CALIBRATED COARSENING: DESIGNING INFORMATION FOR AI-ASSISTED DECISIONS*

Bnaya Dreyfuss† Ruru Hoong‡

This version: July 23, 2026

Abstract

Artificial intelligence (AI) is increasingly used to aid human decision-making across critical applications, but human cognitive biases (e.g., errors in probabilistic reasoning) often undermine its effectiveness. This raises the question of how to provide AI input to humans in a way that improves decision-making outcomes. We propose *calibrated coarsening*—partitioning the signal space into fewer cells at a threshold calibrated to the decision environment rather than chosen arbitrarily—as a way to do so that (i) retains humans' final decision rights, (ii) modifies signals without deception, and (iii) improves outcomes under a range of empirically common biases. Within an information disclosure framework, we derive an approximately optimal binary coarsened policy when the designer does not observe the decision-maker's information. In an experiment with professional loan specialists ($N = 150$), we empirically demonstrate that coarsening AI signals at the theory-motivated threshold improves decision accuracy, outperforming both the human-only (based solely on the loan application) and uncoarsened AI assistance (continuous AI risk-score, in addition to loan application) benchmarks. We uncover substantial heterogeneity among specialists and estimate a Bayesian hierarchical latent-variable model (HLM), which we use to simulate and test the benefits of personalized coarsening.

JEL Codes: C91, C93, D82, D83, D91, O33.

*An earlier version of this paper was Ruru Hoong's job market paper; we are especially indebted to her advisors Katie Coffman, Shane Greenstein, Jesse Shapiro and David Yang for their support. We further thank Ophir Averbuch, John Beshears, Alex Chan, Fiona Chen, Benjamin Enke, Christine Exley, Kris Ferreira, Jeff Gortmaker, Matthew Gentzkow, Yannai A. Gonczarowski, Ayelet Israeli, Baiyun Jing, Andrew Kao, Chris Liao, Pëllumb Reshidi, Josh Schwartzstein, Bennett Smith-Worthington, Clare Suter, Satoru Takahashi, Chris Walker and the behavioral and labor workshops at Harvard for helpful comments. We thank Philipp Chapkovski and Kirill Odintsov for helpful assistance. We are grateful for funding from Harvard Business School and Harvard Economics, NUS Development Grant, and Social Science Research Council Graduate Research Fellowship (SSRC 2025-004) administered by the Ministry of Education, Singapore. Any opinions, findings and conclusions expressed in this material are those of the authors and do not reflect the views of funders. The experiments involved in this paper were exempted under Protocol #IRB24-0395 in April 2024. Refine.ink was used to check the paper for consistency and clarity.

†University of Pittsburgh. Email: bnaya3fuss@gmail.com

‡MIT Sloan School of Management. Corresponding author. Email: ruruh@mit.edu. E62-539, 100 Main Street, Cambridge, MA 02142

1 Introduction

Artificial intelligence (AI) has caught pace with—and in some contexts even surpassed—human capabilities in data-driven prediction, promising to improve decision making across an array of important fields, from medical diagnosis to managerial decision making (Agrawal et al., 2022). Yet real-world deployment of AI tools to support human decision-making has frequently yielded underwhelming results, including in high-stakes contexts like judicial sentencing and healthcare.[1]

A key reason for this lies in the critical role *humans* play. AI typically assists rather than replaces human decision-making, as oversight often remains necessary to implement final decisions for example due to institutional constraints like legal liability.[2] While AI predictions can offer valuable input, its impact on performance depends not only on its informativeness, but also on how humans integrate it with their own beliefs to make decisions.

It has been long documented that humans struggle to process and aggregate information: myriad biases and belief misspecifications distort belief updating (e.g., Tversky and Kahneman 1974; Benjamin 2019). It is therefore unsurprising that even informative AI predictions may sometimes fail to improve outcomes. This raises a natural question: how can we improve outcomes in AI-assisted decision-making? Several design challenges emerge. First, removing humans from the process is often not feasible, as they have to retain final decision authority. Second, transparency is critical, as deceptive strategies have been found to hinder long-term adoption (Mahmud et al., 2022). Third, solutions must be capable of addressing a range of biases, which likely differ across individuals and decision-making contexts.

We propose *calibrated coarsening*—i.e., partitioning information into fewer cells at a threshold calibrated to the decision environment rather than chosen arbitrarily[3]—as a broad approach to improving decision-making outcomes that: (i) retains humans' final decision rights, (ii) transparently modifies signals without deception, and (iii) can address a wide range of cognitive biases. Coarsening is already commonplace in information-provision settings, from discretized scores (e.g., 1,

[1] e.g., Algorithmic risk assessments have not improved sentencing outcomes (Stevenson and Doleac, 2024; Garrett and Monahan, 2020; Imai et al., 2023), nor shown clear benefits in clinical and radiological decision-making (Brocklehurst et al., 2017; Nunes et al., 2017; Gaube et al., 2021; Agarwal et al., 2023).

[2] e.g., Clinicians remain liable in clinical settings (Greenes, 2007; Mezrich, 2022; Cestonaro et al., 2023); recent regulations—e.g., EU's AI Act—mandate human oversight in algorithmic decision-making (European Commission, 2024). Human involvement mitigates litigation risks surrounding disparate impact, where a facially neutral procedure disadvantages a protected class per Title VII of the Civil Rights Act (Kassir et al., 2023).

[3] This is the sense in which we use "calibrated" throughout. We focus on *binary* coarsening (a single threshold). A further layer of calibration—personalizing the threshold to the individual—is the subject of our second experiment, where we also allow finer, trinary partitions.

2, 3) to color-coded risk (e.g., green vs. red),[4] demonstrating its intuitive appeal over raw continuous algorithmic outputs. Yet it is often implemented heuristically rather than systematically. Our contribution is to provide a principled framework for the *calibrated* implementation of this familiar practice to improve outcomes.

A key advantage of our approach is that it does so in a context- and bias-agnostic way: AI predictions can be coarsened—calibrated to the decision environment's cost structure—and recoarsened as objectives change, without needing to identify which bias is at play. When more is known about individual behavior, further calibration is possible.

In this paper, we develop a theoretical framework and propose an approximately optimal policy within the *binary* class—our focus throughout—and empirically test it in an experiment with 150 loan officers, underwriters, and processors (henceforth, "loan specialists") with an average of 12 years of industry experience, recruited through professional channels. In all decision blocks, participants review loan applications and make approval decisions; decision blocks vary in whether they additionally receive an AI signal, and if so, whether it is coarsened and how. We find that coarsening AI signals at the approximately optimal threshold improves approval accuracy, outperforming both human-only (based solely on the application) and continuous AI (assisted with uncoarsened AI risk-score) benchmarks, and two binary benchmarks set at arbitrary (suboptimal) thresholds. It also reduces time-to-decision, and many participants prefer—and are willing to pay for—coarsened over full signals. We further document substantial heterogeneity in specialist behavior, moderated by both demographics and elicited bias measures. This heterogeneity motivates adding another layer of calibration to coarsening on top of choosing a non-arbitrary threshold: personalizing that threshold to individuals based on their observed choices, when feasible. We explore the scope for such personalization by estimating a Bayesian HLM on past data, predicting modest gains of ~ 2 percentage points. In a two-stage experiment, we explore the value of personalizing policies, and find that in our context, there are no detectable gains.

How can coarsening signals—making them strictly *less* informative—lead to better decisions? Consider a specialist whose decisions deviate from those of a fully rational Bayesian, even when given the same information; we refer to these divergences as "disagreement regions." Intuitively, calibrated coarsening shrinks these regions by pooling them with nearby areas of agreement—a core insight from information design (Kamenica and Gentzkow, 2011). For example, if a Bayesian

[4]E.g., Pretrial risk is scored on a discrete 1–5 scale (see Figure 2 in Angelova et al. (2026)), and dental platform VideaAI color-codes disease risk into 3 bands instead of providing the raw probabilities (see Appendix Figure A.1).

would approve a loan given a specific AI signal but a biased specialist would not, pooling that signal with higher ones (where both would approve) can pull the specialist towards approval; the same logic applies to rejections. Note that a coarse binary message and a recommendation are isomorphic objects: both pool signal realizations into a single message, which the specialist either acts on or ignores. Our framework is agnostic between the two readings.[5]

Our framework generalizes this intuition and motivates our experiment. A Receiver (e.g., loan officer) takes a binary action to match a binary state (e.g., approve/deny the loan depending on whether it will be repaid), after observing (i) a (private) *human signal* (e.g., based on looking at the loan application) and (ii) a message from a Sender (e.g., the system designer). The Sender shares the Receiver's prior and payoffs, and privately observes an *AI signal*. Because signals are jointly normally distributed,[6] the Bayesian decision rule is characterized by a linear *decision boundary*, approving cases that lie above it and denying the rest. The Receiver is biased, so their decision boundary deviates from this Bayesian benchmark; rather than requiring the bias to be a fixed distortion of the Bayesian posterior as in de Clippel and Zhang (2022), we treat the boundary itself as primitive and remain agnostic about the source of the Receiver's bias (e.g., misspecified beliefs, updating errors, noisy decision making). The Sender's goal is to align the Receiver's behavior as closely as possible with the Bayesian benchmark by choosing to either fully reveal the AI signal, or instead coarsen it into a binary message, reporting only whether it exceeds a threshold t.

When the Receiver is Bayesian, full revelation is optimal; bias creates scope for coarsening to help. The logic is clearest when the Sender can condition the threshold on the human signal r. A binary Yes/No message splits Receivers into a band of *adhering types* who follow the recommendation, and extreme types outside the band who act on their own signal regardless. The optimal r-contingent binary threshold aligns each Receiver's action with the Bayesian benchmark as closely as adherence allows; we show that whenever the Receiver is biased, it strictly improves on full revelation—and, under mild conditions, is the optimal monotone disclosure mechanism.

In practice, the Sender does not observe r and must commit to a fixed threshold. Coarsening then operates as *endogenous delegation*: moderate types adhere, extreme types act on their own.

[5]The framework treats the specialist's response to messages as a primitive; in particular, the bias also need not stay fixed across information regimes: the form of disclosure (e.g., coarsening) can itself reshape it—a "Yes/No" message may invoke a different bias than a continuous score, due to, e.g., salience and framing effects (Bordalo et al., 2012; Tversky and Kahneman, 1981).

[6]Normality buys tractability (and shapes the conditions for *when* coarsening helps), while shared priors and aligned incentives isolate the role of bias and can be relaxed. The logic for *why* and *how* coarsening helps depends on neither, and carries over to less stylized settings.

Gains from coarsening come from correcting decisions of adhering types. However, the untailored threshold also introduces two types of losses: it misrecommends action for some adhering types, and leads non-adhering types to discard the message entirely. We characterize the optimal fixed threshold and show numerically that it is well approximated by a simple, behavior-free benchmark $\hat{t}$: the *AI-only threshold*, or the cutoff a Bayesian would use given only the AI signal.

Finally, we provide a sufficient condition, in model primitives, for coarsening at $\hat{t}$ to dominate full revelation. Coarsening is more likely to help the more the Receiver attenuates their response to the AI signal (i.e., flatter decision boundary), the wider the adhering types band, and the less residual information the human signal carries beyond the AI. Simulations show that coarsening at $\hat{t}$ delivers meaningful accuracy gains across wide parameter ranges, and under several common biases. Because $\hat{t}$ requires no knowledge of the Receiver's signal or bias, it is the threshold we use in our experiment.

Our main experiment tests the real-world applicability of our approach by evaluating whether coarsening signals at the AI-only threshold improves decision-making outcomes. The 150 loan specialists approve loan applications they believe will be repaid on time and deny the rest. All review simplified versions of real loan applications, allowing them to form beliefs about the probability of repayment (human signal). We use an off-the-shelf model to generate a risk-score for each application (AI signal), expressed as repayment probability.

