all component constraints and movement constraints are removed, every joint distribution of pre-advice and final responses is feasible, recovering the unrestricted interval $[c^\top p - \max_j c_j, c^\top p - \min_j c_j]$. If convex or keep-or-adopt support is retained, all resulting bounds stay within the corresponding support-only set.

Retaining marginal propensity and selection information. The cumulative comparison can be implemented in the same programs. For a finite collection of contexts with observed weights w_z, let $\ell_j(z)$ be the class label associated with response r_j, let $A(z)$ be the AI label, and write

$$m_{kj}(z) = \mathbf{1}\{k = j\} s_j(z) + v_{kj}(z)$$

for the complete joint response mass in context z. Retain keep-or-adopt at the label level by setting $m_{kj}(z) = 0$ whenever $\ell_k(z) \notin \{\ell_j(z), A(z)\}$. Define the initial-disagreement and adoption masses

$$D = \sum_z w_z \sum_{j:\ell_j(z)\neq A(z)} h_j(z),$$

$$T = \sum_z w_z \sum_{j:\ell_j(z)\neq A(z)} \sum_{k:\ell_k(z)=A(z)} m_{kj}(z).$$

For $g < 1$, $D \geq 1 - g > 0$. The pooled propensity constraint is the linear equality $T = \theta D$, or the inequalities $\theta_L D \leq T \leq \theta_U D$ for an interval. This constrains adoption of the AI label, not the total activation mass: changing confidence without changing the label need not contribute to T.

For $0 < g < 1$, put $c_j(z) = \mathbf{1}\{\ell_j(z) = y(z)\}$ and define

$$Q_G = \sum_z w_z \sum_j \sum_{k:\ell_k(z)=A(z)} c_j(z) m_{kj}(z).$$

Then pooled bounded selection is the pair of linear inequalities

$$g(p_{\text{disagree}} - \kappa) \leq Q_G \leq g(p_{\text{disagree}} + \kappa).$$

The proof of Proposition 11 continues to apply after these linear constraints are added, so the resulting bounds are sharp for this componentwise class and lie inside $\mathbf{I}_\Delta(\theta, \kappa)$. The aggregate objective is $\sum_z w_z c(z)^\top [p(z) - h(z)]$. A known kernel uses $m_{kj}(z) = K_{kj}(z) h_j(z)$ directly and admits the same constraints. Additional parameter sharing can further restrict this class; the linear-program bounds are then outer bounds unless that sharing is also enforced.

A constructed example separating the two stages. Take responses $(-1, -1/3, 1/3, 1)$, positive responses as correct, advice $r_{\text{a}} = 1$, and observed final-response probabilities $p = (0.10, 0.20, 0.20, 0.50)^\top$. Thus $c = (0, 0, 1, 1)^\top$ and final accuracy is 0.70. Maintain convex movement throughout. With that support constraint alone, pre-advice accuracy can be any value from 0 to 0.70, giving $\Delta \in [0, 0.70]$: setting every pre-advice response to -1 attains the upper endpoint, while taking it equal to the final response attains the lower endpoint.

For the reference model, set $a_j^0 = 1/2$ for $j = 1, 2, 3$ and $a_4 = 0$. Activated responses move to the next grid point: $B_{j+1,j}^0 = 1$ for $j < 4$; set $B_{44}^0 = 1$ for the unused fourth column. The resulting kernel is

$$K^0 = \begin{pmatrix} 1/2 & 0 & 0 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/2 & 0 \\ 0 & 0 & 1/2 & 1 \end{pmatrix}.$$

It is invertible and gives $h = (0.20, 0.20, 0.20, 0.40)^\top$, hence $\Delta = 0.10$. In the component relaxations, keep $a_4 = 0$ and require activated movement to a strictly higher response for $j < 4$, i.e., $\mathbf{S}_j = \{j + 1, \ldots, 4\}$, with $\mathbf{S}_4 = \{4\}$. Activation intervals apply to the first three states; integration tolerances apply to their allowed destination distributions.

Table 2 evaluates the linear programs in Proposition 11 at these fixed inputs. The accompanying numerical verification checks the endpoints by independent vertex enumeration and exact rational witnesses. The example measures sensitivity to individual component restrictions; its tolerances are not estimated from human data.

Table 2: **Component relaxation at a fixed final-response distribution.** Constructed example; all rows retain convex movement. Tolerances are probabilities, not estimated uncertainty levels.

Additional component information	ε_A	ε_I	Sharp $\mathbf{I}_\Delta$
None (convex support only)	—	—	$[0, 7/10]$
Exact activation and integration	0	0	$\{1/10\}$
Relax activation	$1/10$	0	$[1/30, 1/5]$
Relax integration	0	$1/5$	$[1/10, 7/50]$
Relax both	$1/10$	$1/5$	$[1/30, 7/30]$

A.6 CONTINUOUS JUDGMENTS

Let $\hat{y}_{\mathrm{h}}, \hat{y}_{\mathrm{a}}, \hat{y}$ lie in a bounded scalar judgment domain $\mathbf{D} = [l, u] \subset \mathbb{R}$, let y be the observed outcome entering the loss, and let $\ell(\hat{y}_{\mathrm{h}}, y)$ be bounded and jointly measurable, continuous in $\hat{y}_{\mathrm{h}}$, with measurable minimizing and maximizing selectors. Define

$$\Delta_\ell := \mathbb{E}[\ell(\hat{y}_{\mathrm{h}}, y) - \ell(\hat{y}, y)].$$

Without behavioral constraints on advice use, define

$$\underline{\ell}_0(y) := \min_{\hat{y}_{\mathrm{h}} \in \mathbf{D}} \ell(\hat{y}_{\mathrm{h}}, y), \qquad \overline{\ell}_0(y) := \max_{\hat{y}_{\mathrm{h}} \in \mathbf{D}} \ell(\hat{y}_{\mathrm{h}}, y).$$

Theorem 1 (Sharp continuous identified set without behavioral constraints on advice use)**.** *If the only constraint on the unobserved pre-advice judgment is $\hat{y}_{\mathrm{h}} \in \mathbf{D}$, then*

$$\mathbf{I}^0_{\Delta_\ell} = \left[\mathbb{E}\{\underline{\ell}_0(y) - \ell(\hat{y}, y)\},\ \mathbb{E}\{\overline{\ell}_0(y) - \ell(\hat{y}, y)\}\right] \tag{24}$$

is sharp, with width

$$\gamma_0 := \mathbb{E}[\overline{\ell}_0(y) - \underline{\ell}_0(y)]. \tag{25}$$

For every n, any estimator based only on $(y, \hat{y}_{\mathrm{a}}, \hat{y})$ has worst-case absolute error at least $\gamma_0/2$ and worst-case squared error at least $\gamma_0^2/4$ over compatible joint distributions.

A weight-of-advice model restricts the final response to a convex combination of the initial judgment and advice (Bailey et al., 2023):

$$\hat{y} = (1 - w)\hat{y}_{\mathrm{h}} + w\hat{y}_{\mathrm{a}}, \qquad w \in [0, 1]. \tag{26}$$

This constraint requires the final response to lie between the initial judgment and advice, and is testable when the initial response is observed. The convex-support activation–integration model additionally restricts transition probabilities, refining the support-only analysis below.

Under Eq. 26, the set of feasible pre-advice judgments given $(\hat{y}_{\mathrm{a}}, \hat{y})$ is

$$\mathbf{H}_{\mathrm{cvx}}(\hat{y}_{\mathrm{a}}, \hat{y}) := \begin{cases} [l, \hat{y}], & \hat{y}_{\mathrm{a}} > \hat{y}, \\ [\hat{y}, u], & \hat{y}_{\mathrm{a}} < \hat{y}, \\ [l, u], & \hat{y}_{\mathrm{a}} = \hat{y}. \end{cases} \tag{27}$$

Define

$$\underline{\ell}_{\mathrm{cvx}}(o) := \min_{\hat{y}_{\mathrm{h}} \in \mathbf{H}_{\mathrm{cvx}}(\hat{y}_{\mathrm{a}}, \hat{y})} \ell(\hat{y}_{\mathrm{h}}, y),$$
$$\overline{\ell}_{\mathrm{cvx}}(o) := \max_{\hat{y}_{\mathrm{h}} \in \mathbf{H}_{\mathrm{cvx}}(\hat{y}_{\mathrm{a}}, \hat{y})} \ell(\hat{y}_{\mathrm{h}}, y).$$

Corollary 2 (Identification under the convex-combination constraint)**.** *Under Eq. 26,*

$$\mathbf{I}^{\mathrm{cvx}}_{\Delta_\ell} = \left[\mathbb{E}\{\underline{\ell}_{\mathrm{cvx}}(o) - \ell(\hat{y}, y)\},\ \mathbb{E}\{\overline{\ell}_{\mathrm{cvx}}(o) - \ell(\hat{y}, y)\}\right] \tag{28}$$

is sharp, with width $\gamma_{\mathrm{cvx}} := \mathbb{E}[\overline{\ell}_{\mathrm{cvx}}(o) - \underline{\ell}_{\mathrm{cvx}}(o)] \le \gamma_0$. Also $\mathbf{I}^{\mathrm{cvx}}_{\Delta_\ell} \subseteq \mathbf{I}^0_{\Delta_\ell}$, with corresponding minimax lower bounds $\gamma_{\mathrm{cvx}}/2$ for absolute error and $\gamma^2_{\mathrm{cvx}}/4$ for squared error.

A.7 General bounded losses

The equality in Proposition 1 is especially simple for 0–1 loss. Let $\ell(\hat{y}_{\mathrm{h}}, y)$ be any bounded loss and define the loss reduction relative to the pre-advice judgment by $\Delta_\ell := \mathbb{E}[\ell(\hat{y}_{\mathrm{h}}, y) - \ell(\hat{y}, y)]$. Under the keep-or-adopt assumption, disagreement cells have $\hat{y}_{\mathrm{h}} = \hat{y}$ and contribute no ambiguity. For an agreement cell $o = (y, \hat{y}_{\mathrm{a}}, \hat{y}_{\mathrm{a}})$ with $\hat{y}_{\mathrm{a}} \in \mathbf{D}$, compatible values of $\hat{y}_{\mathrm{h}}$ generate a loss range $\max_{\hat{y}_{\mathrm{h}} \in \mathbf{D}} \ell(\hat{y}_{\mathrm{h}}, y) - \min_{\hat{y}_{\mathrm{h}} \in \mathbf{D}} \ell(\hat{y}_{\mathrm{h}}, y)$. Choosing compatible conditional distributions independently for each observable record therefore gives an identified set of width

$$\mathbb{E}\left[\mathbf{1}\{\hat{y} = \hat{y}_{\mathrm{a}}\}\left(\max_{\hat{y}_{\mathrm{h}} \in \mathbf{D}} \ell(\hat{y}_{\mathrm{h}}, y) - \min_{\hat{y}_{\mathrm{h}} \in \mathbf{D}} \ell(\hat{y}_{\mathrm{h}}, y)\right)\right].$$

For ordinary classification with 0–1 loss, this per-agreement range is one, recovering $\Pr(\hat{y} = \hat{y}_{\mathrm{a}})$. Continuous or graded response models in which the final response can interpolate between $\hat{y}_{\mathrm{h}}$ and $\hat{y}_{\mathrm{a}}$ require a different support model and are not covered by the discrete adoption analysis.

B Human-side information and sensitivity

Section 3 states that the remaining width equals how often agreement hides $\hat{y}_{\mathrm{h}}$ times the uncertainty left about pre-advice correctness on those trials. The results below spell that product out for a bound on pre-advice accuracy, a measurement from another workflow, a sparse correctness bit, selection into agreement, and departures from keep-or-adopt support.

B.1 Information about pre-advice human accuracy

Let $q_{\mathrm{pre}}(z) := \Pr(\hat{y}_{\mathrm{h}} = y \mid z)$ and $p_{\mathrm{final}}(z) := \Pr(\hat{y} = y \mid z)$, and suppose auxiliary information gives bounds for the same target setting, $\underline{q}(z) \le q_{\mathrm{pre}}(z) \le \overline{q}(z)$. Define

$$l(z) := \max\{-p_-(z),\ p_{\mathrm{final}}(z) - \overline{q}(z)\},$$
$$u(z) := \min\{p_+(z),\ p_{\mathrm{final}}(z) - \underline{q}(z)\}.$$

Theorem 2 (Sharp identification from pre-advice accuracy information)**.** *Suppose z has finite or countable support. If $l(z) \le u(z)$ for every context of positive probability, then*

$$\begin{aligned}\mathbf{I}_\Delta(z) &= [l(z), u(z)] \\ &= [-p_-(z), p_+(z)] \cap [p_{\mathrm{final}}(z) - \overline{q}(z),\ p_{\mathrm{final}}(z) - \underline{q}(z)]\end{aligned} \tag{29}$$

is the sharp identified interval conditional on each z, and the aggregate target is sharply identified as

$$\Delta \in [\mathbb{E}l(z),\ \mathbb{E}u(z)]. \tag{30}$$

If $q_{\mathrm{pre}}(z)$ is known exactly, $\Delta(z) = p_{\mathrm{final}}(z) - q_{\mathrm{pre}}(z)$ is point identified whenever the auxiliary information is compatible with the observed workflow.

B.2 Measurements from another workflow

Use the tags tar and aux for the target and auxiliary workflows, respectively. Write $p^{\mathrm{tar}}_{\mathrm{final}}, p_{+,\mathrm{tar}}, p_{-,\mathrm{tar}}$ for the target-workflow quantities and let $q_{\mathrm{tar}}(z) = \Pr(\hat{y}^{\mathrm{tar}}_{\mathrm{h}} = y \mid z)$ and observed $q_{\mathrm{aux}}(z) = \Pr(\hat{y}^{\mathrm{aux}}_{\mathrm{h}} = y \mid z)$. Suppose $|q_{\mathrm{tar}}(z) - q_{\mathrm{aux}}(z)| \le \varepsilon(z)$, and define $\underline{q}_\varepsilon(z) = \max\{0, q_{\mathrm{aux}}(z) - \varepsilon(z)\}$ and $\overline{q}_\varepsilon(z) = \min\{1, q_{\mathrm{aux}}(z) + \varepsilon(z)\}$.

Corollary 3 (Sharp bounds under cross-workflow deviation)**.** *Let*

$$l_\varepsilon(z) = \max\{-p_{-,\mathrm{tar}}(z), p^{\mathrm{tar}}_{\mathrm{final}}(z) - \overline{q}_\varepsilon(z)\},$$
$$u_\varepsilon(z) = \min\{p_{+,\mathrm{tar}}(z), p^{\mathrm{tar}}_{\mathrm{final}}(z) - \underline{q}_\varepsilon(z)\}.$$

If $l_\varepsilon(z) \le u_\varepsilon(z)$, the conditional and aggregate targets have the sharp bounds

$$\Delta_{\mathrm{tar}}(z) \in [l_\varepsilon(z), u_\varepsilon(z)],$$
$$\Delta_{\mathrm{tar}} \in [\mathbb{E}l_\varepsilon(z), \mathbb{E}u_\varepsilon(z)].$$

Writing $\delta_{\mathrm{aux}}(z) = p^{\mathrm{tar}}_{\mathrm{final}}(z) - q_{\mathrm{aux}}(z)$, the sign is identified exactly when $\varepsilon(z) < |\delta_{\mathrm{aux}}(z)|$.

This is Theorem 2 with the transported accuracy range. Since the baseline interval contains zero, the refined interval contains zero exactly when $|p^{\mathrm{tar}}_{\mathrm{final}}(z) - q_{\mathrm{aux}}(z)| \le \varepsilon(z)$, giving the sign threshold.

B.3 SPARSE PRE-ADVICE MEASUREMENT

Sharp bounds for a partially measured sample. Fix $n \geq 2$ trials with at least two decision labels. Write $b_i = \mathbf{1}\{\hat{y}_{\mathrm{h},i} = y_i\}$, $c_i = \mathbf{1}\{\hat{y}_i = y_i\}$, and $a_i = \mathbf{1}\{\hat{y}_i = \hat{y}_{\mathrm{a},i}\}$. The sample accuracy change is $\Delta_n = n^{-1}\sum_i (c_i - b_i)$. For a measurement pattern $r \in \{0,1\}^n$, let $d(r) = n^{-1}\sum_i r_i(c_i - b_i)$ be the observed contribution. Define the unmeasured correct- and incorrect-agreement fractions as $u_+(r) = n^{-1}\sum_i (1-r_i)a_i c_i$ and $u_-(r) = n^{-1}\sum_i (1-r_i)a_i(1-c_i)$. For compatible measured bits, the sharp lower and upper bounds are

$$\Delta_n \in [d(r) - u_-(r),\ d(r) + u_+(r)], \qquad w(r) = u_-(r) + u_+(r). \tag{31}$$

Indeed, disagreement implies $b_i = c_i$. Each unmeasured correct-agreement trial contributes either zero or $1/n$, and each unmeasured incorrect-agreement trial contributes either $-1/n$ or zero. Choosing compatible initial labels separately on these trials attains both endpoints. The attainable values lie on a $1/n$ grid between them; the interval gives their sharp bounds. No sampling assumption is needed for this statement. If the measured set is a uniform subset of size m, then $\mathbb{E}_r(1 - r_i) = 1 - m/n$, giving Eq. 3. Measuring an additional agreement trial reduces the width by $1/n$; measuring a disagreement trial leaves it unchanged. Measuring all agreement trials therefore suffices to determine Δ_n. This is a statement about which missing bits matter; selecting trials for elicitation before advice cannot use their eventual agreement status.

Random measurement also supports inference about unmeasured trials. For a uniform subset of size $1 \leq m \leq n$, let $\rho = m/n$, $\bar{b} = n^{-1}\sum_i b_i$, and $s_b^2 = (n-1)^{-1}\sum_i (b_i - \bar{b})^2$. Conditional on the complete fixed sample, the estimator and its design variance are

$$\begin{aligned} \widehat{\Delta}_{n,m} &= \frac{1}{n}\sum_i c_i - \frac{1}{m}\sum_i r_i b_i, \\ \mathbb{E}_r[\widehat{\Delta}_{n,m}] = \Delta_n, &\qquad \mathrm{Var}_r(\widehat{\Delta}_{n,m}) = \frac{1-\rho}{m} s_b^2. \end{aligned} \tag{32}$$

The identities follow from $\mathbb{E}_r r_i = \rho$, $\mathrm{Var}_r(r_i) = \rho(1-\rho)$, and $\mathrm{Cov}_r(r_i, r_j) = -\rho(1-\rho)/(n-1)$ for $i \neq j$. Because b_i is binary, $s_b^2 \leq n/[4(n-1)]$. For $0 < \alpha < 1$, Chebyshev's inequality therefore gives a conservative confidence interval $[\widehat{\Delta}_{n,m} - \varepsilon_{n,m}, \widehat{\Delta}_{n,m} + \varepsilon_{n,m}]$ with coverage at least $1 - \alpha$, where $\varepsilon_{n,m} = \sqrt{(n-m)/[4\alpha m(n-1)]}$. Intersecting it with Eq. 31 preserves coverage and, when the intersection is nonempty, gives width at most $\min\{w(r), 2\varepsilon_{n,m}\}$. At fixed α, the second term is $\mathrm{O}(\sqrt{(1-\rho)/(n\rho)})$. Full measurement gives zero variance and the exact sample change. With $m = 0$, only the completion bounds apply.

The completion bounds retain all compatible missing judgments, whereas the confidence interval uses the sampling design to control error. For population identification, any positive independent reveal probability makes mean initial correctness equal to its observed mean among revealed trials; it does not leave a population identified-set width of $(1-\rho)g$. The following result describes estimation of that population mean from a growing sample.

Corollary 4 (Sparse pre-advice correctness restores consistency)**.** *Suppose each of n i.i.d. interactions independently reveals only the correctness bit $b_i = \mathbf{1}\{\hat{y}_{\mathrm{h},i} = y_i\}$ with probability $\pi_n > 0$, while $(y_i, \hat{y}_{\mathrm{a},i}, \hat{y}_i)$ are always retained. The reveal indicator only controls recording of an already-defined judgment and does not change the distribution of $(y_i, \hat{y}_{\mathrm{h},i}, \hat{y}_{\mathrm{a},i}, \hat{y}_i)$. If it is independent of the interaction and $n\pi_n \to \infty$, then there exists an estimator $\widehat{\Delta}_n$ satisfying*

$$\mathbb{E}(\widehat{\Delta}_n - \Delta)^2 = \mathrm{O}\left(\frac{1}{n} + \frac{1}{n\pi_n}\right) = o(1).$$

Thus the reveal fraction may vanish while consistent estimation remains possible, provided the number of revealed correctness bits diverges.

When revealing b requires asking for a pre-advice judgment, use the tags tar and aux for unprompted and prompted decisions, respectively, as in the preceding subsection. Random assignment permits $p_{\mathrm{final}}^{\mathrm{tar}}$ to be estimated from unprompted trials and q_{aux} from prompted trials. Their difference identifies the natural-workflow Δ_{tar} only if mean pre-advice accuracy satisfies $q_{\mathrm{tar}} = q_{\mathrm{aux}}$; conditional analyses require the analogous equality within each z. If that condition is weakened to a bound on $|q_{\mathrm{tar}}(z) - q_{\mathrm{aux}}(z)|$, Corollary 3 gives the remaining range. The HAIID mask-and-reveal exercise

uses pre-advice judgments that were already collected and cannot test whether prompting changes pre-advice accuracy. If an unprompted pre-advice judgment is not well defined, Δ_{tar} must first be specified before this comparison is meaningful.

The next two subsections relax keep-or-adopt, first alone and then jointly with the bounded-selection constraint of Proposition 4.

B.4 Support violations

Let $\mathbf{V} := \{\hat{y} \notin \{\hat{y}_{\mathrm{h}}, \hat{y}_{\mathrm{a}}\}\}$, assume $\Pr(\mathbf{V}) \leq \eta$, and define $r_+ := \Pr(\hat{y} \neq \hat{y}_{\mathrm{a}}, \hat{y} = y)$ and $r_- := \Pr(\hat{y} \neq \hat{y}_{\mathrm{a}}, \hat{y} \neq y)$.

Theorem 3 (Sharp robust accuracy-change interval)**.** *Under* $\Pr(\mathbf{V}) \leq \eta$*,*

$$\mathbf{I}_{\Delta}^{(\eta)} = [-p_- - \min\{\eta, r_-\},\ p_+ + \min\{\eta, r_+\}]$$

is sharp, with width $g + \min\{\eta, r_-\} + \min\{\eta, r_+\}$.

The lower and upper endpoints use different compatible joint distributions, so the violation budget is not spent twice within one joint distribution. This relaxation is most relevant in multiclass or deliberative settings, where a final answer can be neither $\hat{y}_{\mathrm{h}}$ nor $\hat{y}_{\mathrm{a}}$.

For the robust-width demonstration we fix $p_+ = 0.28$, $p_- = 0.17$, $r_+ = 0.25$, and $r_- = 0.30$ and vary η from 0 to 0.35. The identified width follows Theorem 3 and increases continuously from 0.45 to 1.00.

B.5 Joint sensitivity to agreement selection and support violations

We now combine support violations with the bounded-selection constraint $|q_{\mathrm{agree}} - q_{\mathrm{disagree}}| \leq \kappa$. Let $d := 1 - g$, $p_{\mathrm{final}} := \Pr(\hat{y} = y)$, and retain $r_+ := \Pr(\hat{y} \neq \hat{y}_{\mathrm{a}}, \hat{y} = y)$ and $r_- := \Pr(\hat{y} \neq \hat{y}_{\mathrm{a}}, \hat{y} \neq y)$. For $d > 0$, define

$$\begin{aligned} \underline{q}_{\mathrm{disagree}}(\eta) &:= \frac{r_+ - \min\{\eta, r_+\}}{d}, \\ \overline{q}_{\mathrm{disagree}}(\eta) &:= \frac{r_+ + \min\{\eta, r_-\}}{d}. \end{aligned} \tag{33}$$

These are the sharp lower and upper bounds on $q_{\mathrm{disagree}} = \Pr(\hat{y}_{\mathrm{h}} = y \mid \hat{y} \neq \hat{y}_{\mathrm{a}})$ under $\Pr(\hat{y} \notin \{\hat{y}_{\mathrm{h}}, \hat{y}_{\mathrm{a}}\}) \leq \eta$.

Proposition 12 (Joint (η, κ) sensitivity region)**.** *Suppose* $0 < g < 1$, $\Pr(\hat{y} \notin \{\hat{y}_{\mathrm{h}}, \hat{y}_{\mathrm{a}}\}) \leq \eta$*, and* $|q_{\mathrm{agree}} - q_{\mathrm{disagree}}| \leq \kappa$*. Then the sharp identified set for* Δ *is*

$$\begin{aligned} \mathbf{I}_{\Delta}^{(\eta,\kappa)} = \Big[& p_{\mathrm{final}} - d\overline{q}_{\mathrm{disagree}}(\eta) - g \min\{1, \overline{q}_{\mathrm{disagree}}(\eta) + \kappa\}, \\ & p_{\mathrm{final}} - d\underline{q}_{\mathrm{disagree}}(\eta) - g \max\{0, \underline{q}_{\mathrm{disagree}}(\eta) - \kappa\} \Big]. \end{aligned} \tag{34}$$

At $\eta = 0$*, Eq. 34 reduces to Proposition 4; at* $\kappa = 1$*, it reduces to the robust interval in Theorem 3.*

C Information requirements and workflow implications

The preceding sections say what observed data identify. This section asks what extra field identifies a given target, when aggregate accuracy change still leaves switching rates unidentified, when a sequential measurement can be reused in another protocol, and how a remaining interval for Δ changes an assistance decision.

C.1 Minimum information for a target claim

For an observable cell o, let $\mathbf{H}(o)$ be its compatible latent judgments, let $t_o(\hat{y}_{\mathrm{h}})$ be a per-instance target contribution, and let an additional recorded field be $m = \psi_o(\hat{y}_{\mathrm{h}})$.

Proposition 13 (Minimum information for a target claim)**.** *The pair (o, m) identifies $\mathbb{E}[t_o(\hat{y}_{\mathrm{h}})]$ for every compatible joint distribution if and only if t_o is constant on every fiber of ψ_o for every observable cell o. If the encoding may depend on o, the minimum alphabet size sufficient for every observable record is*

$$k^* = \max_o |t_o(\mathbf{H}(o))|.$$

If one recorded symbol merges two latent judgments with different target contributions, point masses on those judgments are observationally equivalent but have different targets. Otherwise t_o factors through the recorded symbol and its expectation is observed. Separating the distinct target contributions therefore gives both necessity and attainability of k^*.

For final accuracy, no extra field is required. For the net 0–1 accuracy change, the binary pre-advice correctness indicator $\mathbf{1}\{\hat{y}_{\mathrm{h}} = y\}$ is sufficient and worst-case necessary. In multiclass problems, jointly identifying the accuracy change and whether a switch occurred can require three recorded states; recovering the full pre-advice judgment requires $|\mathbf{D}|$ states in the worst case. A generic switch/no-switch bit is not sufficient for the accuracy change in every multiclass cell: when $\hat{y} = \hat{y}_{\mathrm{a}} \neq y$, switching from $\hat{y}_{\mathrm{h}} = y$ is harmful whereas switching from a third incorrect label is accuracy-neutral.

Table 3 compares information sources for accuracy change. Propensity and sparse correctness measurements are alternative additions to final-only records. The selection and model rows retain propensity information; the remaining rows describe other possible records or summaries.

Table 3: **Information about the person and the accuracy-change range.**

Information	What it reveals or restricts	Consequence for Δ	Source
Final-only data	$q_{\mathrm{agree}} \in [0, 1]$	Width g	Recorded data
Average adoption propensity	Population mean θ in Eq. 4	Eq. 6; exact θ need not identify Δ	Assumption or auxiliary data
Sampled pre-advice correctness	Uniform sample of m correctness bits; $\rho = m/n$	Expected sample completion width $(1-\rho)g_n$; random measurement also supports estimation	Target-workflow data
Propensity and bounded selection	Retain θ; add $\lvert q_{\mathrm{agree}} - q_{\mathrm{disagree}}\rvert \leq \kappa$	Sharp intersection in Eq. 7	Marginal information + assumption
Behavioral model	Retain marginal and relational constraints; add a generative mechanism	A subset of the preceding compatible class	Model + retained information
Marginal confidence	Distribution of C	No guaranteed contraction without a link to judgments or advice use	Assumption or auxiliary data
Pre-advice accuracy range	$q_{\mathrm{pre}} \in [\underline{q}, \overline{q}]$	Eq. 29; exact q_{pre} gives width 0	Assumption or auxiliary data
Full pre-advice judgment	Same-trial $\hat{y}_{\mathrm{h}}$	Identifies Δ and individual switching	Target-workflow data

C.2 Net accuracy change versus beneficial and harmful switching

Define the probabilities of beneficial and harmful switching

$$s_+ := \Pr(\hat{y} = \hat{y}_{\mathrm{a}} = y, \hat{y}_{\mathrm{h}} \neq \hat{y}_{\mathrm{a}}),$$
$$s_- := \Pr(\hat{y} = \hat{y}_{\mathrm{a}} \neq y, \hat{y}_{\mathrm{h}} = y).$$

The first is beneficial switching to a correct recommendation; the second is harmful switching away from a correct pre-advice judgment.