Participants go through five order-randomized decision blocks that vary if and how the AI signal is shown: *Human-only* (no AI), *Probability* (the raw AI risk-score), and three *Binary* conditions which display only whether the AI signal is above or below a given threshold ("Yes, this applicant's AI score is above [threshold]"). Participants are incentivized with a bonus for every correct decision. Given this incentive structure, the AI-only threshold $\hat{t}$ is 50%, which we call *Binary Medium*. To evaluate the value of theoretically-informed coarsening over arbitrary coarsening, we also include two placebo binary conditions, *Binary Low* (30%) and *Binary High* (70%). We first estimate the Bayesian benchmark—the best a specialist could do given their information—and find that in our setting, specialists have little systematic residual information beyond the AI. This reflects a growing trend in many real-world contexts: as data-rich, multi-modal AI systems proliferate, humans increasingly have little systematic informational advantage over AI.[7]

[7]For example, in lending, firms like Upstart and ZestAI use models trained on hundreds of variables, often surpassing what any single underwriter sees—even when in-person interviews are conducted, AI can access video input. In healthcare, clinical decision-support tools embedded in EHRs integrate vitals, labs, clinical history, and imaging in real time—data that physicians may access, but not process as comprehensively or systematically.

We find that providing loan specialists with AI signals coarsened at the model-derived threshold improves decision accuracy compared to all other conditions. Participants are 4 percentage points (pp) more accurate in *Binary Medium* than under *Probability*—closing roughly 40% of the gap to the (in-sample) Bayesian benchmark—and 6pp more accurate than under *Human-only*.

Coarsening not only improves accuracy but also reduces decision time—a critical factor in high-throughput settings like loan underwriting. Participants are 7% faster in *Binary Medium* than in *Probability*. Moreover, many prefer coarsened signals: in an incentivized choice task, over half opt for either binary messages or no AI assistance over a full probability score, and those who choose binary signals were willing to forgo over a third of their per-case bonus to access them. These findings align with literature on aversion to richer information structures and cognitive processing costs (Guan et al., 2025, 2023; Exley and Kessler, 2023), and also echo interviews we conducted with senior lending professionals.

Performance gains are not uniform across specialists. Men benefit several times as much from AI assistance as women, reflecting a lower baseline (Human-only) performance—which mechanically leaves more room to gain—and possibly greater willingness to rely on AI advice. Moreover, consistent with theories of AI underweighting—the tendency to underweight algorithmic information—specialists exhibiting underweighting on an elicitation task perform better with coarsened than full AI risk scores, while those more Bayesian on this scale show the converse. However, other bias moderators also play a role, which suggests that suboptimal AI use likely stems from a combination of behavioral deviations rather than a single one. This highlights the value of calibrated coarsening as an intervention that can address multiple biases simultaneously, and also sets the stage for our next exercise: using past actions to estimate and assign the most effective coarsening policy.[8]

To explore this, we estimate a Bayesian hierarchical model on our main experiment data to predict the gains from personalizing coarsened policies. Our model suggests modest additional gains from personalization relative to the calibrated uniform policy—approximately 2pp—representing an upper bound that remains below the Bayesian benchmark. We then test the value of personalized coarsening in a two-stage experiment with new participants. The model uses decision data from the first stage to estimate individual behavior under different coarsening policies. In the second stage, participants are randomized to receive either their model-predicted best (personalized)

[8]For settings where a single dominant behavioral bias is plausibly stable—and access to historical data permit learning the joint signal distribution and a parsimonious bias parameter—more structural estimation can be informative for characterizing optimal coarsening and its expected gains.

coarsening policy—drawn from a richer set that includes trinary as well as binary coarsenings—or universal coarsened policy benchmarks. We find no average gains in practice—consistent with the modest predicted gains and behavioral nonstationarity across stages; suggestively, personalization performs better only where individual behavior remains stable between estimation and deployment. These results highlight the challenges of operationalizing personalization in dynamic environments where behavioral shifts are common—and suggest that in such settings, a universal calibrated coarsening rule may already offer a simple and effective alternative.

Related Literature

Calibrated coarsening is not the only way to address biases in human decision-making. "Debiasing" has traditionally taken three forms (Roy and Lerch, 1996): (i) modifying information presentation, (ii) replacing humans with models, and (iii) training. Coarsening falls within the first; the second is infeasible when humans must retain final authority—ruling out pure delegation (Raghu et al., 2019; Mozannar and Sontag, 2020; Athey et al., 2020; Bansal et al., 2021); and training can often be costly or context and bias-specific (Kahneman et al., 1982; Fong et al., 1986; Fong and Nisbett, 1991; Milkman et al., 2009; Morewedge et al., 2015; Sellier et al., 2019; Enke and Zimmermann, 2019). Coarsening, by contrast, is adaptable, low-cost, and bias-agnostic.

Our main contributions are three-fold. First, we provide empirical evidence that calibrated coarsening indeed improves decision-making outcomes in a high-stakes professional context: loan approvals. A large experimental literature studies human use of AI signals (Bundorf et al., 2024; Kiani et al., 2020; Maron et al., 2020; Bastani et al., 2026; Lai et al., 2021; Balakrishnan et al., 2022; Biermann et al., 2022; De-Arteaga et al., 2020; Lakkaraju and Farronato, 2022; Imai et al., 2023; Ben-Michael et al., 2025; Kim et al., 2024; Snyder et al., 2026; Otis et al., 2024; Grimon and Mills, 2025; Caplin et al., 2026), and a managerial prediction literature forecasts AI's positive impact on organizational performance (Brynjolfsson and McAfee, 2014; Davenport and Kirby, 2016; Daugherty and Wilson, 2018; Choudhary et al., 2025; Cockburn et al., 2018; Agrawal et al., 2019, 2022). Realizing these gains, however, requires addressing errors in belief-updating—e.g., algorithm aversion (Dietvorst et al., 2015, 2018; Logg and Schlund, 2024), over-adherence (Logg et al., 2019; Banker and Khetani, 2019; Bai et al., 2021; Buçinca et al., 2021; Fügener et al., 2021; Guo et al., 2024), rational inattention (Boyacı et al., 2024), prospect theory (Ye et al., 2022), and correlation neglect (Agarwal et al., 2023)—that shape how humans act upon AI predictions. A key

advantage of coarsening is its potential to address many biases simultaneously.[9] Other proposed solutions include explanations and uncertainty info (Bansal et al., 2019; Green and Chen, 2019; McGrath et al., 2020; Lakkaraju and Bastani, 2020; Taudien et al., 2022), model selection (Guo et al., 2025), hybrid approaches (Karlinsky-Shichor and Netzer, 2024; Chakraborty et al., 2025; Bell et al., 2024), and adaptive techniques (Sun et al., 2022; Caro and de Tejada Cuenca, 2023; Noti and Chen, 2023; Ibrahim et al., 2021; McLaughlin and Spiess, 2024; Dai and Singh, 2025).

We relate most closely to McLaughlin and Spiess (2022) (MS) and Agarwal et al. (2025) (AMW), who apply information design to human-AI collaboration. MS study a principal-agent model where agents incur penalties for deviating from AI recommendations, finding that withholding uncertain signals improves outcomes—effectively a form of coarsening where extreme signals are pooled into a recommendation while the center is fully disclosed without a recommendation. In contrast, our model accounts for errors in belief-updating and shows that coarsening helps even when the agent's payoff depends only on the cost structure, rather than on compliance with AI. AMW propose a framework for how and when to automate or assist decisions, including a form of coarsening that provides the average AI risk-score in each interval. Empirically, the two papers reach opposite results in different contexts: AMW find no benefit to coarsening among Prolific participants performing fact-checking tasks, while we find that coarsening improves approval accuracy among professional loan specialists, relative to full revelation. Theoretically, the frameworks make non-nested assumptions suited to different informational environments. AMW's framework imposes little structure, resting on the assumption that the Receiver's accuracy depends on AI only via the resulting posterior on the state—an assumption that, as they note, fails under Bayesian and quasi-Bayesian (e.g., Grether (1992)) models when human and AI information are non-independent conditional on the state—as is the case in our experiment and other recent work in the area (e.g., Agarwal et al. (2023)).[10] We characterize when and how coarsening improves outcomes through the geometry of the decision boundaries the signals induce, in a jointly normal environment that explicitly allows for conditionally correlated information.

Second, we provide one of the first empirical applications of information design to address human errors in probabilistic reasoning. Our theoretical framework draws on (non-)Bayesian per-

[9]Coarsening also connects tocategorical thinking, where agents simplify complex information by mapping observations into discrete categories and making forecasts using the category's representative belief (Mullainathan, 2002; Mullainathan et al., 2008; Barberis and Shleifer, 2003; Bordalo et al., 2020). We coarsen the *input signal* before the agent updates, which need not be equivalent to coarsening the *belief representation* itself.

[10]Examples of other settings in which human and AI information overlap include clinical diagnosis (Jabbour et al., 2023), child-protection services (Grimon and Mills, 2025), and judicial decision-making (Imai et al., 2023).

suasion (Kamenica and Gentzkow, 2011; Alonso and Câmara, 2016; Kolotilin, 2018; Kamenica, 2019; Levy et al., 2022; de Clippel and Zhang, 2022); a central result in this literature is that when the Sender's and Receiver's incentives are misaligned (Kamenica and Gentzkow, 2011), or when the Receiver deviates from Bayesian updating under a certain class of belief distortions (de Clippel and Zhang, 2022), informational structures other than full revelation—of which coarse partitions is a subset—can be optimal. In recent work, Braghieri (2026) theoretically characterizes when information design recovers the first best under biased updating. Other work treats coarsening as a *constraint* limiting persuasion (Aybas and Turkel, 2019; Chen et al., 2025) or studies other channels through which granularity improves outcomes (Li et al., 2024). We instead use calibrated coarsening as an empirical *tool* to counteract biased belief updating. While the persuasion literature has been largely theoretical, with only a few studies applying information design in empirical settings (Decker, 2022; Xiang, 2026; Avoyan et al., 2026), or focusing on how coarse disclosure can improve outcomes through other channels such as incentives, strategic interaction, or design constraints (Sellier and Dahl, 2011; Goel and Thakor, 2015; Boleslavsky and Cotton, 2015; Harbaugh and Rasmusen, 2018; Corradini, 2024; Zhang and Misra, 2024), we provide, to our knowledge, the first experimental evidence that *calibrated* coarsening improves outcomes for decision-makers whose choices deviate from the Bayesian benchmark.

Third, we offer calibrated coarsening as a practical approach firms and other stakeholders can easily implement. The solution is widely applicable and portable across contexts and decision-making biases. Beyond AI predictions, our framework also applies more broadly to decision-making involving multiple signals;[11] we focus on algorithmic signals because they enable precise, transparent adjustments through threshold choice.[12]

The paper proceeds as follows. Section 2 presents a model of decision-making and illustrates how coarsening can address biases. Section 3 introduces the loan context. Section 4 describes the main experiment, which tests a universal coarsening policy with loan specialists, and Section 5 presents the results. Section 6 estimates gains from personalization. Section 7 concludes.

[11] This also connects to literatures on second opinions (Yaniv, 1997; Yaniv and Kleinberger, 2000; Yaniv, 2004; Surowiecki, 2005; Soll and Larrick, 2009; Sah and Loewenstein, 2015; Staradub et al., 2002; Taylor and Potts, 2008), knowledge transformation (Carlile, 2004; Hardy et al., 2005; Levina, 2005), AI in organizations (Faraj et al., 2018; Von Krogh, 2018; Rai et al., 2019; Kellogg et al., 2020; Anthony et al., 2023), and user engagement with AI (Christin, 2020; Pachidi et al., 2021; Lebovitz et al., 2022; Kelley et al., 2022).

[12] Whereas adjusting the sensitivity or specificity of human second opinions (e.g., a radiologist colleague's input) is difficult, algorithmic models make such adjustments feasible.