Lemma 1 (Sharp joint bounds on beneficial and harmful switching)**.** *For a fixed observed-data distribution,*

$$\mathbf{I}_{s_+, s_-} = [0, p_+] \times [0, p_-]$$

is sharp for binary or multiclass decisions.

Suppose auxiliary experimental information identifies the relevant same-person net accuracy change Δ. This can identify net accuracy change while leaving beneficial and harmful switching rates unidentified.

Theorem 4 (Net accuracy change versus beneficial and harmful switching)**.** *Suppose $\Delta \in [-p_-, p_+]$ is known exactly in addition to the observed-data distribution. Then*

$$s_- \in [\max(0, -\Delta),\ \min(p_-, p_+ - \Delta)], \qquad s_+ = \Delta + s_-,$$

and the probability of an accuracy-changing switch $m_{\mathrm{sw}} := s_+ + s_-$ has sharp set

$$m_{\mathrm{sw}} \in [|\Delta|,\ \min\{2p_+ - \Delta, 2p_- + \Delta\}]\,.$$

In particular, $\Delta = 0$ allows $m_{\mathrm{sw}} \in [0, 2\min(p_+, p_-)]$.

The construction in Proposition 1 varies beneficial and harmful switching independently on the two agreement cells, proving Lemma 1. Intersect its rectangle with $s_+ - s_- = \Delta$ and write $m_{\mathrm{sw}} = \Delta + 2s_-$ to obtain the displayed sharp intervals.

Thus even an exactly known net accuracy change need not reveal how often correctness changed between the pre-advice judgment and final decision. In the HAIID mask-and-reveal data, where the support-violation rate is 0.0034, the pooled values $p_+ = 0.6935$, $p_- = 0.1465$, and the full-data estimate $\Delta = 0.0740$ imply the baseline bounds on the probability of an accuracy-changing switch $m_{\mathrm{sw}} \in [0.074, 0.367]$; the observed frequency of accuracy-changing switches is approximately 0.149 and lies inside this range. Because the support-violation rate is small but nonzero, we use this calculation only to illustrate the distinction between net accuracy change and switching rates.

C.3 Sequential elicitation across protocols

Write $(\hat{y}_{\mathrm{h}}^{\mathrm{conc}}, \hat{y}^{\mathrm{conc}})$ and $(\hat{y}_{\mathrm{h}}^{\mathrm{seq}}, \hat{y}^{\mathrm{seq}})$ for the latent and final responses under concurrent and sequential protocols.

Proposition 14 (Limits of transporting pre-advice judgments across protocols)**.** *Suppose infinite data identify both $p_{\mathrm{conc}}(y, \hat{y}_{\mathrm{a}}, \hat{y}^{\mathrm{conc}})$ and $p_{\mathrm{seq}}(y, \hat{y}_{\mathrm{a}}, \hat{y}_{\mathrm{h}}^{\mathrm{seq}}, \hat{y}^{\mathrm{seq}})$, and the keep-or-adopt assumption holds in the concurrent protocol. Without a constraint linking $\hat{y}_{\mathrm{h}}^{\mathrm{conc}}$ to $\hat{y}_{\mathrm{h}}^{\mathrm{seq}}$, the concurrent accuracy change remains sharply identified only as*

$$\Delta_{\mathrm{conc}} \in [-p_{-,\mathrm{conc}}, p_{+,\mathrm{conc}}],$$

where $p_{+,\mathrm{conc}} = \Pr_{\mathrm{conc}}(\hat{y}^{\mathrm{conc}} = \hat{y}_{\mathrm{a}} = y)$ and $p_{-,\mathrm{conc}} = \Pr_{\mathrm{conc}}(\hat{y}^{\mathrm{conc}} = \hat{y}_{\mathrm{a}} \neq y)$.

For every target value in the concurrent baseline interval, Proposition 1 supplies a compatible concurrent joint distribution. Without a cross-protocol constraint, each can be paired with the same fixed sequential joint distribution. Thus observing the sequential responses excludes none of those target values.

Randomized assignment between interfaces can identify differences in observable protocol outcomes, and a randomized no-assistance arm can identify a causal assistance contrast when it represents the deployment baseline of interest. Randomization between interfaces does not, however, imply that eliciting a pre-advice judgment leaves unchanged the latent judgment that would have existed under a concurrent workflow. Proposition 14 formalizes this identification problem across protocols.

C.4 Context-specific assistance decisions

Before allowing assistance decisions to vary with context, the identified interval itself has a direct decision consequence. Normalize no assistance to 0 and let $\delta \in [l, u]$ be the value of assistance relative to the pre-advice judgment.

Proposition 15 (Minimax regret from an identified interval)**.** *The minimax-optimal assistance probability and regret are*

$$q^*(l,u) = \begin{cases} 1, & l \geq 0, \\ 0, & u \leq 0, \\ u/(u-l), & l < 0 < u, \end{cases}$$

$$r^*(l,u) = \begin{cases} 0, & l \geq 0 \text{ or } u \leq 0, \\ (-l)u/(u-l), & l < 0 < u. \end{cases}$$

When the interval has one sign, always taking the better action gives zero regret. When $l < 0 < u$, worst-case regret is $\max\{u(1-q), -lq\}$; equalizing these two terms gives the stated solution.

This decision criterion follows the minimax-regret tradition in statistical decision theory (Manski, 2004; Stoye, 2009).

Let z be an observable context with finite or countable support and define $p_+(z) := \Pr(\hat{y} = \hat{y}_\mathrm{a} = y \mid z)$ and $p_-(z) := \Pr(\hat{y} = \hat{y}_\mathrm{a},\, \hat{y}_\mathrm{a} \neq y \mid z)$, with $g(z) = p_+(z) + p_-(z)$. Proposition 1 applies conditionally, giving $\Delta(z) \in [-p_-(z), p_+(z)]$. Normalize the value of no assistance to 0 and let δ denote the value of assistance relative to the pre-advice human judgment. A policy that assists with probability q then has regret $r(q,\delta) := \max\{0,\delta\} - q\delta$.

Corollary 5 (Context-specific decision regret relative to $\hat{y}_\mathrm{h}$)**.** *A possibly randomized learner may use any number of final-only training records and the current context z to choose an assistance probability $q_n(z)$. Over the class of compatible joint distributions that permits endpoints to vary across contexts, every such learner has worst-case expected regret at least*

$$r^*_{\mathrm{ctx}} := \mathbb{E}_z\left[\frac{p_+(z)p_-(z)}{p_+(z)+p_-(z)}\right], \tag{35}$$

where the integrand is zero when $p_+(z) + p_-(z) = 0$. If the observed-data distribution is known, $q^*(z) = p_+(z)/(p_+(z)+p_-(z))$ attains the bound. The constant-context case is $p_+p_-/(p_+ + p_-)$.

This criterion becomes a causal assistance/no-assistance comparison only when the pre-advice human judgment represents the relevant no-assistance baseline. Writing $\rho = p_+/g$ in a constant context gives $r^* = g\rho(1-\rho)$, so uncertainty about the magnitude of Δ and uncertainty about the action can differ sharply.

D PROOFS

The arguments below establish the shared sharpness constructions, statistical limits, and sensitivity results. Direct substitutions and consequences already justified beside their statements are omitted.

D.1 BASELINE SHARPNESS AND THE AGREEMENT REDUCTION

On final disagreement, $\hat{y}_\mathrm{h} = \hat{y}$, so the accuracy-change contribution is zero. On correct-advice agreement, choosing the mass with $\hat{y}_\mathrm{h} \neq \hat{y}_\mathrm{a}$ gives any beneficial contribution $b \in [0, p_+]$. On incorrect-advice agreement, choosing the mass with $\hat{y}_\mathrm{h} = y$ gives any harmful contribution $h \in [0, p_-]$. These choices are independent across the two observed cells and preserve the full recorded distribution. Hence $\Delta = b - h$ fills $[-p_-, p_+]$, proving the interval and width statements of Proposition 1.

Behavioral information and identified-set width. Let $\mathbf{K}$ denote compatible additional information, and let $\mathbf{H_K}$ contain all values of q_agree allowed by the observations and $\mathbf{K}$. Their corresponding accuracy changes form the identified set

$$\begin{aligned} \mathbf{I}_\Delta(\mathbf{K}) &= p_+ - g\mathbf{H_K}, \\ \operatorname{diam}(\mathbf{I}_\Delta(\mathbf{K})) &= g\,\operatorname{diam}(\mathbf{H_K})\,. \end{aligned} \tag{36}$$

Without further information, $\mathbf{H_K} = [0,1]$ and the width is g; fixing q_agree identifies Δ. When $g = 0$, $\Delta = 0$.

Conditional form of Eqs. 2 and 36. Let $\mathbf{H}_{\mathbf{K}}(z)$ be the exact projection of all compatible joint distributions satisfying $\mathbf{K}$ onto $\Pr(\hat{y}_{\mathrm{h}} = y \mid \hat{y} = \hat{y}_{\mathrm{a}}, z)$. The identical correctness contributions on final disagreement cancel, leaving

$$\Delta(z) = p_+(z) - g(z)q_{\mathrm{agree}}(z), \qquad \mathbf{I}_{\Delta}(z; \mathbf{K}) = p_+(z) - g(z)\mathbf{H}_{\mathbf{K}}(z).$$

The affine image is sharp by the definition of the projection; its diameter is $g(z)\operatorname{diam}\mathbf{H}_{\mathbf{K}}(z)$. If $g(z) = 0$, then $\Delta(z) = 0$. Pooled restrictions retain their original scope when forming this projection and are not imposed separately within every context. □

D.2 Estimation limits from final-only data

The estimation claim in Proposition 1 takes the following formal form. For $q \in \mathbf{M}_{\mathrm{KA}}(p_o)$, let $\Delta(q)$ be the corresponding accuracy change. For every n and every estimator $\widehat{\Delta}_n$ based only on $o_1, \ldots, o_n$,

$$\sup_{q \in \mathbf{M}_{\mathrm{KA}}(p_o)} \mathbb{E}_q |\widehat{\Delta}_n - \Delta(q)| \geq \frac{g}{2}.$$

Shared endpoint argument. If two compatible targets a, b induce the same sampling law, every estimator T has that same law under both. The pointwise inequalities

$$|T - a| + |T - b| \geq |a - b|, \qquad \frac{(T-a)^2 + (T-b)^2}{2} \geq \frac{(a-b)^2}{4}$$

give worst-case expected absolute and squared errors of at least $|a - b|/2$ and $(a - b)^2/4$. Applying them to the attainable endpoints $a = -p_-$ and $b = p_+$ proves the estimation claim in Proposition 1 and the squared-error floor $g^2/4$.

For a measurable confidence set $\mathbf{C}_n$ with uniform coverage $1 - \alpha$, sharpness supplies the same sampling law for every $\delta \in [-p_-, p_+]$. Fubini's theorem therefore gives

$$\mathbb{E}\operatorname{diam}(\mathbf{C}_n) \geq \mathbb{E}\,\lambda(\mathbf{C}_n \cap [-p_-, p_+]) = \int_{-p_-}^{p_+} \Pr\{\delta \in \mathbf{C}_n\}\, d\delta \geq (1 - \alpha)g,$$

where λ is Lebesgue measure. □

D.3 Finite-sample confidence interval and estimation bound

Let $\widehat{p}_+$ and $\widehat{p}_-$ be the empirical frequencies of correct- and incorrect-advice agreement. For independent observations, one-sided Hoeffding inequalities and a union bound imply that

$$\mathbf{C}_n^{\mathrm{H}} = [-\widehat{p}_- - \varepsilon_n,\, \widehat{p}_+ + \varepsilon_n], \qquad \varepsilon_n = \sqrt{\frac{\log(2/\delta)}{2n}}$$

covers the entire identified set with probability at least $1 - \delta$. Its untruncated expected width is $g + 2\varepsilon_n$. This conservative set-coverage construction differs from intervals aimed at covering the partially identified scalar alone (Imbens & Manski, 2004).

For the midpoint estimator $\widehat{m}_n = (\widehat{p}_+ - \widehat{p}_-)/2$, disjoint agreement cells give $\operatorname{Var}(\widehat{m}_n) \leq g/(4n)$, and hence $\mathbb{E}|\widehat{m}_n - m(p_o)| \leq \sqrt{g}/(2\sqrt{n})$. Every compatible target remains within $g/2$ of the population midpoint $m(p_o) = (p_+ - p_-)/2$, separating sampling error from the identification limit.

D.4 Continuous-response sharpness

For Theorem 1, choose measurable loss-minimizing and loss-maximizing pre-advice judgments in $[l, u]$ for each observed outcome. Keeping the observed distribution fixed while using these conditional choices attains the two endpoints in Eq. 24; mixing the choices fills the interval.

For Corollary 2, the segment constraint gives precisely $\mathbf{H}_{\mathrm{cvx}}$ in Eq. 27: if $\hat{y}_{\mathrm{a}} > \hat{y}$, feasible initial judgments satisfy $\hat{y}_{\mathrm{h}} \leq \hat{y}$, with $w = (\hat{y} - \hat{y}_{\mathrm{h}})/(\hat{y}_{\mathrm{a}} - \hat{y}_{\mathrm{h}})$; the reverse inequality applies when $\hat{y}_{\mathrm{a}} < \hat{y}$, and $w = 1$ permits any initial judgment when $\hat{y}_{\mathrm{a}} = \hat{y}$. Apply the same measurable-selector construction over this smaller compact domain. It proves the sharp interval and its containment in the unrestricted set. Both minimax claims follow from the shared endpoint argument in Appendix D.2, using the respective endpoint separations γ_0 and γ_{cvx}. □

D.5 Pre-advice accuracy information

For Theorem 2, a target value $\delta(z)$ in the baseline interval has a compatible construction by Proposition 1. That construction has initial accuracy $p_{\text{final}}(z) - \delta(z)$, so it satisfies the supplied accuracy bounds exactly when $\delta(z) \in [l(z), u(z)]$. This proves sharpness of the intersection. Endpoint constructions can be chosen independently across the finite or countable contexts; averaging their lower and upper values gives Eq. 30. □

D.6 Sparse correctness measurement

For Corollary 4, let $r_i \sim \text{Bernoulli}(\pi_n)$ be independent reveal indicators, $b_i = \mathbf{1}\{\hat{y}_{\text{h},i} = y_i\}$, and $c_i = \mathbf{1}\{\hat{y}_i = y_i\}$. The inverse-probability estimator

$$\widehat{\Delta}_n = \frac{1}{n}\sum_{i=1}^{n}\left(c_i - \frac{r_i b_i}{\pi_n}\right)$$

is unbiased. The bounds $\text{Var}(c_i) \le 1/4$ and $\text{Var}(r_i b_i/\pi_n) \le 1/\pi_n$ give $\mathbb{E}(\widehat{\Delta}_n - \Delta)^2 \le 1/(2n) + 2/(n\pi_n)$. Thus $n\pi_n \to \infty$ suffices, even when $\pi_n \to 0$. □

D.7 Context-specific decision regret

For Corollary 5, let $m_n(z) = \mathbb{E}[q_n(z) \mid z]$, including training-sample and policy randomization. All compatible distributions induce the same training-sample law, so $m_n(z)$ is common to them. The baseline endpoint constructions can be chosen independently in each finite or countable context. Since regret is convex in the target, choosing the worse endpoint gives

$$\sup_q \mathbb{E}_q[r(q_n(z), \Delta_q(z))] = \mathbb{E}_z \max\{p_+(z)(1 - m_n(z)),\, p_-(z)m_n(z)\}.$$

Proposition 15, applied within each context, bounds this below by Eq. 35. When the observed law is known, $q^*(z) = p_+(z)/g(z)$ attains it wherever $g(z) > 0$; contexts with $g(z) = 0$ have zero regret. □

D.8 Support violations and joint selection sensitivity

Theorem 3. Agreement cells cannot violate keep-or-adopt, so they retain the contribution range $[-p_-, p_+]$. On correct final disagreement, a violation can change initial correctness from one to zero, adding at most $\min(\eta, r_+)$ to the gain. On incorrect final disagreement, it can change initial correctness from zero to one, subtracting at most $\min(\eta, r_-)$. Each endpoint is attained by placing the available violation mass in its respective disagreement cell and using the matching agreement-cell construction. The two endpoints use different distributions and do not spend the same budget twice. Mixtures preserve the budget and fill the interval.

Proposition 12. The same disagreement-cell allocations give the attainable range for q_{disagree} in Eq. 33. For each such value, agreement correctness can vary freely within $[\max(0, q_{\text{disagree}} - \kappa), \min(1, q_{\text{disagree}} + \kappa)]$ without spending violation mass. Since

$$\Delta = p_{\text{final}} - d q_{\text{disagree}} - g q_{\text{agree}},$$

the lower endpoint uses the largest feasible disagreement and agreement correctness rates, and the upper endpoint uses the smallest. These choices give Eq. 34 and are attained by the stated allocations. Convex mixtures give every intermediate value. The special cases follow by setting $\eta = 0$ or $\kappa = 1$. □

E Supporting Analyses for the Constraint Comparison

We organize the supporting evidence around four questions: what remains ambiguous in the recorded data, what summary constraints remove, what fitted mechanisms add after calibration, and what direct measurement resolves. Full-data references use the revealed initial judgments; evaluation intervals use only the information specified below. Widths are in percentage points unless a different scale is stated.

E.1 Data and Baseline Ambiguity

Paired data and encoding. In HAIID (Vodrahalli et al., 2022a), the signs of the initial response, advice, and final response on the correctness-coded slider define the binary labels. Of 35,670 rows from 1,125 participants, 959 have an exact-zero initial or final response, which does not specify a label; none has missing response fields or zero advice. The primary AI-presented arm excludes 452 such midpoint cases from 17,973 rows, leaving 17,521 trials from 567 participants. Human-presented advice is a secondary comparison. For Tejeda, we pair sequential records within participant and image only when the post-advice task index immediately follows the initial index. Excluding two incomplete pairs gives 14,206 trials from 74 participants. Truth is the image category and advice is the argmax of the 16-class probability vector displayed to participants; final choices of a third label count as support violations.

Table 4 reports the six primary task samples and their support-relaxed baselines, followed by pooled HAIID and Pearson as secondary comparisons. The binary tasks have disjoint participants but share one study design. Masking initial judgments leaves both signs feasible in every task and classifier condition, whereas the full-data gains are positive. Paired-data uncertainty uses 5,000 participant-bootstrap replicates. HAIID's pooled gain is 0.0740, with 95% interval $[0.0658, 0.0823]$; all five task-level intervals are also positive.

Table 4: **Task samples and baseline ambiguity.** (a) Six primary tasks from two studies, with task-specific observed support allowances; gains and endpoints are in percentage points. (b) Pooled HAIID and the final-only Pearson application as secondary comparisons; this panel uses probabilities and strict keep-or-adopt intervals. Tejeda records object recognition. Full-data gains use the revealed paired judgments.

(a) Primary task comparisons (percentage points)

Task	Classes	People	Trials	Gain (pp)	Support interval (pp)
Art	2	278	8,691	+9.49	$[-14.82, 69.59]$
Census	2	43	1,352	+3.55	$[-19.97, 72.19]$
Cities	2	79	2,448	+6.45	$[-14.83, 71.04]$
Dermatology	2	20	462	+6.06	$[-6.93, 66.88]$
Sarcasm	2	147	4,568	+5.19	$[-14.73, 68.70]$
Object recognition	16	74	14,206	+10.95	$[-29.03, 60.70]$

(b) Pooled and final-only references (probabilities)

Dataset	Participants	Trials	$g = \Pr(\hat{y} = \hat{y}_a)$	$\mathbf{I}_\Delta$	Full-data Δ
HAIID AI-presented mask-and-reveal	567	17,521	0.840	$[-0.147,\ 0.694]$	0.074
Tejeda mask-and-reveal	74	14,206	0.668	$[-0.158,\ 0.511]$	0.110
Pearson AI guidance	141	11,280	0.605	$[-0.224,\ 0.381]$	–

Table 5 collects the additional baseline diagnostics. Its task and condition regret quantities use Corollary 5; in Tejeda, averaging conditional regret gives 0.1196, close to the pooled value 0.1204 in this sample. The last panel supplies empirical reference scales for selection and support departures used below.

Final-only application and encoding check. Pearson contributes 141 AI-guidance participants with 80 experimental trials each, excluding attention checks. The primary reconstruction matches the published Table-2 totals, including mean agreement of $48.369/80$ trials. The release's A stream agrees with its stored consistency totals, but the B stream has conflicting trial labels and totals. Two alternative reconstructions yield $g \in [0.644, 0.652]$ and endpoints reaching about -0.248 and $+0.405$, preserving sign ambiguity. Under the primary encoding, disagreement accuracy is 0.6989, giving $\Delta_{\text{equal}} = -0.0416$ and sign-breakdown point 0.0688; even $\kappa = 0.30$ leaves $[-0.2230, 0.1398]$ feasible.

Does increasing the final-only sample resolve the ambiguity? We simulate multinomial observed cells with $p_+ = p_- = g/2$ and $r_+ = r_- = (1-g)/2$, using $g \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $n \in \{25, 50, 100, 200, 500, 1000, 2500, 5000\}$. Each cell uses 5,000 repetitions. The 95% Hoeffding set in Appendix D has expected width $g + 2\sqrt{\log(40)/(2n)}$, and midpoint MSE at either compat-

Table 5: **Baseline diagnostics and revealed behavioral summaries.** Rates and gains are probabilities. Here η is the observed support-violation rate. $q^* = p_+/g$ and $r^* = p_+p_-/g$ describe minimax assistance and regret. In (c), $\widehat{\kappa} = |q_{\text{agree}} - q_{\text{disagree}}|$; CIs use participant resampling and the final column allows both selection and support departures.

(a) HAIID tasks

Task	Participants	Trials	g	Baseline $\mathbf{I}_\Delta$	Full-data Δ	η	r^*
Art	278	8,691	0.839	$[-0.145, 0.693]$	0.095	0.003	0.120
Census	43	1,352	0.919	$[-0.198, 0.720]$	0.036	0.001	0.155
Cities	79	2,448	0.853	$[-0.145, 0.708]$	0.065	0.003	0.121
Dermatology	20	462	0.703	$[-0.052, 0.652]$	0.061	0.017	0.048
Sarcasm	147	4,568	0.826	$[-0.143, 0.683]$	0.052	0.004	0.119

(b) Object-recognition classifier conditions

Condition	$g(z)$	Full-data $\Delta(z)$	$\eta(z)$	$q^*(z)$	$r^*(z)$
Below-human	0.538	0.054	0.147	0.715	0.110
At-human	0.727	0.110	0.119	0.761	0.132
Above-human	0.734	0.161	0.133	0.802	0.117

(c) Selection and support departures

Dataset	q_{agree}	q_{disagree}	$\widehat{\kappa}$	95% CI	$\widehat{\eta}$	$\mathbf{I}_\Delta^{(\eta,\kappa)}$
HAIID	0.736	0.430	0.306	$[0.277, 0.335]$	0.003	$[0.060,\ 0.616]$
Tejeda	0.612	0.267	0.346	$[0.307, 0.383]$	0.133	$[-0.290,\ 0.607]$

ible endpoint is $g^2/4 + g/(4n)$. Figure 5 confirms that sampling error vanishes around nonzero identification limits.

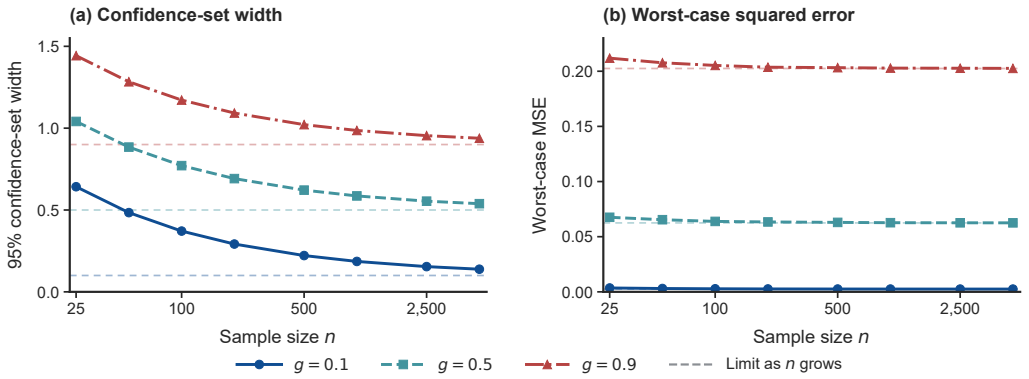


Figure 5: **More final-only observations reduce sampling error, not the identification limits.** Displayed curves use $g \in \{0.1, 0.5, 0.9\}$; dashed lines mark limiting width g and worst-endpoint MSE $g^2/4$. Each point averages 5,000 repetitions.

E.2 Summary Constraints: Contributions, Relaxation, and Calibration

One joint optimization for all stages. For observed cell $o = (y, a, f)$, optimize over nonnegative initial-label masses $w_{o,h}$ with $\sum_h w_{o,h} = \widehat{p}_o$. Write

$$V(w) = \sum_{o,h} w_{o,h} \mathbf{1}\{f \notin \{h, a\}\},$$

$$D(w) = \sum_{o,h} w_{o,h} \mathbf{1}\{h \neq a\}, \qquad S(w) = \sum_{o,h} w_{o,h} \mathbf{1}\{f = a, h \neq a\},$$

and let $Q_G(w)$ and $Q_D(w)$ be the initial-correctness masses on final agreement and disagreement. Successive stages retain

$$\begin{aligned} &V(w) \le \eta, \\ &\theta_L D(w) \le S(w) \le \theta_U D(w), \\ &-\kappa \le Q_G(w)/\widehat{g} - Q_D(w)/(1-\widehat{g}) \le \kappa. \end{aligned} \tag{37}$$

Minimizing and maximizing $\sum_{o,h} w_{o,h}(\mathbf{1}\{f=y\} - \mathbf{1}\{h=y\})$ gives the bounds. All behavioral-summary cases below have $\eta < 1-\widehat{g}$, so $D(w) > 0$. The separate direct-correctness branch does not constrain propensity and permits larger support allowances. Grouping cells by equality among truth, advice, final, and initial labels reduces the program to eight variables for binary labels and twenty for 16 labels. Distributing each pattern's completion proportionally over its observed cells preserves the full observed law, making this reduction lossless for the pooled constraints. Checks against closed-form bounds and full-label optimization on 180 random laws verify the implementation.

Revealed summaries and controlled relaxation. The full-sample values (η, θ, κ) are $(0.0033674, 0.4875817, 0.3058521)$ for HAIID and $(0.1327608, 0.4285714, 0.3455629)$ for Tejeda. Supplying them exactly guarantees inclusion of the revealed joint law. Table 6 places the pooled chain, subgroup chains, and sensitivity to these inputs together. Selection removes a further 69.3% of the propensity width in HAIID and 2.5% in Tejeda. The combined constraints determine a positive gain in four HAIID tasks; dermatology retains $[-2.70, 6.27]$ points, and no Tejeda condition determines a sign. The reference selection magnitudes and bootstrap intervals in Table 5 describe these workflows, rather than justify transfer to another study. Without propensity, the joint (η, κ) range contracts in HAIID but remains equal to the support-relaxed range in Tejeda.

For the sensitivity panel, hold the observed law and η fixed, expand propensity to $[\widehat{\theta} - \varepsilon, \widehat{\theta} + \varepsilon]$, and increase selection to $\widehat{\kappa} + \varepsilon$, clipping to $[0, 1]$. All intervals expand monotonically. HAIID retains a positive sign at $\varepsilon = 0.05$ but not 0.10; Tejeda recovers the support-only range at 0.20. These are deterministic relaxations of supplied summaries, not calibration uncertainty.

Participant calibration and reporting conventions. The summary and model protocols use 100 participant splits at calibration fractions $25\%, 50\%, 75\%$, stratified by task or classifier quality; all trials of a person stay together. The main direct-measurement sweep additionally uses smaller fractions (Appendix E.4). The summary-only protocol rounds the calibration count down within strata, with at least two people in each part. At 50%, HAIID has 282 calibration and 285 evaluation participants; Tejeda has 36 and 38. The optimizer receives evaluation final-only frequencies; evaluation initial judgments are reserved for checks. Within each split, 1,000 stratified bootstrap draws independently sample pseudo-calibration and pseudo-evaluation participants from the calibration group at the respective partition sizes. Quantiles of their summary differences, added to calibration estimates and clipped to $[0, 1]$, set prediction ranges. Behavioral-summary allowances retain the original six-tail allocation: each required one-sided tail receives $0.05/6$, with two for propensity, one for support, and one for selection. The displayed direct-correctness intervals instead use the finite-population construction in Appendix E.4. Seeds are 20260924 (HAIID) and 20260925 (Tejeda).