2 Model

We develop a framework of human-AI decision-making to build intuition for when, why and how coarsening an AI signal can improve the decisions of a biased decision-maker. Two assumptions run throughout: that signals are jointly normal, and that the Sender and Receiver have aligned incentives. The first is a matter of tractability, underlying the conditions for *when* coarsening helps; the second is a matter of focus, isolating the role of deviations from Bayesian updating, and can be relaxed—indeed, coarsening can also serve a misaligned Sender.[13] The logic for *why* and *how* coarsening helps depends on neither, and carries over to less stylized settings.

2.1 Setup

A *Receiver* (e.g., a loan specialist) chooses between two actions $a \in \{0, 1\}$ (e.g., denying or approving a loan), given a binary state $\omega \in \{0, 1\}$ (e.g., whether the applicant repays on time). The Receiver's payoff from $a = 1$ is:

$$U(\omega) = -c_{FP}(1 - \omega) + c_{TP} \cdot \omega, \tag{1}$$

where $c_{FP} > 0$ is the cost of a false positive and $c_{TP} > 0$ the benefit of a true positive. The payoff from $a = 0$ is normalized to zero.[14] The Receiver holds prior $p = \Pr(\omega = 1)$ and privately observes a *human* signal r (their own assessment of the application) which we also call their *type*. A *Sender* (the system designer, or a manager controlling AI output) shares the Receiver's prior and payoffs but privately observes a distinct *AI* signal s (the applicant's AI risk score). Conditional on ω, the signals are jointly normal, with means $\mu_r^\omega = \mu_s^\omega = \mu^\omega = \omega$, state-independent variances $\sigma_r^2 \geq \sigma_s^2 > 0$ (human signal weakly less precise than the AI), and correlation $\rho \geq 0$ (signals weakly positively correlated conditional on ω). We further assume $\sigma_s \geq \rho\sigma_r$, ensuring that the Bayesian posterior $\pi(r, s) = \Pr(\omega = 1 \mid r, s)$ is weakly increasing in both arguments.

Bayesian benchmark. A Bayesian Receiver observing (r, s) chooses $a = 1$ iff their posterior $\pi(r, s)$ exceeds $\frac{c_{FP}}{c_{FP}+c_{TP}}$. Since s and r are jointly normal, the log posterior ratio is linear:

[13] The framework can accommodate misaligned utilities and heterogeneous priors as in Kamenica and Gentzkow (2011); Alonso and Câmara (2016).

[14] When utility only depends on the action-state pair (a, ω), this normalization is without loss.

$$\log \frac{\pi(r,s)}{1-\pi(r,s)} = \log \frac{p}{1-p} + \frac{1}{1-\rho^2}\left[\frac{r-\frac{1}{2}}{\sigma_r^2} + \frac{s-\frac{1}{2}}{\sigma_s^2} - \frac{\rho(r+s-1)}{\sigma_r\sigma_s}\right].$$

Denote the standard deviation of r given s by $\sigma_{r|s} = \sigma_r\sqrt{1-\rho^2}$ and let $\chi = \frac{\sigma_s - \rho\sigma_r}{\sigma_{r|s}}$. The *residual information coefficient* χ measures how much information r carries beyond s: it is the weight the Bayesian rule places on the human signal orthogonal to the AI's.[15]

When $\chi > 0$, a Bayesian chooses $a = 1$ whenever $r > r^*(s)$, where

$$r^*(s) = \gamma_0^* + \gamma_1^* s \tag{2}$$

is a (linear) *decision boundary* (DB) separating the s-r plane into $a = 1$ and $a = 0$ regions. It slopes downward: a higher AI signal lowers the human signal needed to act. Its steepness reflects the two signals' residual informativeness, rising with the AI's residual (what s adds beyond r) and falling with the human's.[16] When $\chi = 0$, the human adds no information beyond the AI (r is a garbling of s), and the boundary is vertical and the decision depends on s alone. The intercept γ_0^*, by contrast, shifts with the prior and payoffs as well.[17]

Biased Receivers. The Receiver's DB may deviate from the Bayesian one, and is given by:

$$\tilde{r}(s) = \tilde{\gamma}_0 + \tilde{\gamma}_1 s. \tag{3}$$

The deviation of the Receiver's DB from the Bayesian DB can stem from a biased posterior mapping $\tilde{\pi}(r,s) \neq \pi(r,s)$, due to, e.g., belief misspecification, under- or over-reaction (Grether, 1992), correlation neglect (Enke and Zimmermann, 2019), confirmation bias, salience and framing effects (Bordalo et al., 2012), and other errors in probabilistic reasoning (Benjamin, 2018). It may also be driven by distortions in the belief-action mapping (Enke and Graeber, 2023), optimization errors (Gabaix, 2014), or a preference for or against relying on algorithmic output.[18] We remain agnos-

[15] To see this, let $u = (r - \frac{1}{2}) - \frac{\rho\sigma_r}{\sigma_s}(s - \frac{1}{2})$ be the component of r orthogonal to s. The log posterior odds can be written as $\log\frac{p}{1-p} + \frac{1}{\sigma_s}\left(\frac{s-\frac{1}{2}}{\sigma_s} + \chi\frac{u}{\sigma_{r|s}}\right)$, so χ is the coefficient on the standardized human residual. Our assumption $\sigma_s \geq \rho\sigma_r$ is equivalent to $\chi \geq 0$, with equality iff r is a garbling of s.

[16] In particular, $\gamma_1^* = -\chi_s/\chi$, where $\chi_s = \frac{\sigma_r - \rho\sigma_s}{\sigma_s\sqrt{1-\rho^2}} > 0$ is the AI's residual weight, defined symmetrically to χ (swapping r and s); it is positive since $\sigma_r \geq \sigma_s$. A larger AI residual thus requires a bigger offsetting move in r to keep the Bayesian indifferent.

[17] The algebraic expression for the intercept is $\gamma_0^* = \frac{\sigma_r^2\sigma_s(1-\rho^2)}{\sigma_s - \rho\sigma_r}\left(\log\frac{c_{FP}}{c_{TP}} - \log\frac{p}{1-p}\right) + \frac{1}{2}(1-\gamma_1^*)$.

[18] In the last case, utility would need to depend directly on the signals, in addition to (1).

tic about the source of the bias, and treat $\tilde{\gamma}_0$ and $\tilde{\gamma}_1$ as primitives. We maintain throughout that $\tilde{\gamma}_1 \leq 0$: the Receiver may under- or over-respond to the AI signal, but does not flip the sign of her response—treating higher AI scores as evidence *against* $\omega = 1$ (recall $\gamma_1^* < 0$).[19]

Illustration: AI misweighting. Our leading illustration is *AI misweighting*: the Receiver's posterior map $\tilde{\pi}$ updates from the AI signal *as if* it has variance σ_s^2/α_s rather than σ_s^2, while correctly perceiving all other parameters. When $\alpha_s < 1$, the Receiver perceives the AI as less precise than it is, capturing *AI underweighting* or algorithmic aversion (Dietvorst et al., 2015); when $\alpha_s > 1$, the Receiver perceives it as more precise, capturing *AI overweighting* or algorithmic appreciation (Logg et al., 2019). Underweighting flattens $\tilde{r}(s)$ relative to $r^*(s)$, while overweighting steepens it; motivated by our later empirical findings in Section C.2.2, our main-text illustration focuses on the former. Figure I visualizes the main concepts in the s-r plane. In all panels, the DGP is calibrated to our experimental data via Maximum Likelihood: $\hat{\sigma}_r = 2.27$, $\hat{\sigma}_s = 0.79$, $\hat{\rho} = 0.25$, and $p = 2/3$ by design.[20] The shading represents the joint density. To simplify the interpretation of units, and as is the case in our experiment, we set $c_{FP} = c_{TP} = 1$ throughout our illustration and later simulations, so expected utility reduces to decision accuracy. Figure Ia visualizes the Bayesian DB (solid black) and the AI underweighting DB (setting $\alpha_s = 0.5$, i.e., discounting the precision of AI signal by half; solid teal): the Receiver would choose $a = 1$ above the boundary and $a = 0$ below. The region where the DBs disagree (dotted blue) is the scope for improvement.

Other biases. In Appendix B.2 we reproduce the panels of Figure I for other frequently studied biases: AI *over*weighting, overconfidence (treating own information, r, as more precise than it is), correlation neglect (ignoring the conditional correlation between r and s) and base-rate neglect (underweighting the prior when updating from signals). These biases all generate analogous shifts and/or rotations and lead to qualitatively similar analysis. For our purposes, what matters is the disagreement region between the Bayesian and biased DBs (i.e., the set of signal pair realizations for which the biased Receiver acts differently from a Bayesian), regardless of its exact psychological drivers—and, as we show next, whether disclosure can be structured to reduce it.

[19] Unlike de Clippel and Zhang (2022), we do not require the biased posterior to be a fixed distortion of the Bayesian posterior: a distorted DB can stem from a biased posterior map $\tilde{\pi}$ where two signal pairs with the same Bayesian posterior yield different biased posteriors.

[20] See Appendix B.1. In Section 5.2.1 we estimate this boundary directly, by a logistic regression of ω on the signals' log-odds; it is nearly identical but, unlike the joint-normal estimate, does not impose equal covariance across states.

Signal disclosure. Since incentives are aligned, the Sender's objective is to maximize the Receiver's (Bayesian) expected utility. The Sender publicly commits to a disclosure rule. The Receiver then observes their own signal r together with the Sender's message and chooses an action. The Sender can either reveal s in full—*full revelation*, under which the biased Receiver acts according to (3)—or coarsen it. A *binary coarsening at threshold* t maps s to a binary signal $k_t(s) = \mathbf{1}\{s > t\}$, revealing only whether s exceeds t.[21] Let $\tilde{r}_1(t)$ denote the lowest r type that acts upon observing $k = 1$ ("Yes" message) and $\tilde{r}_0(t)$ the highest type that does not act upon observing $k = 0$ ("No" message); we call $\tilde{r}_1(\cdot)$ and $\tilde{r}_0(\cdot)$ the two *adherence curves*.

While $\tilde{r}_k(t)$ and $\tilde{r}(s)$ can arise from a single underlying model in principle,[22] a Receiver's bias may itself depend on the form of disclosure: mechanisms such as salience and framing (Bordalo et al., 2012; Tversky and Kahneman, 1981) make beliefs sensitive to how information is presented, so coarsening may alter the bias directly beyond pooling signals together.

We accordingly use "message" and "recommendation" interchangeably, as the framework is agnostic as to whether the Receiver processes $k_t(s)$ as information about s or as advice about the action: both readings are subsumed by treating the adherence curves $\tilde{r}_k(t)$ as primitive. We impose a simple sandwich condition: for all t,

$$\tilde{r}_1(t) < \tilde{r}(t) < \tilde{r}_0(t),$$

where $\tilde{r}(t)$ is the cutoff type that just acts under full revelation (seeing the exact signal $s = t$). The condition orders the messages by how strongly they induce action: a "Yes" ($s > t$) is more favorable to action than the raw signal $s = t$ (full revelation) since it pools higher signals together, while the converse holds for "No" ($s \leq t$). We can therefore split the Receiver population (Figure Ib y-axis) into three groups: *adhering types* in the *adherence band* $[\tilde{r}_1(t), \tilde{r}_0(t)]$ that act if and only if they see a "Yes" message, types below $\tilde{r}_1(t)$ that never act, and those above $\tilde{r}_0(t)$ that always act.

How strongly each message moves the Receiver is captured by the *adherence bandwidths*

$$\delta_1(t) = \tilde{r}(t) - \tilde{r}_1(t) > 0, \qquad \delta_0(t) = \tilde{r}_0(t) - \tilde{r}(t) > 0,$$

[21] In Section 2.2.1, the threshold t can be conditioned on the Receiver's signal r—the "known r" case. In Section 2.2.2, t is chosen without observing r—the "unknown r" case.