Throughout the calibrated experiments, width conditions on feasibility; an empty set has no width and fails gain inclusion. Sign determination means exclusion of zero. Restriction validity tests whether the revealed evaluation joint law satisfies a constraint, whereas gain inclusion tests only its aggregate accuracy change. Bootstrap behavioral allowances have no finite-sample coverage guarantee; the separate direct construction has the conditional design-based guarantee specified in Appendix E.4. Repeated splits reuse people and retain the original tasks and items; they establish neither independent replication nor new-item or cross-workflow transfer. The model protocols below retain these conventions but use their own matched summaries and uncertainty allocations.

Table 7 shows the half-split results. At 25% and 75%, combined-summary widths are 23.51 and 23.33 points in HAIID and 80.29 and 82.37 in Tejeda. All summary intervals include the evaluation gain across all fractions; no split determines the wrong sign. Nevertheless, support is valid in only 94% of HAIID and 89% of Tejeda half splits. Figure 6 compares these summaries with direct correctness on the same participant partitions. The refreshed direct widths at $25\%, 50\%, 75\%$ are 33.06, 24.58, and 26.47 points for HAIID and 86.78, 75.02, and 81.92 for Tejeda. Under these calibration procedures, combined summaries give narrower mean bounds than direct measurement at all three fractions in pooled HAIID. In Tejeda, direct measurement gives narrower mean bounds at 50% and 75%, but

Table 6: **Contributions of supplied summaries and sensitivity to their precision.** All stages retain the empirical support allowance. Panels (a,b) supply exact empirical propensity and selection; (c) relaxes both by ε on fixed pooled data. Endpoints and widths are in percentage points; ε is a probability. These within-sample ranges include the revealed gain by construction.

(a) Pooled cumulative bounds

Dataset	Information	Interval (pp)	Width (pp)	Total reduction
HAIID	Support allowance	$[-14.99, 69.69]$	84.68	0.0%
	+ Propensity θ	$[-14.74, 15.24]$	29.98	64.6%
	+ Selection κ	$[6.04, 15.24]$	9.20	89.1%
Tejeda	Support allowance	$[-29.03, 60.70]$	89.73	0.0%
	+ Propensity θ	$[-25.76, 34.54]$	60.29	32.8%
	+ Selection κ	$[-25.76, 33.01]$	58.76	34.5%

(b) The same chain within tasks and conditions

Dataset	Task / quality	Trials	Support	+ Propensity	+ Selection
HAIID	Art	8,691	84.41	29.68	9.71
	Census	1,352	92.16	20.49	6.70
	Cities	2,448	85.87	29.70	11.78
	Dermatology	462	73.81	13.85	8.97
	Sarcasm	4,568	83.43	21.50	6.82
Tejeda	Below human	4,608	83.16	62.57	58.14
	At human	4,799	90.69	59.94	59.94
	Above human	4,799	92.67	59.41	59.37

(c) Relaxing propensity and selection

	HAIID		Tejeda	
ε	Interval (pp)	Width (pp)	Interval (pp)	Width (pp)
0.00	$[6.04, 15.24]$	9.20	$[-25.76, 33.01]$	58.76
0.01	$[5.20, 15.85]$	10.65	$[-26.52, 34.89]$	61.41
0.02	$[4.36, 16.49]$	12.13	$[-27.26, 36.64]$	63.90
0.05	$[1.84, 18.60]$	16.76	$[-29.03, 40.11]$	69.14
0.10	$[-2.36, 22.79]$	25.15	$[-29.03, 46.86]$	75.89
0.20	$[-10.76, 35.21]$	45.97	$[-29.03, 60.70]$	89.73

Table 7: **Summary constraints and direct measurement on half the participants.** Means over the same 100 splits. Behavioral summaries retain six-tail bootstrap calibration; direct correctness uses finite-population calibration of correctness and support. All intervals are feasible.

Dataset	Information	Mean width (pp)	Inclusion	Sign determined
HAIID	Support allowance	85.28	100%	0%
	+ Propensity θ	35.06	100%	0%
	+ Selection κ	21.32	100%	31%
	Direct correctness	24.58	100%	9%
Tejeda	Support allowance	95.04	100%	0%
	+ Propensity θ	79.57	100%	0%
	+ Selection κ	79.57	100%	0%
	Direct correctness	75.02	100%	0%

wider bounds at 25%. All direct intervals include the evaluation gain. Direct sign-determination rates are $0\%, 9\%, 3\%$ for HAIID and zero at all three fractions for Tejeda. Increasing the calibration fraction also shrinks the evaluation sample, so this is not a learning curve with fixed evaluation size.

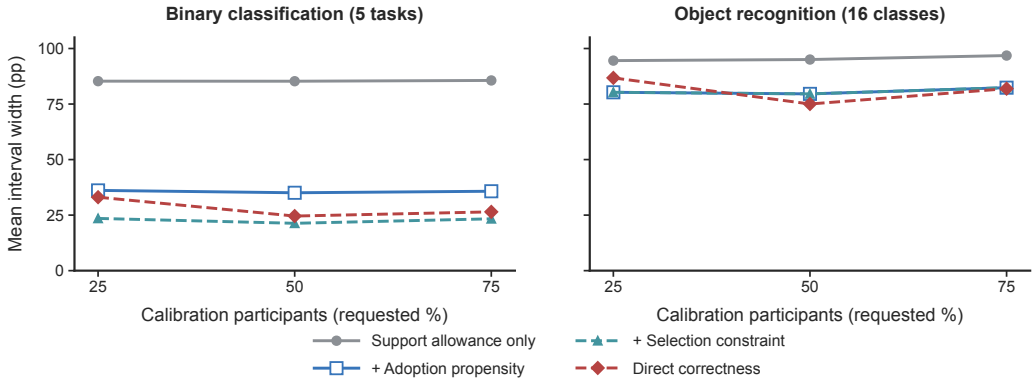


Figure 6: **Summary constraints and direct correctness across calibration fractions.** Mean widths over the same 100 participant partitions per fraction, without a fitted mechanism. Behavioral summaries use bootstrap calibration; direct correctness uses finite-population calibration. Propensity and propensity-plus-selection curves overlap for object recognition. Fractions are requested allocations; both partition sizes change.

E.3 Fitted Mechanisms: Contraction, Attribution, and Validity

Each model analysis combines its specification, calibrated component contributions, and assumption checks. Ablations remove only the named restriction while retaining the other budgets and all summary constraints. The full-sample references include the revealed joint law by construction; the calibrated comparisons test transfer to other participants.

E.3.1 HAIID Tasks: Activation and Integration

Fit and bound. Each binary task is fitted independently using four equal-width, sign-preserving response bins on $[-1, 1]$ and four contexts crossing advice correctness with advice magnitude above 0.5. No summary, component, or budget is pooled across tasks. Activation is any slider change exceeding 10^{-8}, including changes within a bin; inactivity is an unchanged response. The release's 0.07 threshold would require modeling small inactive changes. Calibration frequencies estimate activation a^0_{zj} and activated destination B^0_{zkj} separately in each context and initial bin, without parameter sharing or demographics. This fits the distributional extension, rather than reproducing the neural conditional-mean model of Vodrahalli et al. (2022b). The pooled HAIID comparison below uses the same construction with 20 task-crossed contexts on all 17,521 binary-analysis trials.

For globally normalized inactive masses s_{zj} and activated masses v_{zkj}, define $t_{zj} = \sum_k v_{zkj}$, $h_{zj} = s_{zj} + t_{zj}$, and fix $p_{zk} = s_{zk} + \sum_j v_{zkj}$ to observed context–final-bin frequencies. If $\mathcal{J}_A$ indexes initial columns observed in calibration and $\mathcal{J}_I$ indexes columns with observed activation, impose

$$\begin{aligned} \sum_{(z,j)\in\mathcal{J}_A} |t_{zj} - a^0_{zj} h_{zj}| &\le \varepsilon_A, \\ \tfrac{1}{2} \sum_{(z,j)\in\mathcal{J}_I} \sum_k |v_{zkj} - B^0_{zkj} t_{zj}| &\le \varepsilon_I. \end{aligned} \tag{38}$$

Unsupported components impose no restriction. These aggregate mass-weighted budgets replace the per-column tolerances in Proposition 11; its reconstruction $a_{zj} = t_{zj}/h_{zj}$, $B_{zkj} = v_{zkj}/t_{zj}$ on positive-mass columns proves sharpness for this class. Absolute-value auxiliaries yield a linear program for final minus initial correctness. Every model retains Eq. 37; removing both components numerically recovers its summary-only bounds, so context and final confidence alone do not explain contraction.

Task calibration and small samples. Each task uses 100 splits per fraction, seed 20260928 with task/fraction/repetition keys, and calibration counts rounded down with at least two people in each part. In task order art, census, cities, dermatology, sarcasm, half-split calibration/evaluation counts

are $139/139$, $21/22$, $39/40$, $10/10$, and $73/74$. Each of 1,000 bootstrap draws refits on pseudo-calibration people and evaluates discrepancies on independently drawn pseudo-evaluation people, sampled only from calibration participants at the respective partition sizes. Behavioral summaries and model discrepancies retain the original $0.05/8$ per-tail allocation. Direct correctness uses the same participants with the separate calibration in Appendix E.4.

The saved protocol documents handling of empty bootstrap conditioning cells: an undefined propensity or selection statistic retains $[0, 1]$. Lower target-minus-calibration differences use lower-target minus upper-calibration, and upper differences use upper-target minus lower-calibration. This equals the original rule on nondegenerate draws. Ten pseudo-calibration draws across five 25% census splits needed the extension; all actual split statistics and all half-split draws were defined. Dermatology's smaller partition contains only five people at the 25% and 75% fractions; trials are not treated as independent participants.

Contraction, attribution, and prediction. Table 8 combines interval validity, component ablations, and predictive checks on the same half splits. Exact supported components collapse each binary-task full-data reference to its revealed gain. Calibrated contraction beyond the summaries is 0.010, 6.331, 1.311, zero, and 0.010 points in task order. Integration supplies most of the census contraction; in cities, both components outperform either alone. Yet census joint validity is only 63%, despite 99% gain inclusion: support holds in 69% of splits, activation in 95%, and integration in 97%. Zero-error gain inclusion is zero in four binary tasks and 15% in dermatology.

Fitted kernels beat context-only and unchanged-response Brier baselines in all five tasks. These are retrospective predictions using revealed initial responses and truth-coded context. Columns unobserved in calibration remain unrestricted in the bounds; activation 0.5 and uniform destinations are scoring fallbacks only. Panel (c) reports the evaluation mass in these unsupported columns. Predictive improvement therefore coexists with heterogeneous bound contraction.

To compare the magnitudes of predictive and identification improvements, Figure 3 normalizes each change to its own within-task baseline. Green marks compare the fitted kernel's Brier loss with the context-only baseline; red marks compare gain-interval widths under behavioral patterns plus the model with those under behavioral patterns alone. Dots and labels show $100(1-\overline{x}_{\text{after}}/\overline{x}_{\text{before}})$, using means over the same 100 participant splits with 50% calibration and four response bins. Whiskers show the 10th–90th percentiles of splitwise relative changes, not confidence intervals.

Prediction-to-bound regression. To assess whether larger predictive improvements help explain additional bound contraction within a task, we fit an exploratory regression with task intercepts to the same 500 paired split records. Let $r^L_{ts} = 100(1 - L^{\text{model}}_{ts}/L^{\text{context}}_{ts})$ and $r^W_{ts} = 100(1 - W^{\text{model}}_{ts}/W^{\text{patterns}}_{ts})$ denote the splitwise relative reductions in Brier loss and interval width. We fit

$$r^W_{ts} = \alpha_t + \beta(r^L_{ts} - \overline{r}^L_t) + \varepsilon_{ts},$$

using four response bins and 50% calibration. The estimated slope is $\widehat{\beta} = 0.185$, with partial $R^2 = 0.0172$ relative to task intercepts alone. A two-sided permutation test of within-task independence, using the common slope as its statistic, gives $p = 0.0068$ from 19,999 within-task permutations with the plus-one correction. Resampling paired split records within each task 10,000 times gives a percentile 95% interval of $[0.077, 0.300]$ for the slope.

This inference conditions on the observed participants and items: each repetition uses a separate seed for partitioning and calibration, so the resampling quantifies uncertainty over those randomized repetitions, not population sampling uncertainty. It establishes neither a causal effect of model quality nor transfer to new participants or tasks. In 20 repetitions of task-stratified five-fold validation over split records, the linear regression reduces prediction squared error by 1.40% relative to task intercepts alone; adding a quadratic term or allowing task-specific slopes gives 1.04% and 3.59%, respectively. These checks concern new partitions of the same datasets and support limited predictive usefulness under this protocol despite the positive association.

Calibration size and representation checks. Table 9 places calibration-fraction and binning sensitivity together. At matched calibration fractions, direct mean bounds in Census are slightly wider than model bounds at 25% (60.15 versus 59.33 points), but narrower at 50% and 75%. Census retains a larger model contribution but has low joint validity, especially at the smallest calibration size. At two, four, and eight bins, its extra contraction is 8.15, 6.33, and 2.02 points; dermatology gains 2.23 points only at two bins. All resolutions use identical half splits and summary comparators;

Table 8: **Task-level model contributions and their checks.** All panels use the same 100 half-participant splits within each task. (a) Joint validity includes all summary and model restrictions; inclusion concerns only the gain. No determined sign is wrong. (b) Ablations retain summaries and the other component allowance; widths are in percentage points. (c) Conditional-response Brier loss (lower is better) and evaluation mass in columns absent from calibration.

(a) Interval diagnostics and joint validity

Task	Model gain inclusion	Joint validity	Sign determined	Zero-error incompatible
Art	100%	92%	23%	56%
Census	99%	63%	2%	69%
Cities	100%	81%	9%	88%
Dermatology	100%	81%	0%	42%
Sarcasm	100%	72%	0%	1%
Object recognition	100%	89%	57%	100%

(b) Component attribution within binary tasks

Binary task	Summaries	Remove activation	Remove integration	Both components
Art	30.81	30.81	30.81	30.80
Census	63.46	57.34	63.46	57.13
Cities	48.91	48.00	48.91	47.60
Dermatology	24.57	24.57	24.57	24.57
Sarcasm	29.31	29.31	29.31	29.30

(c) Prediction and unsupported states

Task	Fitted kernel	Context baseline	Unchanged baseline	Unseen initial mass	Unseen active mass
Art	0.355	0.546	0.646	0.00%	0.00%
Census	0.272	0.451	0.468	0.04%	0.07%
Cities	0.369	0.537	0.658	0.00%	0.00%
Dermatology	0.278	0.590	0.309	0.22%	0.52%
Sarcasm	0.290	0.539	0.440	0.00%	0.00%

four bins remain the prespecified primary representation. Changing calibration fraction also changes evaluation size. Saved checks reconstruct raw task counts, verify nesting and validity, and match comparators across bin counts for 20,000 intervals from 2,500 binary-task splits; source and protocol hashes accompany the outputs.