[22] For example, a biased posterior mapping $\tilde{\pi}$ governing belief updating from both the fully revealed signal s and its binary coarsening $k_t(s)$. In particular, the posterior takes the form of a weighted average: $\Pr(\omega = 1 \mid r, k) = \int \tilde{\pi}(r, s)\tilde{\beta}_t(s \mid r, k)ds$, where $\tilde{\pi}(r, s)$ is the biased posterior under full revelation of s, and $\tilde{\beta}_t(s \mid r, k)$ is a bias-consistent weighting over signals satisfying $s > t$ conditional on r.

the distances by which the "Yes" ("No") message lowers (raises) the action boundary relative to full revelation at $s = t$; the sandwich condition requires both to be positive. The bandwidths capture the strength of the adherence: many types act if and only if they are recommended to.

Figure Ib illustrates behavior under binary coarsening, with adherence curves $\tilde{r}_k(t)$ generated from an AI underweighting posterior mapping ($\alpha_s = 0.5$) arising from the same underlying model as $\tilde{r}(s)$. For a given threshold t, the DB is a step function whose two levels are read off the adherence curves at $s = t$. Raising the threshold from t_1 to t_2 has two effects: it changes *who* follows the signal's implied recommendation (vertical shifts of the DB from $\tilde{r}_k(t_1)$ to $\tilde{r}_k(t_2)$), and the *content* of each message (threshold slides horizontally along t_1 to t_2).

Why coarsen? When the Receiver is Bayesian, full revelation is optimal (Kamenica and Gentzkow (2011); Proposition 6(i) in Kolotilin (2018)). When the Receiver is biased, full revelation need not be first-best. In the next subsection, we explore how binary coarsening can improve outcomes.

2.2 Binary coarsening

2.2.1 Binary coarsening with known r

To build intuition, suppose first that the Sender can condition the threshold t on the Receiver's signal r. If the Receiver always adhered, the Sender would simply set $t(r)$ to $s^*(r) = (r - \gamma_0^*)/\gamma_1^*$, the first-best threshold that recommends action if and only if a type-r Bayesian would act. But the Receiver only follows the binary message when they fall within the adherence band, $r \in [\tilde{r}_1(t), \tilde{r}_0(t)]$. The Sender therefore sets the personalized threshold $t^*(r)$ closest to $s^*(r)$ under which the Receiver would actually adhere:

$$t^*(r) = \underset{t:\, r \in [\tilde{r}_1(t), \tilde{r}_0(t)]}{\operatorname{argmin}} |t - s^*(r)|,$$

i.e., the cutoff closest to the Bayesian subject to adherence. When r falls in the adherence band at $s^*(r)$, the constraint does not bind: $t^*(r) = s^*(r)$, delivering first-best. Otherwise, $t^*(r)$ lands on the nearest adherence curve, minimizing the deviation from the Bayesian DB.

Figure Ic plots the optimal r-conditioned threshold $t^*(r)$ (solid purple; note the inverted axis), which coincides with its induced decision boundary—each type r acts iff $s > t^*(r)$—together with the associated gains in our AI underweighting illustrative example. For r values around the center, the Receiver adheres at $s^*(r)$, so the threshold coincides with the Bayesian DB. In the tails

where the Receiver would not adhere, the Sender must shift the threshold outward onto the nearest adherence curve (away from black Bayesian DB). Quantitatively, binary coarsening almost fully corrects the bias in this example, capturing 96% of the potential gain.

Proposition 2 in Appendix B.3 formalizes this: under mild regularity conditions on the Receiver's response to messages (Assumption 1 in Appendix), whenever the Receiver is biased, binary coarsening strictly improves on full revelation (part i), is optimal among deterministic monotone disclosure rules (part ii), and fully implements the first-best when the bias and the Receiver's default action point in opposite directions (part iii).

2.2.2 Binary coarsening with unknown r

Conditioning the threshold on r has two advantages: (i) it tailors the recommendation to each type, achieving the optimal (adherence-constrained) decision, and (ii) it guarantees adherence by construction, since $t^*(r)$ is chosen so that r falls in the adherence band.

Both advantages hinge on observing r and conditioning t on it. In practice the Sender rarely observes r: eliciting it may be costly and susceptible to misreporting, and varying thresholds may be impractical to implement.[23] The Sender must then commit to a single fixed threshold t. Gains from coarsening nonetheless can remain.

Returning to the illustration, Figure Id plots the gains (blue) and losses (orange) from coarsening at the fixed optimal threshold t^*, relative to biased full revelation. Coarsening functions as endogenous delegation: Receivers with relatively moderate r adhere to the implied AI recommendation, while those with extreme r act on their own signal. Here, the gains (blue) come from the adhering types: by pooling a moderate AI score with stronger ones, the "Yes" message flips them from their biased action toward the Bayesian one. This is the delegation channel at work—the recommendation, though untailored, points moderate types the right way.

The losses (orange) reflect the two advantages forgone by not conditioning on r. The first is *misrecommendation*: with one threshold for everyone, the untailored cutoff overshoots for some adhering types, flipping actions they would have gotten right under full revelation. The second is *non-adherence*: extreme types ignore the message, forgoing the correction a strong AI score would have induced under full revelation. In the illustration the gains dominate (net $+0.82$pp accuracy).

[23] Related work on Sender–Receiver problems with Receiver private information includes Guo and Shmaya (2019); Alonso and Câmara (2016); on Sender elicitation, see Kolotilin et al. (2017); Li and Shi (2017).

The optimal fixed threshold t^*. Define the (Bayesian) welfare contribution of acting at (r, s):

$$Q(r, s) = ph_1(r, s)c_{TP} - (1 - p)h_0(r, s)c_{FP}. \tag{4}$$

$Q(r, s) \geq 0$ where a Bayesian would act, so the Bayesian DB is the contour $Q(r, s) = 0$.[24]

With t fixed, the Sender's objective is the total Q over everyone who acts—the green regions of Figure Ib: types above $\tilde{r}_0(t)$ after a "No" ($s < t$), and above $\tilde{r}_1(t)$ after a "Yes" ($s > t$):

$$V(t) = \int_{S_0}\int_{\tilde{r}_0(t)}^{\infty} Q(r, s)\, dr\, ds + \int_{S_1}\int_{\tilde{r}_1(t)}^{\infty} Q(r, s)\, dr\, ds, \qquad S_0 = (-\infty, t),\ S_1 = (t, \infty).$$

Raising t moves the purple boundary in Figure Ib, with any interior optimum satisfying:

$$-V'(t) = \underbrace{\int_{\tilde{r}_1(t)}^{\tilde{r}_0(t)} Q(r, t)\, dr}_{\text{recommendation (RC)}} + \underbrace{\sum_{k\in\{0,1\}} \tilde{r}_k'(t)\int_{S_k} Q(\tilde{r}_k(t), s)\, ds}_{\text{adherence selection (AD)}} = RC(t) + AD(t) = 0.$$

The *recommendation* term RC captures the horizontal change of the threshold from t_1 to t_2, i.e., how it changes the recommendation for those that fall within the adherence band. The *adherence-selection* term AD captures the resulting vertical shift, i.e., *who* adheres to the recommendation.

A bias-agnostic approximation $\hat{t}$. Computing t^* requires the adherence curves $\tilde{r}_k(\cdot)$, and hence data on how the Receiver behaves under a continuum of thresholds. We propose instead a threshold that requires no knowledge of the Receiver's signal or behavior: the *AI-only threshold* $\hat{t}$, defined by $\int Q(r, \hat{t})\, dr = 0$: the cutoff a Bayesian would use on seeing the AI signal alone.[25]

As we verify numerically below, $\hat{t}$ approximates t^* remarkably well across the parameter space. The logic is simple: the welfare loss from using $\hat{t}$ is second-order in $|V'(\hat{t})| = |RC(\hat{t}) + AD(\hat{t})|$.[26] When both terms are small, $\hat{t}$ nearly solves the Sender's first-order condition.

The $RC(\hat{t})$ term is small when the adhering types are representative of the population. By construction, the integral of $Q(r, \hat{t})$ over *all* r is zero; $RC(\hat{t})$ restricts this to the adherence band, dropping the two tails of the r-distribution, so is small whenever the two tails are individually small

[24] Formally, $Q(r, s) = h(r, s) \cdot \mathbb{E}(U(\omega) \mid r, s)$ is the Bayesian expected utility from acting at (r, s), scaled by the signal density h.

[25] In our joint normal setting, $\hat{t} = \frac{1}{2} + \sigma_s^2\big(\log\frac{1-p}{p} + \log\frac{c_{FP}}{c_{TP}}\big)$.

[26] If $V'' \leq -\kappa < 0$ on the interval between $\hat{t}$ and t^*, then $V(t^*) - V(\hat{t}) \leq |V'(\hat{t})|^2/(2\kappa)$.

or offset one another.[27]

The $AD(\hat{t})$ term captures how welfare changes as the threshold shifts *who* adheres. The term multiplies two factors: how much the marginal adhering type shifts with the threshold ($\tilde{r}_k'(t)$), and how much welfare those marginal types carry given the message ($\int_{S_k} Q(\tilde{r}_k(t), s)\, ds$). It is small when either factor is.[28] The figures in Appendix B.4 quantify and illustrate why both channels are small in our AI misweighting illustration.

When binary coarsening at $\hat{t}$ dominates full revelation. We now give a sufficient condition, in terms of primitives, for coarsening at $\hat{t}$ to dominate full revelation. Write $g = |\tilde{\gamma}_1|$ for the Receiver's slope and $\delta = \big(\delta_1(\hat{t}), \delta_0(\hat{t})\big)$ for the bandwidths at $\hat{t}$.

Proposition 1. *There exist a threshold $\bar{\chi}(g, \delta) \in \mathbb{R}$ and a* slope tolerance *$\Gamma(\delta) \geq 0$, independent of χ, such that:*

(i) If $\chi < \bar{\chi}(g, \delta)$, binary coarsening at $\hat{t}$ strictly dominates full revelation.

(ii) $\bar{\chi}$ is continuous, strictly increasing in δ_1 and in δ_0, and strictly decreasing in g.

(iii) $\bar{\chi}(g, \delta) > 0$ if and only if $g < \Gamma(\delta)$. The slope tolerance Γ is continuous and strictly increasing in each band, with $\Gamma(0, 0) = 0$, $\Gamma(\delta) < \infty$ for every δ, and $\Gamma(d, d) \uparrow \infty$ as $d \to \infty$. In particular, $\bar{\chi} > 0$ for any $\delta \neq (0, 0)$ if $g = 0$, and, for any g, once both bands are wide enough.

The proof in Appendix B is constructive, delivering $\bar{\chi}(\cdot, \cdot)$ and $\Gamma(\cdot)$ in closed form. The proposition organizes the gains and losses of Figure Id through its three primitives: the adherence bandwidths δ (gaps between adherence curves and biased DB), the slope g (steepness of biased DB), and the residual weight χ (which governs slope of Bayesian DB).[29] Loosely, the three primitives pull in intuitive directions: coarsening at $\hat{t}$ is more likely to beat full revelation the wider the adherence, the flatter the Receiver's DB, and the less residual information the human carries.

[27] The omitted low-r tail carries negative welfare (it typically lies below the Bayesian DB, where $Q < 0$) and the high-r tail positive (typically $Q > 0$). The two cancel—and $RC(\hat{t})$ is small—whenever the adherence band is representative of the population, so that the average welfare of acting at $s = \hat{t}$ is close to its population average of zero. $RC(\hat{t})$ vanishes *exactly* when $\chi = 0$, where r carries no information beyond s so $Q(r, \hat{t}) = 0$ at every r; it vanishes *approximately* for small χ, where $Q(\cdot, \hat{t})$ is nearly flat and close to zero across the band.

[28] Under-reaction to s (e.g., underweighting) flattens the adherence curves, so $|\tilde{r}_k'(\hat{t})|$ is small. Over-reaction (overweighting) pushes $\tilde{r}_1$ down and $\tilde{r}_0$ up into the tails of the joint density, where little mass lies, so the welfare integral is small. For a Bayesian Receiver the second factor vanishes: the marginal type is exactly indifferent given recommendation k, so $\int_{S_k} Q(\tilde{r}_k, s)\, ds = 0$—an envelope property—and $AD(t) = 0$ at *every* threshold.

[29] While Figure Id plots the gains and losses from coarsening at t^* rather than $\hat{t}$, the same geometry applies.

These directions are sharpest at $\chi = 0$, where the human adds no residual information over the AI: the Bayesian DB (black line) is vertical and coincides with $\hat{t}$, so coarsening at $\hat{t}$ recommends action exactly when a Bayesian would for every adhering type. Coarsening therefore dominates full revelation when the gains from the adhering types outweigh non-adherence losses—the condition $g < \Gamma(\delta)$ of part (iii). The comparative statics follow: $\bar{\chi}$ rises with adherence bandwidths δ and falls with the slope g. Widening adherence turns some non-adhering types into adhering ones, and at $\chi = 0$ each then sits on the exactly-right boundary. Attenuating the Receiver's DB (teal) slope helps coarsening twofold: it widens the disagreement (gap between the biased and Bayesian actions) coarsening corrects among adhering types, and shrinks the non-adherence losses.[30]

When $\chi > 0$, the Bayesian DB flattens, and losses appear among adhering types: the threshold now overshoots for some, pushing them *away* from the Bayesian benchmark. This erodes some of the gain, but only so far; in particular, the welfare improvement survives for all χ up to $\bar{\chi}$.[31] The bound is conservative: coarsening can help even for some $\chi > \bar{\chi}$.[32]

The sufficient condition $\chi < \bar{\chi}(g, \delta)$ for coarsening to beat full revelation is not necessary: it requires a sufficiently attenuated Receiver DB ($g < \Gamma(\delta)$) and is silent on overly steep DBs (e.g., AI overweighting). Coarsening can still help there,[33] as Appendix B.2 and simulations below show.

Gains from coarsening: simulations across parameters. Proposition 1 treats residual information, slope, and adherence as independent primitives, for tractability and transparency. Yet they may co-move in practice. First, slope and adherence both vary with residual information: as a biased Receiver's r becomes more or less informative, she may adjust her responsiveness to the AI signal.[34] Second, even fixing χ, slope and adherence tend to move together: a Receiver who responds less to the AI signal under full revelation (flatter slope) typically also responds less under coarsened disclosure (narrower adherence bands). Since $\bar{\chi}$ rises with the bands but falls with the

[30] Less is lost when non-adhering types always act on their own signal, since fewer of them would have gotten it right under biased full revelation anyway.

[31] Proposition 1 relates to Theorem 2 in Braghieri (2026), where coarsening attains the first best exactly when it is *safe*—Receiver's response induces the objectively optimal action (analogous to our wide adherence bandwidths δ)—and *lossless*—pooling destroys no first-best decision value (analogous to small residual info χ—at $\chi = 0$, coarsening at $\hat{t}$ pools only within decision faces). Here, residual information is what makes coarsening lossy: as finer signal partitions become decision-relevant, the value of coarsening is gradually eroded. There, with the Receiver holding no private information, greater curvature of the decision problem plays a similar role.

[32] It is, however, generically tight at zero: as shown in the proof, $\bar{\chi} > 0$ precisely when coarsening dominates for a nondegenerate set of χ near zero (equivalently, $g < \Gamma(\delta)$). So while $\bar{\chi}$ understates how large the true dominance threshold may be, it correctly identifies whether any such threshold above zero exists at all.

[33] Under AI overweighting, coarsening corrects overreaction to extreme signals by withholding the exact value of s.

[34] e.g., Under AI misweighting, full-revelation DB $\tilde{r}(s)$ and adherence curves $\tilde{r}_k(t)$ all move with χ for any $\alpha_s > 0$.

slope, such co-movement pushes it in opposite directions, so the net effect remains ambiguous.

We address this question numerically, simulating welfare under AI misweighting across parameter values. Figure II plots welfare differences relative to biased full revelation (V^{full}) for three benchmarks: binary coarsening at the optimal threshold ($V(t^*)$; dashed teal with circles), the approximation ($V(\hat{t})$; dotted teal), and the Bayesian first-best under full information (V^{FB}; solid black). The black line is the maximum potential gain—the analog, at each parameter value, of the potential gain in Figure Ia—while each teal line depicts the gain coarsening delivers. How close the teal lines sit to the black therefore shows how much of that gap coarsening closes.

First, across all panels, the AI-only threshold $\hat{t}$ achieves virtually the same welfare as the optimal t^*; we treat them together below. The top row varies one observable covariance parameter $(\rho, \sigma_r, \sigma_s)$ at a time, holding the others at the baseline estimated from our experimental data [35] and fixing the AI-misweighting parameter α_s at 0.5 under both full revelation and binary coarsening.[36] Consistent with Proposition 1, coarsening helps more when the human signal adds little residual information (χ small): as correlation rises, r grows more noisy, and s more precise.

The bottom row instead holds the DGP fixed and varies the Receiver's behavior.[37] Its left and middle panels allow the Receiver to (mis)weight the continuous and binary AI signals differently, isolating two comparative statics of Proposition 1. The left varies misweighting under full revelation α_s^{full} (which sets the DB slope), the middle panel coarsened-signal misweighting α_s^{coarse} (which sets the adherence bandwidths), each holding the other fixed at 0.5. Consistent with the proposition, a flatter DB slope and wider adherence favors coarsening.[38] The right panel varies a single α_s governing behavior under both regimes, so δ and g co-move (at fixed χ): lowering α_s flattens the slope (favoring coarsening) while narrowing the adherence bandwidth (against it). The simulation signs the net effect: coarsening helps under underweighting, where the slope channel dominates, and even under overweighting in relative terms.[39]

Deviations from Bayesian decision-making may arise from many combinations of misspec-

[35] See figure note. We cap the range of the DGP parameters to preserve our monotonic posterior assumption $\chi \geq 0$.

[36] This is consistent with the AI underweighting we empirically document in Figure VIb.

[37] We cap the α_s ranges in these panels to preserve our monotonicity assumptions on the Receiver's behavior ($\tilde{\gamma}_1 < 0$, and the sandwich condition on $\tilde{r}_k$). These fail under severe underweighting, where the perceived weight on s flips sign, and severe overweighting, where the perceived weight on r does.

[38] Here, coarsening helps when $\alpha_s^{\text{full}} < 0.6$, with gains rising steeply as the slope flattens; as adherence widens, gains rise and turn positive at $\alpha_s^{\text{coarse}} \approx 0.4$, and almost fully close the gap to first-best at $\alpha_s^{\text{coarse}} \approx 0.7$.

[39] Under AI underweighting ($\alpha_s < 1$), coarsening raises accuracy by up to 1.5pp—a quarter of the gap to the first-best. At $\alpha_s = 1$, the Receiver is Bayesian and full revelation is optimal. Under AI overweighting ($\alpha_s > 1$), absolute gains shrink—small χ keeps even an over-steep DB close to the near-vertical Bayesian one—but in relative terms coarsening closes nearly the entire gap.

ification and bias. Our simulations show that coarsening can generate meaningful gains under empirically relevant ones: beyond AI misweighting, Appendix Figure B.4 replicates the bottom-row exercise of Figure II for three other common biases: overconfidence, correlation neglect, and base-rate neglect. Whether coarsening helps depends on the bias—under our DGP, it beats full revelation under some (e.g., base-rate neglect) but not others (e.g., overconfidence). Whether coarsening beats full revelation in practice is ultimately an empirical question. In the next section, we test coarsening at our proposed threshold $\hat{t}$ in an experiment with professional loan officers, comparing it against several benchmarks: full revelation (access to full AI score), human-only (no AI input), and two alternative coarsenings—Binary High ($t \gg \hat{t}$) and Binary Low ($t \ll \hat{t}$).[40]

3 Industry Context: Consumer Loan Approvals

Our experiments test the deployment of coarsened signals in consumer loan approvals, an important use case of AI and an industry with a long history of leveraging algorithmic predictions. Loan decisions are economically and socially consequential: timely repayment not only generates revenue through interest and fees but also promotes financial inclusion, while defaults lead to losses for lenders. This is the real-world setting in which we conduct our main experiment.

3.1 Background on algorithmic inputs in consumer loans

Consumer loan underwriting has relied heavily on algorithmic models for decades, most prominently the credit-scoring system FICO introduced in 1989 (FICO, 2024). Algorithmic approaches were met with resistance, with one senior mortgage banking executive we interviewed describing the initial introduction of the FICO system as "challenging," noting that while it has eventually become "an important tool," it remains "an input that can be overridden" by human underwriters.

While FICO scores remain widely used, the landscape has since shifted significantly toward integrating automated and, increasingly, AI-driven algorithms. For instance, the Government-Sponsored Enterprises (GSEs) Fannie Mae and Freddie Mac employ automated underwriting systems like Desktop Underwriter (DU) and Loan Product Advisor (LPA). These systems summarize recommendations regarding loan eligibility into categories like "Approve/Eligible," "Refer," or "Ineligible" (Freddie Mac, 2025)—effectively a trinary coarsening—for lenders to use in decision-

[40] There, we also recover the empirical analogs of Figure Ia—the Bayesian and specialist decision boundaries, and the disagreement region between them—directly from specialists' choices.

making. LPA has been integrated with third-party AI-powered platforms to enhance underwriting assessments (BusinessWire, 2021; PRNewswire, 2024).

3.2 Automated underwriting vs. humans-in-the-loop

Although many consumer-credit products—like credit card applications and auto loans—are evaluated using automated algorithms with minimal human oversight, higher-stakes loans like mortgages continue to rely on human decision-makers as final arbiters who use algorithms as advisory tools. For example, DU and LPA recommendations are being used as input in lending decisions, but do not replace the lender's obligation to review borrower documentation and ultimately approve the loan. Importantly, liability still lies with the lenders if they sell non-compliant loans to Fannie or Freddie—even if DU/LPA recommends "Approve".

In practice, concerns over legal liability and regulatory scrutiny—especially around disparate impact under the Fair Housing Act (FHA) and the Equal Credit Opportunity Act (ECOA)—create strong incentives for lenders to preserve human involvement in underwriting to avoid being exposed to greater legal risk.[41] Black-box algorithms make it difficult to explain, defend, or audit credit decisions against claims of discrimination or unfairness. For instance, Wells Fargo's Enhanced Credit Score (ECS) algorithm has faced litigation alleging discrimination (*Williams et al. v. Wells Fargo Bank, N.A., Case No. 3:22-cv-00990-JD (N.D. Cal.)*), with Wells Fargo defending it as an internal workflow tool rather than an automated decision system (Bloomberg, 2024). Such legal pressures strongly suggest that human-in-the-loop practices will persist in consumer lending,[42] preserving the central role of human underwriters.

4 Main Experiment: Design

In our main experiment, we recruit loan specialists to make hypothetical approval decisions based on real loan applications, testing whether providing AI signals universally coarsened at the approximately optimal threshold $\hat{t}$ from Section 2 improves decision-making outcomes.

[41] The Consumer Financial Protection Bureau (CFPB) explicitly states that advanced technologies do not exempt institutions from compliance with federal consumer financial laws like ECOA (CFPB, 2025).

[42] Non-consumer loans like small business loans receive less protection. Even so, the FDIC's Small Business Lending Survey shows that FinTech primarily augments rather than replaces human decision-making (FDIC, 2024).

4.1 Experimental design

Figure III summarizes the flow of our main experiment. Participants are presented with a series of incentivized binary choices: approve the loan to applicants expected to make timely payments, and deny the loan otherwise. Each participant completes six blocks of ten loan-approval decisions: one practice block (without AI assistance and unincentivized), followed by five order-randomized treatment blocks that vary the type of AI assistance participants receive (if any) in addition to seeing loan applications:

1. *Human-only*: no AI assistance.
2. *Probability*: the AI's estimate of the likelihood the person would repay the loan on time, displayed as a numerical score from 0–100.
3. *Binary Low (30)*: a binary message coarsening the AI score at threshold $t = 30$.
4. *Binary Medium (50)*: a binary message coarsening the AI score at threshold $t = 50$. Given the incentive structure, this is the approximately optimal threshold $\hat{t}$ from Section 2.[43]
5. *Binary High (70)*: a binary message coarsening the AI score at threshold $t = 70$.