Pooled HAIID reference. The secondary pooled protocol uses 100 task-stratified splits per fraction with seed 20260925, the same bootstrap refitting rule, and $0.05/8$ per tail. Its summary constraints and budgets target the pooled population; these values do not replace task-level results. Zero-error fits fix supported components, and single-component fits remove the other restriction.

Table 10 combines the bound comparison and calibrated-budget checks. The calibrated model and either component leave the half-split summary width unchanged at 22.16 points. This also holds at 25% and 75%, with sign-determination rates $22\%, 15\%, 11\%$ and no wrong signs. All calibrated intervals are feasible and include the gain, but all summary and model restrictions jointly hold in only 89% of half splits. Zero-error fits are incompatible in 299 of 300 splits; the sole feasible width is 0.057 points and misses the gain. This checks compatibility with finite empirical frequencies, rather than rejecting the population behavioral theory. On the same half splits, the refreshed direct branch has mean width 24.56 points and determines the sign in 6% of splits; all its intervals contain the evaluation gain.

Prediction and representation. At the half split, the fitted four-bin kernel has Brier loss 0.331, versus 0.536 for a context-only final distribution and 0.570 for no response change. Activation loss is 0.215 versus 0.239 for a task constant; activated-destination loss is 0.439 versus 0.543 for a context-only destination distribution. Baselines are calibrated or fixed in advance. These retrospective predictions condition on initial responses and truth-coded context, so better prediction does not establish latent identification or deployable outcome prediction. Unsupported states use activation

Table 9: **Task-level sensitivity to calibration size and response binning.** Each row averages 100 splits; widths are in percentage points. (a) Both partition sizes change; direct widths use finite-population calibration, while the last two columns describe the bootstrap-calibrated model. (b) Bin counts use identical half splits and summary comparators. No representation is selected using evaluation outcomes.

(a) Calibration fractions, four bins

Task	Calibration	Summary width	Model width	Direct width	Gain included	Joint validity
Art	25%	33.38	33.38	14.46	100%	85%
Art	50%	30.81	30.80	10.28	100%	92%
Art	75%	34.98	34.96	11.37	100%	96%
Census	25%	66.48	59.33	60.15	98%	37%
Census	50%	63.46	57.13	44.74	99%	63%
Census	75%	74.68	66.04	47.41	100%	85%
Cities	25%	51.91	49.80	40.32	98%	61%
Cities	50%	48.91	47.60	28.85	100%	81%
Cities	75%	55.19	52.98	30.53	99%	95%
Dermatology	25%	24.49	24.47	89.26	98%	59%
Dermatology	50%	24.57	24.57	76.97	100%	81%
Dermatology	75%	26.87	26.87	85.89	99%	90%
Sarcasm	25%	29.92	29.92	24.76	100%	69%
Sarcasm	50%	29.31	29.30	17.31	100%	72%
Sarcasm	75%	32.59	32.58	18.53	100%	81%

(b) Response bins, half-participant calibration

Task	Bins	Summary width	Model width	Gain included	Joint validity
Art	2	30.81	30.76	100%	91%
Art	4	30.81	30.80	100%	92%
Art	8	30.81	30.81	100%	91%
Census	2	63.46	55.31	99%	63%
Census	4	63.46	57.13	99%	63%
Census	8	63.46	61.44	99%	63%
Cities	2	48.91	48.01	100%	81%
Cities	4	48.91	47.60	100%	81%
Cities	8	48.91	48.78	100%	76%
Dermatology	2	24.57	22.34	100%	90%
Dermatology	4	24.57	24.57	100%	81%
Dermatology	8	24.57	24.57	100%	72%
Sarcasm	2	29.31	27.87	100%	73%
Sarcasm	4	29.31	29.30	100%	72%
Sarcasm	8	29.31	29.31	100%	70%

0.5 and uniform destinations only for scoring. Repeating the same half splits with two and eight bins still yields no calibrated contraction (Table 11a). Transport of these coarsened kernels requires stable within-bin initial mixtures, beyond a fixed continuous mechanism.

Tighter budgets and continuous support. Table 11b multiplies both budgets by $\lambda \in \{0, 0.25, 0.5, 0.75, 1\}$ on the same half splits, leaving summaries fixed. This is exploratory; $\lambda = 1$ remains primary. At 0.25, width falls to 19.28 points and every interval contains the gain, yet neither component is valid for the revealed evaluation behavior in any split. Panel (c) checks the original slider geometry on all 17,973 AI-presented rows, including midpoints. Only 78.75% place the final response between the initial response and advice: 8.94% move away, 12.27% overshoot, and six trials change despite initial agreement. A tolerance of 0.02 still leaves 15.56% outside the segment. Imposing convex support narrows absolute-loss width from 2.000 to 0.660 and squared-loss width from 4.00 to about 1.35 (full-data squared-loss reduction 0.1983). These are restrictions on the continuous response, separate from the discrete fitted model. Together the budget and geometry checks show why narrower bounds require checking the behavior they exclude.

E.3.2 Tejeda: Initial Accuracy, Error Allocation, and Response Policy

Fit and prediction. We fit the correctness, confidence, and adoption components of Tejeda et al. (2022) to the 14,206 sequential pairs. Initial judgments and confidence of fitting participants supply no-advice observations; no concurrent data or cross-workflow parameter link is used. Penalized

Table 10: **Pooled HAIID components: contraction and validity.** (a) Full-sample reference and 100 half-participant splits. Model rows retain the same bootstrap-calibrated summaries; direct correctness uses separate finite-population calibration. Reference components have zero error. (b) Mean probability-mass budgets and revealed component validity across calibration fractions. Width conditions on feasibility; empty sets fail inclusion.

(a) Bounds and component ablations

Information	Reference Width (pp)	Held-out participants Width (pp)	Incompatible	Inclusion
Support allowance	84.68	85.37	0%	100%
+ Propensity	29.98	35.10	0%	100%
+ Selection	9.20	22.16	0%	100%
Zero-error fitted model	0.00	0.06	99%	0%
Activation only	9.20	22.16	0%	100%
Integration only	9.20	22.16	0%	100%
Both components	0.00	22.16	0%	100%
Direct correctness	0.00	24.56	0%	100%

(b) Budgets and component validity

Calibration fraction	ε_A	ε_I	Activation valid	Integration valid	Gain included
25%	0.0903	0.0628	98%	91%	100%
50%	0.0799	0.0568	97%	95%	100%
75%	0.0926	0.0665	100%	100%	100%

Table 11: **Pooled HAIID checks of representation and imposed behavior.** Panels (a,b) use the same 100 half splits and summary constraints; widths are in percentage points. "Fixed" means exact supported components; multiplier one is the calibrated model. Panel (c) uses original continuous responses, including midpoints, and absolute-loss units; convex share is the fraction satisfying segment support.

(a) Response binning

Bins	Summary width	Model width	Model inclusion	Fixed incompatible	Fixed inclusion
2	22.16	22.16	100%	29%	0%
4	22.16	22.16	100%	99%	0%
8	22.16	22.16	100%	100%	0%

(b) Exploratory budget tightening

Multiplier	Width (pp)	Incompatible	Gain included	Activation valid	Integration valid
0.00	0.06	99%	0%	0%	0%
0.25	19.28	0%	100%	0%	0%
0.50	22.16	0%	100%	1%	0%
0.75	22.16	0%	100%	68%	21%
1.00	22.16	0%	100%	97%	95%

(c) Continuous support and absolute-loss bounds

Task	Trials	Convex share	Full-data $\Delta_{\lvert\cdot\rvert}$	Unrestricted width	Convex width
All AI	17,973	0.788	0.0918	2.000	0.660
Art	8,890	0.770	0.1141	2.000	0.637
Census	1,376	0.867	0.0324	2.000	0.901
Cities	2,528	0.669	0.0952	2.000	0.567
Dermatology	480	0.769	0.0916	2.000	0.705
Sarcasm	4,699	0.863	0.0653	2.000	0.678

likelihood replaces the original concurrent-study MCMC estimator. The first stage jointly fits $q_{ij} = \text{logit}^{-1}(s_j b_i - d_j)$ and ordered-probit confidence (Appendix A.4), with penalties $b_i^2/2$ and $(s_j - 1)^2/2$, difficulty bounds $[-10, 10]$, discrimination $[0.001, 10]$, cutpoint and gap fractions $[0.001, 0.999]$, and precision $[0.05, 15]$. The second stage fits person-specific adoption logits for three confidence levels and four AI-confidence bins (cutoffs $0.35, 0.57, 0.78$), with classifier-specific group logits. Logits increase with confidence following the published order. Person-to-group penalty scale is fixed at one, group prior variance at ten, base logits lie in $[-10, 10]$, and ordered increments in $[0, 10]$.

A classifier-specific uniform lapse $\ell_k \in [10^{-5}, 1 - 10^{-5}]$ retains third answers in the likelihood:

$$k_\psi^{\text{lap}}(f \mid h, r, z) = (1 - \ell_k)\big[(1 - \alpha_\psi(i, r, c, k))\mathbf{1}\{f = h\} + \alpha_\psi(i, r, c, k)\mathbf{1}\{f = \hat{y}_{\text{a}}\}\big] + \ell_k/16.$$

Zero lapse recovers the original kernel. Lapse differs from observed support violation because it can also return the initial answer or advice. For new people, predictions average the joint fitted person parameters within classifier condition and integrate all three confidence levels, preserving the dependence between correctness and adoption.

Joint-law bounds and attribution. Retain all 16 labels, recoding truth as 0, incorrect advice as 1 when applicable, and other labels alphabetically. Context crosses classifier, AI-confidence bin, and advice correctness (24 cells). Predictions average over actual target items within each context; target initial judgments and confidence are unavailable to fitting and calibration. Let m_{zfh} be normalized initial/final masses, $h_{zh} = \sum_f m_{zfh}$, and $w_z = \sum_h h_{zh}$. Observed frequencies fix $\sum_h m_{zfh} = p_{zf}$. For fitted correctness q_z^0 and final kernel K_{zfh}^0, add

$$\begin{aligned} Q(m) &:= \sum_z |h_{z0} - q_z^0 w_z| \le \varepsilon_Q, \\ E(m) &:= \tfrac{1}{2} \sum_z \sum_{h \neq 0} \left| h_{zh} - \frac{w_z - h_{z0}}{15} \right| \le \varepsilon_E, \\ P(m) &:= \tfrac{1}{2} \sum_{z,f,h} |m_{zfh} - K_{zfh}^0 h_{zh}| \le \varepsilon_P. \end{aligned} \tag{39}$$

Q carries initial-accuracy information from auxiliary judgments, since $|q_{\text{pre}} - \sum_z q_z^0 w_z| \le \varepsilon_Q$; E constrains error allocation and P the confidence-averaged policy including lapse. Label-independent adoption holds before averaging heterogeneous mechanisms and is not reimposed after pooling. Linear optimization of final minus initial correctness, retaining Eq. 37, gives sharp bounds for this convex discrepancy class, not for the nonlinear IRT parameter family or a posterior credible region. Removing all components numerically recovers the summary-only interval.

Independent calibration and results. The 100 classifier-stratified splits per fraction use seed 20260927. Roughly two thirds of outer calibration participants fit the model; the remainder calibrate discrepancies, with at least two people per classifier in each part. Holding the model fixed, 1,000 participant-bootstrap samples from the inner calibration group match target group sizes and recompute context mixtures. Upper discrepancy quantiles set Q, E, P allowances; each of six summary and three discrepancy tails receives $0.05/9$. All outer calibration participants supply the summary ranges and, through the separate finite-population construction, the direct-correctness ranges. Thus fitting, discrepancy calibration, and evaluation use disjoint people. The full-sample reference instead supplies observed fitted discrepancies; the zero-error row sets them to zero.

Table 12 combines component ablations, validity, and calibration-fraction checks. At 50%, the model narrows the summary width from 81.07 to 19.70 points. Removing Q restores 81.07; removing E or P leaves 19.70. Every calibrated interval includes the gain, but the full set of summary and model constraints holds for the revealed joint law in 89% of splits. Fixing all components exactly is incompatible in every half split. Panel (c) shows similar model contraction at the other fractions, while direct measurement gives wider mean bounds and determines no sign at all three reported fractions. The ablation attributes the model's contraction to its initial-accuracy restriction; the comparison with direct measurement also reflects the different calibration procedures.

Checks of the unrelaxed model. Table 13 reports three checks on the same sequential pairs. First, 1,886 third answers violate strict keep-or-adopt, motivating the lapse extension. Second, among

Table 12: **Tejeda fitted components: attribution, validity, and calibration size.** (a) Full-sample references allow observed discrepancies, except the zero-error row; held-out values use 100 half splits. A dash denotes infeasibility. (b) Mean probability-mass budgets and revealed component validity. (c) Widths in percentage points and sign rates across calibration fractions. Model rows retain all bootstrap-calibrated summaries; direct correctness uses separate finite-population calibration. Width conditions on feasibility; empty sets fail inclusion and sign determination.

(a) Bounds and component ablations

Information	Reference Width (pp)	Held-out participants Width (pp)	Incompatible	Inclusion
Support allowance	89.73	95.19	0%	100%
+ Propensity	60.29	81.07	0%	100%
+ Selection	58.76	81.07	0%	100%
Zero-error fitted model	–	–	100%	0%
Remove initial-accuracy constraint	51.52	81.07	0%	100%
Remove uniform-error constraint	5.56	19.70	0%	100%
Remove response-policy constraint	5.56	19.70	0%	100%
All three components	5.56	19.70	0%	100%
Direct correctness	0.00	75.00	0%	100%

(b) Budgets and component validity

Calibration	ε_Q	ε_E	ε_P	Q valid	E valid	P valid	Inclusion
25%	0.100	0.190	0.331	95%	100%	100%	100%
50%	0.098	0.170	0.320	98%	100%	100%	100%
75%	0.104	0.177	0.326	97%	100%	99%	100%

(c) Calibration-fraction sensitivity

Calibration	Summary width	Full-model width	Direct width	Model sign	Direct sign
25%	81.71	20.07	86.55	53%	0%
50%	81.07	19.70	75.00	57%	0%
75%	83.85	20.89	81.94	58%	0%

7,139 initial errors, uniform error allocation implies expected true-to-wrong-label counts $n_y/15$. The likelihood-ratio statistic $2\sum_{y,h\neq y} n_{yh}\log[n_{yh}/(n_y/15)]$, omitting zero terms, is compared with draws allocating each true class's fixed error count uniformly across 15 labels. Its total-variation distance is descriptive, not corrected for sampling noise.