In each of the three *Binary* conditions, participants are shown one of two messages: "Yes, this applicant's AI score is above [threshold]" or "No, this applicant's AI score is below [threshold]". Note that Yes/No may connote a recommended action, even if not explicitly stated (they are statements about repayment likelihood). In general, any practical implementation of coarsening an information structure inherently functions as an implicit recommendation. While both the pooling and the recommendation channel (e.g., framing) are likely at play, our framework is agnostic to the two readings, as made clear in Section 2.

To increase power, all participants complete all five treatment blocks: treatment blocks appear in randomized order across participants.[44] The order of the underlying loan applications is also randomized, so cases differ across treatment conditions for different individuals.

For each decision, participants are shown the loan application (described in Section 4.1.1) and, where applicable, the AI message (described by treatment blocks above and in Section 4.1.2).

[43] In particular, when s is a calibrated repayment probability, $\hat{t} = 100 \cdot \frac{c_{FP}}{c_{FP}+c_{TP}} = 50$ which coincides with the natural midpoint heuristic. But this cuts both ways: the same features make "approve above 50" an equally obvious rule of thumb under *Probability*—indeed, we calibrated the score precisely so its numbers would be easy to interpret. Design features that help participants act correctly under *Binary Medium* thus operate similarly under *Probability*, and it is not clear whether communicating an uncalibrated score would have raised or lowered the effect of coarsening relative to full revelation.

[44] As a robustness check, we report block-level results in Appendix C.3.3 and show that *Binary Medium* consistently outperforms *Probability* across all five blocks.

After reviewing the available information, they respond to two questions:

- "What do you think is the % chance that the applicant will make their future payments on time? [Click on the blue bar to choose a number between 0 and 100.]"
- "Do you want to approve this applicant for a loan? [Approve / Deny]"

Participants can view the application details and the AI signal (if applicable) and take as much time as they need to submit answers. Before starting the incentivized blocks, participants are briefed on (i) the repayment rate in the sample of loans (set to 66%), (ii) their own performance on the practice block, and, when applicable, (iii) the AI model used in the condition (including its training data, score interpretation, performance metrics, and signal density conditional on repayment outcomes), and (iv) the AI's decision rule and behavior (e.g., threshold and repayment rate by message for *Binary* formats, or average scores by repayment status for *Probability* scores). This information is available to participants throughout the relevant treatment block via a "More details" tab.[45] Screenshots from the experimental interface are in Appendix C.1.

Incentives are structured to equalize the benefits of true positives/negatives: a \$0.10 bonus is awarded for each loan correctly denied (default applicants) and correctly approved (on-time payers). We make this incentive structure salient, and participants undergo comprehension checks. Posteriors are additionally incentivized via a binarized scoring rule (Hossain and Okui, 2013).

To measure beliefs formed based on the loan application alone (i.e., human signal r from the theory section), we cross-randomize two additional design features at the case-level: (i) whether the AI signal or loan application is shown first {AI (10%), Loan (90%)}, and (ii) whether posterior beliefs are additionally elicited before the second signal is shown {Elicit after only (10%), Elicit both before and after (90%)}. This also allows us to test for potential anchoring effects. Appendix C.3.2 shows that neither dimension systematically affects participants' decisions or final posteriors.

4.1.1 Loan application details

The loans in the experiment come from a dataset of real personal loan and mortgage applications from the multinational financial institution Home Credit, which contains extensive variables including historical loan applications and repayment history, credit bureau records (previous loans,

[45] These briefings necessarily differ with the message format: the *Binary* conditions disclose repayment rates conditional on each message—an action-relevant reliability statistic with no exact analogue in *Probability*.

credits, cash), demographic information (e.g., gender, age, car/house ownership, residential information, family, education, and employment), credit card information, normalized credit scores.

To reduce cognitive load, we choose a subset of the loan application to display, identified through expert interviews with loan specialists and results from pilot studies. We chose variables pertaining to the primary factors that loan specialists use to assess applications, such as the loan amount, loan annuity, age, months in current employment, income (yearly), and normalized credit scores from three sources, each ranging from 0–100, with higher values indicating greater creditworthiness (see Figure C.6 in the Appendix for an example).

4.1.2 AI signal details

The AI predictions are generated from a top-performing AI model trained on the Home Credit dataset of 300,000 loan applicants and their repayment outcomes.[46] The model is trained on all variables in the loan application available to human participants, as well as additional ones, as is explicitly made clear in the instructions (see Appendix Figure C.1). The model scores each loan application from 0–100; we convert this score to a Bayesian posterior, so what subjects see is an empirically calibrated model-derived repayment probability (expressed as a percentage). Crucially, the model's scores are genuine out-of-sample predictions: we randomly sample the 60 experimental applications (including 10 used for practice) from the model's held-out test set, stratified by repayment outcome so that the base repayment rate is $p = 2/3$ by design.

Repayment outcomes in the Home Credit dataset are observed only for applicants approved in the original Home Credit setting, and our AI model is trained on this subset. This does not threaten the internal validity of our experiment, as treatment is randomly assigned within the same application pool. Participants evaluate a random holdout sample from the Home Credit dataset with an application distribution that matches that of the training data, mitigating traditional selective label concerns.[47] External validity is only threatened if differences in this application pool (as a result of the initial selected labels) interact with our treatments to skew results. Moreover, firms deploying calibrated coarsening will typically have access to richer datasets—including data on applicants for whom the repayment outcome is unobserved—to more directly mitigate these issues

[46] This model had the best performance in a public competition on the same dataset. It employs Bayesian optimization to automate the tuning of hyper-parameters, resulting in a cross-validation AUC-ROC score of 0.79.

[47] Participants were informed that the repayment rate in the experiment was 66%—substantially lower than the 90%+ rates typical in their day-to-day work—making them aware of the selected nature of the sample; we also conducted comprehension checks to verify participants understood this base rate.

through reweighting or other methods.

4.1.3 Post-decision elicitations

Participants complete a post-decision block where we elicit perceived residual info, demand for AI, biases, and demographics, detailed in Appendix C.1.1: perceived residual info is elicited right after completing the *Probability* block; the rest after completing all six decision blocks.

Perceived Residual Info. At the end of the *Probability* block, participants estimate how many of 100 loan applications they believe they would decide correctly under each of three information conditions: with both the application information and the AI score, with the AI score alone, and with the application information alone. We take the ratio of the first estimate to the second—the perceived proportional gain from adding the application to the AI score—as a proxy for the participant's perceived residual information.

Demand for AI. For two final decisions, participants choose whether to receive AI assistance and, if so, in which form: the continuous *Probability* score or a *Binary* message at a threshold of their choosing. We then elicit their willingness to pay for each form through an incentivized multiple price list (offers up to 10 cents, the bonus for a correct decision), randomly implementing one row for the two decisions.

Biases and Demographics. Participants complete a battery of incentivized cognitive tasks measuring susceptibility to the belief-updating errors most relevant to combining signals—base-rate neglect, correlation neglect, AI underweighting, and balls-and-urns updating—adapted from Enke et al. (2023); Piccione and Rubinstein (2026). We elicit these outside the loan context to avoid priming, which may attenuate their measured association with loan decisions relative to their true influence. Lastly, we collect standard demographics and loan-industry experience.

4.2 Survey sample and logistics

Our main experiment took place between October 3 and November 4, 2024. We recruited 150 US-based loan specialists with at least one year of experience and at least 18 years of age. Given our focus on improving real-world decision-making outcomes, we prioritized recruiting practicing

professionals that use such tools in practice. For the majority of participants ($N = 97$), recruitment primarily took place through temporary job postings on LinkedIn and Indeed and by targeted outreach through industry channels to members of the National Association of Mortgage Underwriters (NAMU) and the National Association of Mortgage Processors (NAMP). The rest of the sample was recruited through contract hiring platforms like Upwork (N = 14) and custom screening on Prolific (N = 39), with stringent (pre-registered) screening processes.[48] Appendix C.3.1 shows that our main results remain qualitatively similar when excluding Prolific and Upwork participants, whose professional status as practicing loan officers is less verifiable.

Table I presents the demographic summary of those who completed our experiment. Loan specialists in our sample have an average of 12 years of experience (median = 10). Most report prior exposure to any form of AI, though only a fifth had experience using supervised AI models, such as those commonly employed in loans, as is the one used in our experiment.

Once through with screening, qualified participants were directed to complete the 60–90 minute survey hosted on the Otree platform via Heroku, comprising 60 loan decisions described above and post-decision elicitations. They were paid a base rate which varied from $16 (Prolific) to $60 per survey (Indeed, LinkedIn), and earned on average an extra 10–25% in bonuses.

4.3 Empirical strategy

We focus on loan approval accuracy as our main outcome: by equalizing the cost of false positives and negatives, our incentive structure maximizes accuracy. Our main specification is as follows:

$$Y_{ilt} = \beta_0 + \sum_t \beta_t \text{Treatment}_{il}^t + \epsilon_{ilt} \tag{5}$$

where Y_{ilt} denotes the outcome variable for case i decided by loan officer l in treatment block t and $t \in \{\text{Human Only, Probability, Binary Low, Binary Medium, Binary High}\}$. $\text{Treatment}_{il}^t \in \{0, 1\}$ indicates that the case i decided by loan officer l is in the treatment block t, setting *Binary Medium* as the reference base level. We estimate the regression with and without specialist and case (and block-order) fixed effects. The treatment effect is identified by random assignment of cases to treat-

[48] Prolific loan specialists were recruited after October 21, 2024, when we updated our pre-registration to include Prolific participants that: (1) report > 1 year of loan-related experience, (2) pass a check that excludes those who have taken the survey twice or failed screening previously, and (3) either confirm their years of work experience or specific workplace branch, *or* demonstrate familiarity with loan specialist responsibilities in their decision descriptions (e.g., reference income-to-debt ratio, income stability, capacity, or collateral). Besides the first condition, these filtering criteria were not made known so as to reduce gaming.

ment blocks. Our main objects of interest are the β_t coefficients. In particular, as pre-registered, we test (i) whether participants perform differently in *Binary Medium* relative to *Probability* (i.e., testing the null hypothesis that $\beta_{\text{probability}} = 0$) and (ii) whether participants perform differently in *Binary Medium* relative to *Human-only* treatment (testing the null $\beta_{\text{Human-only}} = 0$). Since the theoretical prediction being tested is directional—coarsening at $\hat{t}$ should *improve* outcomes—so the alternative of interest is $\beta_t < 0$ for each comparison condition. We therefore also report left one-sided randomization-inference p-values, which test this directional prediction.

The above tests were pre-registered in our pre-analysis plan (AEARCTR-0013716). The plan also included a prediction that *Binary Medium* would perform better than the *Binary High and Low*, although we remained agnostic to whether *Binary High/Low* would outperform *Human-only* or *Probability*.

5 Main Experiment Results: Universal Calibrated Coarsening Improves Outcomes

5.1 Main results

We start by testing our main hypothesis: providing loan specialists with signals coarsened at the AI-only threshold $\hat{t}$ leads to better decision-making outcomes than human judgment alone, providing the full AI output, or other binary AI coarsenings.

5.1.1 Impact of coarsening on decision-making outcomes

Figure IV compares the decision accuracy of participants across the different treatment blocks, and Table II presents regression results from estimating Equation 5, using *Binary Medium* as the reference. *Binary Medium* outperforms all other treatments: the share of correct decisions in this treatment is 74%, compared to 70% in *Probability* and 68% in Human only. The other Binary treatments, Binary High and Low achieve 67% and 69%, respectively. Each alternative has a negative coefficient, and the ranking is stable across specifications—our pre-registered specification with individual and case fixed effects (Column 3) and one that additionally includes block-order fixed effects (Column 2).[49] Testing the theory's directional prediction, the left one-sided randomization-

[49] We focus on the latter for inference, since treatment was randomized across block positions within participants, so block order is the natural unit to condition on. Reflecting this design, we also report randomization-inference

inference p-values in the preferred specification (Column 2) show *Binary Medium* outperforming *Probability* ($p = 0.030$), *Human-only* ($p < 0.001$), *Binary High* ($p < 0.001$), and *Binary Low* ($p = 0.005$). Under the two-sided nulls, all contrasts remain significant except *Probability*, which is marginal ($p = 0.059$; 0.063 in Column 3).

To contextualize the magnitude of these effects, accuracy under *Human only* is 68%—only 1.3pp, and not statistically significantly, higher than the accuracy from approving every case (the prior/repayment rate is 66.7%; dashed red line). Access to the continuous AI score raises accuracy by (non-statistically-significant) 2pp; the effect of *Binary Medium*—6pp over *Human Only*—is three times as large and highly statistically significant. We can think of this additional 4pp over *Probability* as the value of calibrated coarsening over the provision of full information. *Binary Medium* narrows the gap to the Bayesian benchmark of 80%, which represents the accuracy a fully Bayesian agent would achieve with complete access to AI probability signals (dashed green line estimated from the empirical joint distribution; further details in Section 5.2.1 below).[50] In other words, as a rough gauge of magnitude, *Binary Medium* closes about 40 percent of the gap between *Probability* and the Bayesian benchmark, and about half the gap between *Human-only* and the Bayesian benchmark.

However, for binary coarsening to improve accuracy, it needs to be *calibrated*: coarsening at the arbitrary alternative thresholds *Binary High* and *Binary Low* yield accuracy rates comparable to *Human-only*, suggesting that not all forms of coarsening are beneficial.

5.1.2 Impact of coarsening on time spent

Binary coarsening improves decision accuracy, but how does it affect decision time? In real-world applications where specialists are evaluated on both quality and quantity, this is an important consideration. Because decision times are right-skewed by occasional pauses, we summarize each participant by their median round in each condition and report the across-participant mean of these medians.[51] Figure V compares the average median decision time across conditions. In *Binary*

p-values from 10,000 design-consistent permutations of the treatment–block assignment.

[50] We use the Bayesian benchmark purely as a descriptive yardstick to convey the magnitude of the treatment effects, not for inference: the point estimate is computed in-sample from the experiment's realized outcomes and carries sampling uncertainty. Also note that this is slightly above the AI-only benchmark of 79%, as the human has slight residual information over the AI here. Finally, because coarsening at $\hat{t} = 50$ preserves the AI-only optimal action and the human signal carries little residual information, the analogous benchmark for a Bayesian who observes only the coarse *Binary Medium* message is similar to the full-information benchmark.

[51] Per-case decision time is measured as the interval between the onset of consecutive cases (nine intervals per ten-case block). We use between-onset intervals because a bug left the direct per-screen timer unreliable.

Medium, the median decision takes 37.5 seconds on average, roughly 7% faster than in *Probability* (40.5; paired t-test, $p = 0.043$), consistent with a reduction in cognitive load.

5.1.3 Demand for coarsened AI

Beyond efficiency gains, a practical question is whether, given the choice, specialists would choose *Binary Medium*— or more generally, whether they choose the signal presentation that is most effective for them. When given a choice between receiving no AI assistance, a binary signal (at their preferred threshold), or the full probability score, fewer than half (46%) chose the full probability score, with the rest preferring a binary signal (30%) or no AI at all (23%; Appendix Table C.1). These choices are mirrored in participants' incentivized willingness to pay (WTP): those who opted for binary and probability assistance were willing to forgo 38% (3.8 cents) and 25% (2.5 cents) of their per-case bonus to access them, respectively—*more* than the accuracy gains these signals deliver—whereas those who declined AI would pay less than a cent for a binary signal and would have to be paid to take the probability score.

Participants also appear to select into the conditions most effective for them in terms of whether an AI output is provided, despite receiving no feedback on their performance across those blocks. The majority of those whose highest-payoff treatment block was *Human-only* chose not to use AI, while the majority of those who did best under *Probability* chose to use the probability score. However, results among those who did best in one of the three binary treatments are more mixed: about half prefer the continuous *Probability*, and only a third the binary presentation. This suggests that participants can reasonably anticipate *whether* AI assistance benefits them, but are less able to predict that a coarser presentation would serve them best. Appendix C.2.1 details the distribution of preferred thresholds, the full choice breakdown, and how choices relate to demographics.

5.2 Understanding how participants make decisions

How do loan specialists make decisions in our experiment? One important aspect is whether participants' behavior is consistent with the incentive structure (given beliefs). Since correct approvals and denials are rewarded equally, payoff-maximizing participants should approve when their posterior exceeds 50% and deny otherwise. Appendix Figure C.11 shows that pooled approval rates drop sharply around the 50% posterior threshold, and Appendix Figure C.12 shows that this pattern holds for most participants. This suggests that most deviations from the Bayesian benchmark

stem from belief formation, rather than non-standard preferences, decision noise, or miscalibrated thresholds.

Having established that participants generally apply consistent action thresholds, we examine how they respond to variation in the binary AI thresholds. Appendix Figure C.13 shows that approval rates increase systematically with the binary threshold: higher thresholds increase approvals under a "Yes" recommendation while weakening the force of a "No" and reducing denials. Thus, participants respond to changing thresholds. Moreover, they do not follow the AI blindly—approval rates remain well below 100% for "Yes" recommendations and above 0% for "No" recommendations—indicating that participants incorporate, but do not fully adhere to, the AI binary signal.

5.2.1 The Bayes-optimal benchmark

A natural benchmark for these decisions is the *Bayes-optimal* boundary—what a Bayesian would do if they had access to the continuous AI score. We estimate it directly from realized repayment outcomes, as a logistic regression of the repayment state on the two signals (r, s), whose 0.5 posterior contour is the Bayesian decision boundary in Figure VIa.[52] Because both signals are expressed on a common 0–100 probability scale, we fit this regression in the signals' log-odds and, alternatively, in the raw, 0–100 (r, s) units; the two specifications yield boundaries that broadly coincide over the support of the data.[53] The weight on r in the estimated boundary is statistically non-zero, so specialists do have *some* residual information about repayment beyond the AI score—but the boundary is close to vertical: the coefficient on r is more than an order of magnitude smaller than β_s, and far below the $\beta_r = 1$ implied by conditionally independent signals.[54] Elicited r thus contains little residual information over s. This may reflect the information structure in our setting, where specialists observe a subset of the variables the AI was trained on, but can also be partly driven by noise in the elicitation of r, which would attenuate its estimated informativeness.

[52] This regression is fit on the cases in which the human signal r is elicited over the loan application. In the *Human-only* condition r is elicited every round; in the AI-assisted rounds it is elicited for $\sim 81\%$ of cases, since 90% present the loan application first (before the AI signal) and we independently randomize whether the posterior is elicited after the first signal in 90% of those—about 85% of all cases in total. The estimated posterior $\Pr(\omega = 1 \mid r, s)$ then yields the payoff-maximizing rule: approve whenever it exceeds 0.5.

[53] We draw the log-odds boundary only over the central 5th–95th percentile of the elicited human signal r. Because this specification is linear in the *log-odds* of the signals, it is highly sensitive to the logit transform near the endpoints of the $[0, 1]$ scale—precisely the region where observations are sparsest.

[54] In the log-odds specification $\Pr(\omega = 1 \mid r, s) = \Lambda\big(\beta_0 + \beta_r \operatorname{logit}(r) + \beta_s \operatorname{logit}(s)\big)$, we estimate $\hat{\beta}_r = 0.09$ (s.e. 0.03) alongside $\hat{\beta}_s = 1.22$ (s.e. 0.04).

5.2.2 Estimating decision boundaries: a hierarchical model of specialist decisions

How do participants in our experiment actually combine their own information with AI signals to make decisions? To investigate this, we estimate each specialist's decision boundary from their choices using a hierarchical latent-variable model (HLM) (Gelman et al., 1995; Zitzmann et al., 2021), which does not rely on asymptotic approximations and can yield more reliable estimates in settings with sparse data. This allows us to adjust flexibly to individual heterogeneity while performing well under the data constraints of our setting, where we have less than 10 observations per specialist-treatment for which the human signal (r) is elicited. Specialist i makes an approval decision $a_{i,T,n} \in \{0, 1\}$ (approve or deny) on loan n under treatment T, which we model as arising from a latent utility $y^*_{i,T,n}$ that is not directly observed. The specialist approves when this utility is positive, $a_{i,T,n} = \mathbb{1}[\, y^*_{i,T,n} \geq 0\,]$, where

$$y^*_{i,T,n} = \alpha_{i,T} + r_{i,T,n} + \beta_{i,T}\, \Phi_T(s_n) + \epsilon_{i,T,n}, \qquad \epsilon_{i,T,n} \sim \mathcal{N}(0, \sigma^2_{\epsilon,T}).$$

Here $r_{i,T,n}$ is the specialist's human signal, s_n is the underlying continuous AI score for loan n, and $\Phi_T(s_n)$ is the AI message it induces under treatment T (under binary coarsening at threshold t, $\mathbb{1}[s_n \geq t]$; under *Probability*, the raw score s_n; under *Human-only*, 0). The intercept $\alpha_{i,T}$ captures baseline approval propensity, $\beta_{i,T} > 0$ is the weight placed on the AI signal, and $\epsilon_{i,T,n}$ is idiosyncratic decision noise with treatment-specific variance $\sigma^2_{\epsilon,T}$. The coefficients $\alpha_{i,T}$ and $\beta_{i,T}$ vary across specialists and are drawn from a normal distribution whose mean depends on specialist-specific covariates. We estimate the model separately for each treatment condition T, recovering one decision boundary per specialist-treatment.[55] The hierarchical structure pools information across specialists; we defer the full specification to Appendix C.4, which details the general K-coarsening message Φ_T (including the trinary policies used in Section 6), the treatment of r when unelicited, the monotonicity constraint $\beta_{i,T} > 0$, the per-treatment decision-boundary expressions, and estimation via Hamiltonian Monte Carlo.

Using data from the main experiment,[56] we estimate the model and draw $S = 1000$ posterior samples of the model parameters. For each specialist-treatment pair, we compute and plot the *median* decision boundary in the (r, s)-plane—the set of points at which their latent utility $y^*_i =$

[55] This preserves the flexibility for the Receiver's bias to vary with the form of disclosure.

[56] The HLM is estimated on the $N = 149$ specialists who completed the end-of-survey demographics, which enter as covariates in the hierarchical prior; one of the 150 attrited right before the end of the survey.

0.[57] We then plot the median $\tilde{r}(s)$ across loan specialists, i.e., the specialists' biased *full-revelation* decision boundary under *Probability* (Figure VIb, teal) and their *coarsening* decision boundary under *Binary Medium* (Figure VIc, purple step). In Appendix C.2.6 we produce similar plots from a non-parametric Nadaraya-Watson kernel regression of the approval decision on (r, s), which yields similar boundaries.

Figures VIb—VId are the empirical analogs of the model-section boundary graphs Figures Ia, Ib and Id, estimated from observed behavior. In panel (b), the specialists' *Probability* boundary (teal) is far flatter than the near-vertical Bayes-optimal boundary (black): compared to a Bayesian decision-maker, specialists place too little weight on the AI signal, the empirical counterpart of the (biased) AI underweighting boundary in the model. The shaded wedge between the two DBs is the *scope for improvement*—decisions on the wrong side of the Bayesian boundary depart from the benchmark, as *deny* (red) decisions to its right should have been approvals and *approve* (green) decisions to its left should have been denials. Coarsening to *Binary Medium* (panel (c)) pulls specialists toward the benchmark in some regions and pushes them away in others,[58] which nets out as the accuracy gains and losses in panel (d).