Third, on 5,326 trials with incorrect advice and initial disagreement, the original policy implies equal adoption across initial labels conditional on participant, initial confidence, and AI-confidence bin. Conditioning on correctness and adoption margins leaves 285 informative strata (3,545 trials). Observed minus expected initially-correct adoptions sum to -120.75; a hypergeometric score tests the deviation. Raw adoption is 13.07% for initially correct answers and 41.28% for other errors. Third answers remain nonadoptions (24.00% of this subset), so the check cannot distinguish lapse from label-dependent adoption. Both conditional Monte Carlo tests use 9,999 draws and plus-one probabilities.

Finally, 100 classifier-stratified half splits compare uniform wrong-label prediction with probabilities fitted by true class using a half-count per incorrect label. Fitted conditional-on-error log loss improves in all splits. This predicts error destinations given truth and an error, without estimating initial accuracy or supplying a cognitive-model bound.

Table 13: **Tejeda checks of unrelaxed restrictions.** Monte Carlo probabilities use 9,999 conditional draws and the plus-one correction. Wrong-label prediction is evaluated on separate participants. These diagnostics do not themselves supply a bound.

Check	Value	Evaluation basis
Keep-or-adopt violations	13.28%	1,886 / 14,206 paired trials
Uniform-error likelihood ratio	2127.53	Conditional Monte Carlo $p = 0.0001$
Wrong-label allocation TV	0.216	Conditional on an initial error
Conditional policy score	-11.96	Conditional Monte Carlo $p = 0.0001$
Uniform wrong-label log loss	2.708	Held-out initial errors
Fitted allocation log loss	2.616	100 participant splits; lower is better

E.4 Direct Measurement: Calibration Budgets, Precision, and Prior Sensitivity

Participant measurement budgets. To evaluate how much direct measurement is needed, we vary the number of calibration participants while retaining the participant-disjoint evaluation design. The main sweep requests $1\%, 2\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%$ of participants, with 100 random partitions per fraction and at least two people per stratum in each partition. All of a participant's trials remain together. The smallest measured groups contain two people in each binary task and six in object recognition, which has three classifier strata. Figure 2 uses the actual measured percentages; fractions giving the same participant allocation are represented by the largest requested fraction, independently of the results. Evaluation uses the remaining participants, so both its size and its trial-weighted gain change with the partition. The 25% and 75% appendix comparisons preserve their original partitions as well.

Finite-sample uncertainty from measured participants. Very small calibration samples can give degenerate empirical bootstrap distributions, so all reported direct branches instead use the Hedged-WoR-CI construction of Waudby-Smith & Ramdas (2024) (Section 5.2, Theorem 4 and Remark 4, with $c = 1/2$). Within stratum s, let N_s denote the total number of participants before splitting. Normalize each participant's initial-correct count C_i by the known maximum trial count M_s, giving bounded observations $X_i = C_i/M_s \in [0, 1]$. A uniformly sampled calibration group, independently randomized in order, then supplies confidence bounds on the finite-population mean. Multiplying those bounds by $N_s M_s$ and subtracting the measured correct counts gives bounds on the remaining participants' correct totals. Intersecting with their possible trial counts, summing across strata, and dividing by the total evaluation trial count preserves the trial-weighted target even when participant trial counts differ. Apply the same construction to support-violation counts to obtain a separate upper allowance. The joint-mass program then retains these correctness and support restrictions, replacing propensity, selection, and model constraints. Only calibration outcomes, public trial counts, and stratum membership enter calibration; evaluation initial judgments supply diagnostics afterward.

For H strata, each correctness tail and each support upper tail receives $\delta = 0.05/(3H)$. Their joint event has probability at least 95% under random participant sampling and ordering, conditional on the fixed released dataset. On that event the evaluation joint law is feasible in the program, so its gain is contained in the reported interval. This guarantee is per budget, does not require independent trials within participants, and is neither simultaneous across the budget curve nor a guarantee for another population or workflow. Behavioral intervals retain their original bootstrap calibration; they do not inherit this guarantee. Matching participant partitions therefore does not equalize uncertainty constructions, retained information, or elicitation burden.

Results and checks. Table 14 gives the 50% endpoint of the main comparison. In task order Art, Census, Cities, Dermatology, Sarcasm, and object recognition, direct mean widths are 10.28, 44.74, 28.85, 76.97, 17.31, and 75.00 points; sign-determination rates are $100\%, 0\%, 5\%, 0\%, 24\%, 0\%$. The main sweep contains $4{,}800$ intervals at 38 distinct task/allocation combinations. Across all protocols, $6{,}900$ intervals include the $25\%/75\%$ and pooled comparisons. Every interval is feasible and contains its evaluation gain, with no wrong signs; both direct restrictions hold for every revealed evaluation joint law. These empirical frequencies do not imply 100% coverage beyond the observed splits. At every main task's minimum allocation the width is 100 points: identical measurements from two people do not imply zero uncertainty. The implementation reproduces the original participant counts, trial totals, and evaluation gains; exact small-population checks also verify coverage and the conversion for unequal trial counts. The same direct construction is used in Tables 7, 9, 10, and 12.

Direct mean widths increase from 50% to 75% requested calibration in all six tasks; more calibration participants also leave fewer participants in the evaluation sample.

Table 14: **Direct measurement within each task.** Cells report mean width in percentage points and evaluation-gain inclusion over the same 100 participant partitions at the 50% requested calibration fraction. Direct correctness uses finite-population calibration of correctness and support; summaries and models retain bootstrap calibration. All intervals are feasible. This table gives the 50% endpoint of Figure 2.

Task	Summary constraints Width / inclusion	Fitted model Width / inclusion	Direct correctness Width / inclusion
Art	30.81 / 100%	30.80 / 100%	10.28 / 100%
Census	63.46 / 99%	57.13 / 99%	44.74 / 100%
Cities	48.91 / 100%	47.60 / 100%	28.85 / 100%
Dermatology	24.57 / 100%	24.57 / 100%	76.97 / 100%
Sarcasm	29.31 / 100%	29.30 / 100%	17.31 / 100%
Object recognition	81.07 / 100%	19.70 / 100%	75.00 / 100%

One reveal design, two checks. We now consider sparse recording within a fixed sample, separating sampling precision from participant-transfer interval widths. For each of 5,000 random reveal designs on the 17,521 HAIID trials, retain a simple random subset of initial-correctness bits, estimate q_{pre} from it, and subtract from the fully observed final accuracy. Table 15 reports design-based error conditional on this dataset: 5% revelation supplies 876 bits and gives RMSE 0.0154. Equation 32 explains how this error falls with the reveal count and reaches zero under full revelation.

Table 15: **Precision from sparse correctness measurement.** Errors over 5,000 random reveal designs on the fixed HAIID sample, on the probability scale.

Reveal fraction	Revealed correctness bits	RMSE	Median absolute error
0.5%	88	0.0497	0.0394
1%	175	0.0348	0.0243
2%	350	0.0244	0.0157
5%	876	0.0154	0.0103
10%	1,752	0.0104	0.0069
20%	3,504	0.00695	0.00461
50%	8,760	0.00348	0.00244

For the prior check, truncate three Beta priors for q_{pre} to the compatible interval $[0.0676, 0.9077]$, with untruncated means $0.55, 0.70, 0.85$ and common effective sample size 20. These illustrative priors are unchanged by final-only likelihood (Proposition 6); revealed correctness bits supply a Bernoulli likelihood. Figure 7 averages posterior means over the same reveal design. Both estimation error and between-prior spread fall as initial correctness is observed. This exercise studies missing information within the recorded workflow, not the behavioral effect of eliciting a response.

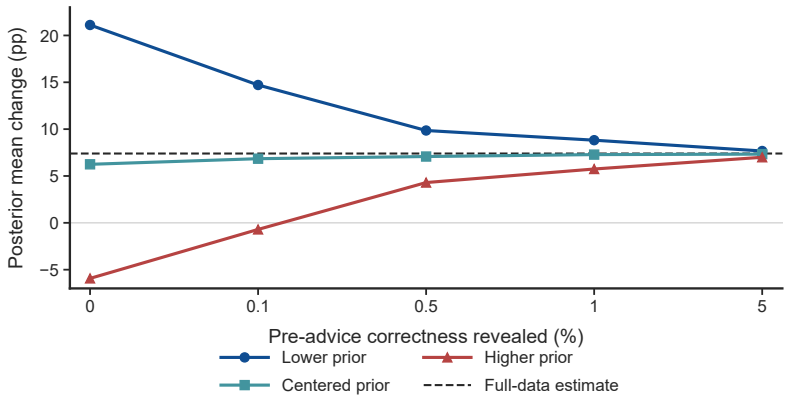


Figure 7: **The same correctness measurements reduce prior sensitivity.** Posterior mean gain under three illustrative priors, averaged over 5,000 HAIID reveal designs. The dashed line is the full-data gain, 7.4 points.

F CLAIM-LEVEL LITERATURE AUDIT

This appendix records the claim–workflow units, the codebook, the two independent coding passes, and the exact final-agreement fractions used to quantify the baseline identified-set width. The audited papers are listed in the bibliography.

F.1 CORPUS AND UNIT OF ANALYSIS

The study set is the 56 empirical papers in Raees et al. (2026). A complete pre-advice judgment $\hat{y}_{\mathrm{h}}$ is coded only when the participant gives a same-trial, independently interpretable answer in the same decision domain before seeing the AI. Confidence reports, partial responses, gaze traces, and unaided answers on different cases do not count. Figure 4 reports claim-level identification status. Separately, 29 of the 56 papers contain at least one assisted workflow with no complete same-trial $\hat{y}_{\mathrm{h}}$.

The unit of analysis is the claim–workflow instance, not the paper. Within a paper we keep one instance for each distinct combination of workflow or condition, reported target, and identifying information; repeated restatements of the same claim are collapsed. Each instance is coded on five axes: claim family, availability of complete same-trial $\hat{y}_{\mathrm{h}}$, primary identifying information, identification status, and operational versus latent interpretation. Source locations are stored for every row. This rule yields 96 instances. The counts compare information requirements across claim types; they are not estimates of field-wide prevalence.

Claim-identification rule. To compare claims with different targets, we use one definition of identification (Guo et al., 2024; Agarwal et al., 2025; Guo et al., 2026). Let o be the record relevant to a claim and p_o its distribution. A target $\theta(q)$ is a functional of a fuller description q of the study, which may include latent human states or counterfactual outcomes. Information from the study design, from human-side measurements, and from maintained behavioral constraints, denoted $\mathbf{E}_{\mathrm{design}}$, $\mathbf{E}_{\mathrm{meas}}$, and $\mathbf{E}_{\mathrm{beh}}$, limits which descriptions are admissible. The compatible descriptions and their identified set are

$$\begin{aligned}&\mathbf{C}(p_o; \mathbf{E}_{\mathrm{design}}, \mathbf{E}_{\mathrm{meas}}, \mathbf{E}_{\mathrm{beh}})\\ &\quad := \{q : q \text{ induces } p_o \text{ and is compatible with } \mathbf{E}_{\mathrm{design}}, \mathbf{E}_{\mathrm{meas}}, \mathbf{E}_{\mathrm{beh}}\},\\ &\mathbf{I}_\theta(p_o; \mathbf{E}_{\mathrm{design}}, \mathbf{E}_{\mathrm{meas}}, \mathbf{E}_{\mathrm{beh}})\\ &\quad := \{\theta(q) : q \in \mathbf{C}(p_o; \mathbf{E}_{\mathrm{design}}, \mathbf{E}_{\mathrm{meas}}, \mathbf{E}_{\mathrm{beh}})\}.\end{aligned} \tag{40}$$

The identified set collects every value of the target produced by some admissible description that reproduces the record. A reported interval is *sharp* when it equals this set, so that every value in it is attained, and the target is point identified when the set is a single value. New data, design information, or behavioral knowledge therefore matter only insofar as they exclude joint distributions that remain compatible with the observed data.

F.2 INDEPENDENT CODING AND ADJUDICATION

The first author coded all 96 instances and checked the first pass against the original papers. A second author independently coded the same instances on five axes, without the first-pass labels or aggregate counts. Two axes began as free-text notes (primary identifying information and identification status). After the second pass was finished and before the two files were compared, both passes were mapped onto the shared codebook in Table 16. Table 17 reports agreement on that codebook, before adjudication. The authors resolved disagreements against the source papers; all reported counts use the consensus labels.

Table 16: **Shared codebook for the five coding axes.** Values are mutually exclusive on each axis.

Axis	Value	Coded when
Claim family	Perf. / Assist. effect / Switch/change / Operational / Latent mech. / Outside	The reported target is final accuracy or team performance; an effect of offering assistance; same-person revision; observed agreement, override, or following; a latent construct such as overreliance or anchoring; or none of these.
Same-trial $\hat{y}_{\mathrm{h}}$	Complete / None / Partial/aux. / N/R	The assisted workflow records an independent same-trial pre-advice judgment; records none; records only a proxy or a judgment from another case; or $\hat{y}_{\mathrm{h}}$ is irrelevant to the stated target.
Primary source	Recorded / Design / Direct $\hat{y}_{\mathrm{h}}$ / Model / Aux.+model	The identifying information is final-only data; an assigned or randomized condition; the same-trial $\hat{y}_{\mathrm{h}}$; a maintained behavioral model; or auxiliary unaided data plus a model.
Identification status	Data-ID / Design-ID / Op.-ID, latent unresolved / Model-rel. / Outside	The stated claim is pinned down by the recorded data; by the design; only as an operational statistic, with a stronger latent reading left open; only under a behavioral model; or outside the audit.
Op./latent	Operational / Latent / Both	The paper reports an observed statistic; a latent construct; or both.

Table 17: **Inter-coder agreement after freezing the revised codebook, before adjudication.** Both authors independently recoded all 96 claim–workflow instances using the frozen definitions.

Coding axis	Agreement	Agreement rate	Cohen's κ
Claim family	88/96	91.7%	0.889
Same-trial pre-advice judgment	90/96	93.8%	0.889
Primary identifying information	87/96	90.6%	0.861
Identification status	86/96	89.6%	0.834
Operational vs. latent	91/96	94.8%	0.845

For the headline binary result, Data-ID and Design-ID were grouped as claims supported by recorded data or study design, whereas Op.-ID/latent-unresolved and Model-rel. were grouped as claims requiring additional information or assumptions for the stronger interpretation. C070 was excluded from this binary comparison because both coders independently classified it as outside the audit scope. On the remaining 95 instances, the coders agreed on 89 (93.7%), yielding Cohen's $\kappa = 0.804$. The six disagreements were adjudicated only after the independent recoding was complete. This binary agreement is distinct from the finer five-category identification-status agreement reported in Table 17 (86/96, 89.6%, $\kappa = 0.834$): disagreements between Data-ID and Design-ID count as disagreements in the five-category coding but as agreement in the headline binary, and C070 is retained as Outside in the former but excluded from the latter.

F.3 CONSENSUS IDENTIFICATION RESULTS

Table 18: **Final consensus identification status for the 96 audited claim–workflow instances.** Categories refer to identification of the stated claim. The first two rows sum to the 80 instances whose targets are identified from recorded data or study design.

Claim-level status	Instances
Identified from recorded data	49
Identified by randomized or assigned study design	31
Operational quantity identified; stronger latent interpretation unresolved	12
Latent target identified only under a behavioral model	2
Outside the present audit target	2
Total	96

The coded source of identifying information provides a complementary view. Recorded observables and study design account for 70 of the 96 instances, direct same-trial $\hat{y}_{\mathrm{h}}$ accounts for 24, and only two instances rely primarily on model-based information. Direct $\hat{y}_{\mathrm{h}}$ is separated from other recorded variables because it can reveal the individual decision path rather than only an aggregate outcome. The detailed consensus table below carries the complete instance-level coding.

The distinction is visible within individual papers. Bansal et al. (2021), for example, report identifiable final team performance and observed agreement; interpreting the latter as the level of "blind trust" additionally requires separating pre-advice agreement from switching to AI advice. In Rastogi et al. (2022), final agreement is observable and randomized timing comparisons remain valid, while interpreting agreement itself as AI-induced anchoring requires the missing pre-advice state. Wysocki et al. (2023) similarly report observable clinician–model concordance and correctness, whereas an automation-bias interpretation is a claim about how automated advice changed the human decision. We preserve each paper's operational definition: when reliance is defined as an agreement statistic, we treat that statistic as identified and distinguish it from stronger latent adoption or influence interpretations.

F.4 FULL CLAIM-LEVEL CODING

For Table 19, claim-family codes are: Perf. = final performance; Assist. effect = effect of assistance or an experimental condition; Switch/change = same-person revision; Operational = observed agreement/adherence/override/follow behavior; Latent mech. = latent reliance or influence mechanism; Outside = a residual target outside the five preceding families. $\hat{y}_{\mathrm{h}}$ codes are Complete, None, Partial/aux., and N/R (not relevant to the stated target). Primary-source codes are Recorded, Design, Direct $\hat{y}_{\mathrm{h}}$, Model, and Aux.+model. Status codes are Data-ID, Design-ID, Op.-ID/latent unresolved, Model-rel., and Outside. The final column distinguishes operational-only, latent-only, and mixed operational-plus-latent interpretations. For C008, we code the study's "appropriate trust" measure as final performance because it is defined as final-decision accuracy; C007 records its separate agreement measure.

Table 19: **Final consensus coding for all 96 claim–workflow instances.** Claim IDs link to the source records used for verification; abbreviations are defined immediately above.

ID	Study	Claim family	$\hat{y}_{\mathrm{h}}$	Primary source	Identification status	Op./latent
C001	Yin et al. 2019	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Op.-ID / latent unresolved	Both
C002	Zhang et al. 2020	Assist. effect	Complete	Design	Design-ID	Both
C003	Bansal et al.	Perf.	N/R	Recorded	Data-ID	Operational
C004	Bansal et al.	Operational	None	Recorded	Data-ID	Operational
C005	Grgić-Hlača et al.	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C006	Lu & Yin 2021	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C007	Wang & Yin	Operational	Complete	Recorded	Data-ID	Operational
C008	Wang & Yin	Perf.	Complete	Recorded	Data-ID	Operational
C009	Buçinca et al.	Assist. effect	N/R	Design	Design-ID	Operational
C010	Buçinca et al.	Operational	Partial/aux.	Recorded	Data-ID	Operational
C011	Buçinca et al.	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational

Continued on next page

ID	Study	Claim family	$\hat{y}_{\mathrm{h}}$	Primary source	Identification status	Op./latent
C012	Schemmer et al. 2023 (IUI)	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C013	Salimzadeh et al. 2024	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C014	de Jong et al. 2025	Assist. effect	N/R	Design	Design-ID	Operational
C015	de Jong et al. 2025	Operational	None	Recorded	Data-ID	Operational
C016	de Jong et al. 2025	Assist. effect	N/R	Design	Design-ID	Operational
C017	de Jong et al. 2025	Operational	None	Recorded	Data-ID	Operational
C018	de Jong et al. 2025	Assist. effect	N/R	Design	Design-ID	Operational
C019	Buçinca et al. 2025	Assist. effect	N/R	Design	Design-ID	Operational
C020	Buçinca et al. 2025	Operational	None	Recorded	Data-ID	Operational
C021	Buçinca et al. 2025	Operational	Complete	Recorded	Data-ID	Operational
C022	Swaroop et al. 2025	Latent mech.	None	Model	Model-rel.	Latent
C023	Swaroop et al. 2025	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C024	Swaroop et al. 2025	Assist. effect	N/R	Design	Design-ID	Operational
C025	Rastogi et al. 2022	Latent mech.	None	Recorded	Op.-ID / latent unresolved	Both
C026	Rastogi et al. 2022	Assist. effect	N/R	Design	Design-ID	Operational
C027	Wysocki et al. 2023	Latent mech.	None	Recorded	Op.-ID / latent unresolved	Both
C028	Wysocki et al. 2023	Perf.	N/R	Recorded	Data-ID	Operational
C029	Dikmen & Burns 2022	Assist. effect	N/R	Design	Design-ID	Operational
C030	Dikmen & Burns 2022	Latent mech.	None	Recorded	Op.-ID / latent unresolved	Both
C031	Schemmer et al. 2023 (ICIS)	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C032	Morrison et al. 2023	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C033	Morrison et al. 2024	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C034	Morrison et al. 2024	Assist. effect	N/R	Design	Design-ID	Operational
C035	Cao & Huang 2022	Operational	None	Recorded	Op.-ID / latent unresolved	Both
C036	Lee & Chew 2023	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C037	Lee & Chew 2023	Assist. effect	N/R	Design	Design-ID	Operational
C038	Cabitza et al. 2023	Latent mech.	Complete	Direct $\hat{y}_{\mathrm{h}}$	Op.-ID / latent unresolved	Both
C039	Ma et al. 2023	Operational	None	Recorded	Op.-ID / latent unresolved	Both
C040	Ma et al. 2023	Outside	Partial/aux.	Design	Outside	Operational
C041	Ma et al. 2023	Assist. effect	N/R	Design	Design-ID	Operational
C042	He, Kuiper & Gadiraju 2023	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C043	He, Buijsman & Gadiraju 2023	Assist. effect	Complete	Design	Design-ID	Operational
C044	Cau et al. 2023 (IUI)	Assist. effect	N/R	Design	Design-ID	Operational
C045	Cau et al. 2023 (IUI)	Latent mech.	None	Recorded	Op.-ID / latent unresolved	Both
C046	Cao, Gomez & Huang 2023	Assist. effect	Complete	Design	Design-ID	Operational
C047	Peng et al. 2024	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C048	Ma et al. 2024	Assist. effect	Complete	Design	Design-ID	Operational
C049	He, Bharos & Gadiraju 2024	Assist. effect	Complete	Design	Design-ID	Operational
C050	Kim et al. 2024	Assist. effect	N/R	Design	Design-ID	Operational
C051	Kim et al. 2024	Operational	None	Recorded	Data-ID	Operational
C052	Kim et al. 2024	Operational	None	Recorded	Data-ID	Operational
C053	Kim et al. 2024	Operational	None	Recorded	Data-ID	Operational
C054	Cabitza et al. 2025	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C055	He, Aishwarya & Gadiraju 2025	Assist. effect	Complete	Design	Design-ID	Operational
C056	Ma et al. 2025	Assist. effect	Complete	Design	Design-ID	Operational
C057	Lai & Tan 2019	Assist. effect	N/R	Design	Design-ID	Operational
C058	Buçinca et al. 2020	Assist. effect	N/R	Design	Design-ID	Operational
C059	Yang et al. 2020	Operational	None	Recorded	Op.-ID / latent unresolved	Both
C060	Nourani et al.	Assist. effect	N/R	Design	Design-ID	Both
C061	Poursabzi-Sangdeh et al.	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C062	Chiang & Yin 2022	Assist. effect	Complete	Design	Design-ID	Operational
C063	Brachman et al. 2022	Operational	None	Recorded	Data-ID	Operational
C064	Tejeda et al.	Latent mech.	Partial/aux.	Aux.+model	Model-rel.	Latent
C065	Tejeda et al.	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C066	Vasconcelos et al. 2023	Operational	None	Recorded	Data-ID	Operational
C067	Vasconcelos et al.	Assist. effect	N/R	Design	Design-ID	Operational
C068	Vered et al. 2023	Latent mech.	None	Recorded	Op.-ID / latent unresolved	Both
C069	Chen et al. 2023	Latent mech.	Complete	Direct $\hat{y}_{\mathrm{h}}$	Op.-ID / latent unresolved	Both
C070	Prabhudesai et al. 2023	Outside	N/R	Recorded	Outside	Latent
C071	Cabrera et al. 2023	Assist. effect	N/R	Design	Design-ID	Operational
C072	Cabrera et al. 2023	Assist. effect	N/R	Design	Design-ID	Both
C073	Cau et al. 2023 (TiiS)	Operational	None	Recorded	Op.-ID / latent unresolved	Operational
C074	Chiang et al.	Assist. effect	N/R	Design	Design-ID	Operational
C075	Chiang et al. 2023	Assist. effect	N/R	Design	Design-ID	Operational
C076	Cabitza et al. 2024	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C077	Cao et al. 2024	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C078	Pareek et al. 2024	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C079	Schoeffer et al. 2024	Operational	None	Recorded	Data-ID	Operational

Continued on next page

ID	Study	Claim family	$\hat{y}_{\mathrm{h}}$	Primary source	Identification status	Op./latent
C080	Swaroop et al. 2024	Operational	None	Recorded	Data-ID	Operational
C081	Swaroop et al. 2024	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C082	Lu, Wang & Yin 2024	Operational	None	Recorded	Data-ID	Operational
C083	Lu, Wang & Yin 2024	Assist. effect	N/R	Design	Design-ID	Operational
C084	Jeon et al. 2025	Switch/change	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C085	Jeon et al. 2025	Operational	None	Recorded	Data-ID	Operational
C086	Cau & Spano 2025	Assist. effect	N/R	Design	Design-ID	Operational
C087	Chiang et al. 2023	Operational	None	Recorded	Data-ID	Operational
C088	Chiang et al. 2023	Operational	None	Recorded	Data-ID	Operational
C089	Chiang et al. 2023	Outside	Complete	Direct $\hat{y}_{\mathrm{h}}$	Data-ID	Operational
C090	Schöffer et al. 2025	Operational	None	Recorded	Data-ID	Operational
C091	Schöffer et al. 2025	Operational	None	Recorded	Data-ID	Operational
C092	Schöffer et al. 2025	Operational	None	Recorded	Data-ID	Operational
C093	Lu, Wang & Yin 2024	Operational	None	Recorded	Data-ID	Operational
C094	Lu, Wang & Yin 2024	Operational	None	Recorded	Data-ID	Operational
C095	Lu, Wang & Yin 2024	Operational	None	Recorded	Data-ID	Operational
C096	Brachman et al. 2022	Operational	None	Recorded	Data-ID	Operational

F.5 RECOVERABLE FINAL-AGREEMENT FRACTIONS

The quantitative extraction targets $g = \Pr(\hat{y} = \hat{y}_{\mathrm{a}})$ for eligible final-only workflows. We accept trial-level public data, exact tables or numerical text, or algebraic reconstruction from exact conditional rates and denominators, and exclude plot-only values, post-treatment selected subgroups, and fitted marginal probabilities that do not equal raw agreement. To avoid condition-count weighting, the primary aggregation uses one median exact g per paper; the study/domain-balanced median is a sensitivity analysis and the condition-level pool is descriptive.

Under these rules, 13 papers yield exact paper-level g summaries, including 12 of the 21 papers with no complete $\hat{y}_{\mathrm{h}}$ in any assisted arm. Twelve papers contribute 49 separately identified condition values; one additional paper contributes an exact paper median from equal-block information without separately identified block values. Table 20 reports these sources. Its Tejeda row uses the concurrent final-only experiment; Section 4.1 analyzes the separate sequential experiment.

Table 20: **Sources for exactly recoverable final-agreement fractions.** Rows give the evidence used to recover paper-level g, the number of separately identified conditions, and the paper median. *Notes.* "Conditions" counts distinct final-only g values; Wysocki et al. contribute an exact paper median from two equal blocks without separately identified block values.

Study	Exact-g evidence	Conditions	Paper median g
Vered et al. (2023)	Exact Tables 6 and 12 with the fixed 6/4 wrong/correct-advice task mix	8	0.250
de Jong et al. (2025)	Exact reported reliance percentages across two studies	4	0.460
Schoeffer et al. (2024); values from Schöffer et al. (2025)	JAIR Table 2 reanalysis reports unconditional Study-2 adherence	3	0.596
Chiang et al. (2023)	Exact main-text overall-reliance means for individual and group AI-assisted conditions	2	0.665
Lu et al. (2024)	Supplement reports exact Overall Reliance means for three complete randomized treatments	3	0.670
Buçinca et al. (2020)	Exact conditional final accuracies plus the fixed 75/25 binary AI-correctness mix	2	0.708
Lai & Tan (2019)	Trust is defined as final agreement; Figure 4 prints five exact numeric labels	5	0.748
Rastogi et al. (2022)	Exact numerical text plus fixed binary confidence/correctness strata	1	0.761
Kim et al. (2024)	Exact reported agreement estimates for the three AI-access conditions	3	0.776
Tejeda et al. (2022)	Trial-level public data from the concurrent final-only experiment	3	0.776
Brachman et al. (2022)	Exact conditional reliance rates plus fixed AI accuracy reconstruct unconditional agreement	1	0.812
Wysocki et al. (2023)	Exact prose counts and equal five-case blocks identify the paper median	–	0.815
Bansal et al. (2021)	Trial-level public Complementary-Performance data across 14 assisted conditions	14	0.815

Table 20 carries the full paper-level distribution. Its median is 0.748 (published-rounding range $[0.7475, 0.7485]$), while the study/domain-balanced sensitivity median is 0.728. Source locations, extraction algebra, and exclusions are documented by the coding and extraction rules above.