5.2.3 Where gains from coarsening are concentrated

Which decisions are specialists good at, or particularly struggle in? Appendix Tables C.23–C.25 break down performance by true/false positives and negatives, stratified by AI and human signal bins. In general, loan specialists are better at correctly approving than correctly denying applicants. *Binary Medium* primarily improves performance by converting false negatives—helping participants approve creditworthy applicants who might otherwise be denied (also seen in Figure VIb). Consistent with the theory, the effect is concentrated in the middle bands (31–70): (i) where participants' own signal reflects high uncertainty, strengthening the AI message has more room to affect decisions, and (ii) where the AI signal is itself uncertain, pooling sharpens it in both directions—scores just above 50 are pooled upward into a firmer "Yes," and those just below into a firmer "No."

[57] For the *Probability* treatment, we use the median slope and intercept; for the other treatments, we take the median of each relevant point of the adherence curve (e.g., $\tilde{r}_0(t), \tilde{r}_1(t)$).

[58] Appendix Figure C.14 provides supplemental evidence: it plots the difference in approval probability between *Binary Medium* and *Probability* by underlying AI score. Participants under *Binary Medium* are more likely to approve when the AI score exceeds 50 (following a "Yes"), with the largest effects for AI scores close to the threshold.

5.3 Heterogeneous Treatment Effects

For whom is AI assistance—whether coarsened or not—most effective, and what mechanisms drive differential responsiveness to AI? To explore these questions, we examine heterogeneous treatment effects (HTEs). We emphasize at the outset that the results in this section should be taken as suggestive; our experiment was not powered to detect HTEs as per our pre-analysis plan. Figures VII and VIII show the effect of each AI assistance on accuracy, relative to the *Human-only* baseline, across subgroups defined by demographic and bias-related moderators. Corresponding regressions (Equation 5) are in Appendix C.2.2.

5.3.1 Heterogeneous treatment effects by demographics

First, notice that in nearly all of these cuts, *Binary Medium* outperforms *Probability*, which in turn outperforms *Human-only*.

Men benefit more from AI assistance: *Probability* increases men's accuracy by 4.3 percentage points (pp) over *Human-only*, compared to 1.3pp for women. The gap is even higher for *Binary Medium*, where men gain 9.4pp relative to just 2.5pp for women. These differential gains are partly explained by differences in baseline performance: without AI assistance, women exhibit higher accuracy (69.7%) than men (65.3%), suggesting that men have more room to benefit. The gender gap in gains from AI aligns with existing literature suggesting that men may be more willing to adopt, adapt to and leverage AI-generated recommendations (Bick et al., 2026; Carvajal et al., 2024; Humlum and Vestergaard, 2024; Liu and Wang, 2024), though some evidence suggests that over time, this gender gap in adoption is narrowed (Chatterji et al., 2025).

Other demographic factors also matter. Older loan specialists perform somewhat better across all treatments (0.9pp advantage in *Human-only*). However, age does not meaningfully moderate treatment effects. In contrast, experience seems to matter: more experienced specialists have higher baseline accuracy (1.1pp), but their gains from both *Probability* and *Binary Medium* are smaller (1.6 vs. 3.1pp, and 3.8 vs. 7.1pp, respectively). Existing evidence on whether experience complements or substitutes for AI assistance is mixed: in some contexts, less experienced workers benefit disproportionately (Dell'Acqua et al., 2023; Brynjolfsson et al., 2025) and other contexts, domain expertise complements AI (Otis et al., 2024).

Perceived residual information also shapes responsiveness to AI. Baseline accuracy in *Human-only* is nearly identical for those above- and below-median perceived residual information (68.0%

vs. 67.7%), yet, below-median types gain more from both *Probability* (3.8 vs. 1.9pp) and *Binary Medium* (7.5 vs. 4.2pp gains). This suggests that those who (correctly) recognize their limited residual information are more responsive to the raw AI signal, and more adherent to binary one.

5.3.2 Heterogeneous treatment effects by cognitive biases

Moving on to cognitive biases, while we cannot isolate a single dominant bias driving the gap between actual performance and the Bayesian benchmark, examining how different bias measures moderate performance sheds light on the mechanisms through which AI assistance—whether coarsened or not—yields differential gains. For definitions of the bias elicitation variables, as well as the corresponding survey questions, see Appendix C.1.2.

The clearest pattern emerges for AI misweighting. Because there is little residual human information beyond the AI in this setting, the relevant distortion is not AI overweighting but AI underweighting—relying too heavily on one's own information rather than the AI's signal. Baseline accuracy under *Human-only* is similar across groups, but responses to AI diverge sharply: *Binary Medium* outperforms *Probability* (a $+7.0$ vs. $+2.4$pp gain over *Human-only*). This pattern reverses among those who underweight AI less, who perform better with the *Probability* treatment ($+4.3$ vs. $+2.1$pp). This reversal aligns with the conceptual insights from Proposition 1: coarsening helps more when specialists underweight the AI signal more heavily.

Most other measures point a similar way—the more biased subgroup tends to gain more from *Binary Medium* relative to *Probability*—though none of these interactions is individually distinguishable from zero. It is sharpest for correlation neglect: among high-CN specialists, *Probability* delivers essentially no gain ($+0.7$pp) while *Binary Medium* raises accuracy by 5.9pp. None of the remaining splits is statistically distinguishable.

Taken together, these findings suggest that suboptimal use of AI is not driven by any single bias, but by a mix of behavioral deviations. Coarsening can mitigate distortions driven by multiple biases at once, motivating our next step: rather than making structural assumptions about any particular bias, we adopt a data-driven personalization strategy that uses participants' choices to identify the most effective coarsening policy for each participant.

5.4 Heterogeneity in specialists' estimated individual decision boundaries

The substantial heterogeneity in how specialists respond to different types of AI assistance suggests that a one-size-fits-all approach may leave additional gains on the table. This variation is evident in the specialists' estimated decision boundaries: Figure IX plots each specialist's decision boundary (gray) based on posterior samples from the HLM-estimated distribution, making the heterogeneity in individual behavior evident: slopes and intercepts vary widely under *Probability* (Figure IXa), as does adherence under *Binary Medium* (Figure IXb). Appendix Figure C.18 shows the analogous plots for the remaining treatments.

5.5 Simulating the gains from personalized policies

To gauge whether we can use this variation to further increase the gains, we estimate the HLM detailed in 5.2.2 and simulate each specialist's accuracy under each policy—the three binary coarsenings, *Probability*, and *Human-only*. Assigning each specialist their predicted best policy (process detailed in Appendix C.4) allocates 52% to *Binary Medium*, 34% to *Probability*, 10% to *Binary High*, and 3% to *Binary Low*, and raises accuracy by $\sim$ 2pp over the universal *Binary Medium* rule—a suggestive upper bound on the value of personalizing disclosure policies.

6 Estimating the Gains from Personalized Coarsening Policies

We then run a second, two-stage experiment deploying personalized policies with a new set of loan specialists. Our focus is on *how* to coarsen: since our previous experiment shows that coarsening improves on *Probability* and *Human-only* on average, the next question we address is whether personalizing the coarsening policy delivers in practice. In this experiment, we also expand the set of disclosure policies we consider to trinary coarsenings. We find no average improvement from personalization over the universal benchmarks. Several forces plausibly contribute: the predicted gains were modest ($\sim$2pp)—below what the experiment was powered to detect—and decision patterns shifted between estimation and deployment. These findings underscore the practical difficulty of personalizing in dynamic environments, and suggest that a well-calibrated universal coarsening rule may already capture much of the available benefit: a simple, robust, and effective way to improve AI-assisted decisions.

6.1 Design

In *Stage 1*, each participant makes 50 incentivized decisions (plus 10 practice) across five coarsening conditions in randomized order—*Binary Low*, *Binary Medium*, *Binary High*, *Trinary Low (30, 50)*, and *Trinary High (50, 70)*—with AI signals shown on a dial to hold visual design fixed across binary and trinary formats. We oversampled cases in the 10-30th and 60-80th percentiles of the *predicted* human signal—regions near the margins of adherence, where non-*Binary Medium* treatments are expected to have greater impact—to increase statistical power (sampling weight = 2). We use Stage 1 data to estimate the model of Section 5.2.2 and form each participant's potential personalized assignment based on out-of-sample predictions from the model.

Stage 2, run within a month of stage 1, forms our main treatment randomization: each participant is randomly assigned to one of three arms—*Personalization* (50%; the AI signal shown using the individual's personalized assignment, estimated using Stage 1 data as detailed by the process in Section 5.2.2), *Binary Medium* (25%; the universal policy from the main experiment), or *Top-performing* (25%; the best Stage 1 population policy, here *Trinary High*)—and decides 30 new applications, all shown under their assigned policy. This compares individualized assignment against two universal "off-the-shelf" benchmarks. Our specification follows Equation 5, with the difference being the set of treatment conditions $T \in \{$*Binary Medium, Top Performing, Personalization*$\}$. We use *Personalization* as the reference arm and test whether it differs from *Binary Medium* and from *Top-performing*.

Logistics. The experiment ran between May 17 and June 24, 2025 over Prolific, using a two-stage design; full logistics are in Appendix C.5. We recruited 218 participants under screening consistent with that for Prolific participants in Section 4, relaxing the workplace-familiarity requirement to scale recruitment but adding a safeguard that filters out—before the Stage 2 randomization—participants whose Stage 1 accuracy falls below 60% (pre-registered, AEARCTR-0015758).[59] 173 of the 218 completed both stages (60 min and 40 min on average, respectively), reporting 6 years of experience as loan specialists on average, with more than half having prior experience working with supervised AI models (Appendix Table C.28). Participants received base payments of $16 (Stage 1) and $10 (Stage 2), plus an average bonus of about $6.

[59]The prior probability of repayment is 66%, so a participant approving every loan would score 66%; we set the filtering benchmark at 60% to account for decision noise. Applying it before the Stage 2 randomization ensures it introduces no selection bias across treatment arms.

6.2 Main results

Figure X presents the main results from Stage 1 (left) and Stage 2 (right) of the personalization experiment.[60] In Stage 1, the *Trinary High (50, 70)* condition slightly outperforms the other coarsened policies and is therefore used as the top-performing benchmark going into Stage 2.

Average performance across conditions is low and not significantly different from the prior, even after excluding participants who scored below 60% (as pre-registered).[61] This reflects the fact that we (i) recruited only Prolific participants in this experiment, and (ii) intentionally oversampled harder cases.[62] In addition, since the pre-registered exclusion rule may disproportionately drop lower-performing participants in some treatment arms, we caution against overinterpreting treatment differences—such as those between *Binary Medium* and *Binary High*—in Stage 1.

In Stage 2, we observe no evidence of treatment effects: whether assigned to a personalized policy or not, participants' average performance stays flat at the prior.[63] The red bars represent performance predicted by the decision model introduced in Section 5.2.2; while the model suggests gains from personalization, these gains are not realized in practice.[64]

6.2.1 Model validation

We validate the model in two ways (full details in Appendix C.6.4). First, the model overpredicts personalization gains: the gap between predicted and observed personalization accuracy gain (personalized minus universal) is statistically significant, with a 95% bootstrap confidence interval of $[0.009, 0.057]$ and a 95% Bayesian credible interval of $[0.021, 0.047]$. Second, this overprediction is unlikely to stem primarily from poor model fit—out-of-sample, the model predicts held-out approve/deny decisions 77–84% of the time, and in-sample, specialists perform significantly better under their model-predicted optimal treatment (Appendix C.6.8).[65] Recall that the model is fit to

[60] We present further breakdown of Stage 2 outcomes by specific personalized assignments in Appendix C.6.1.

[61] The corresponding regressions can be found in Appendix Table C.29.

[62] When we reweight the data to reflect the original case distribution, performance levels are higher across the board, as seen in Appendix Figures C.23a and C.23b.

[63] Note the design was not powered to detect the predicted gain: with a clustered standard error of 1.5pp on the personalization contrast, the minimum detectable effect is roughly 4pp, twice the predicted $\sim$2pp.

[64] Model predictions (red) include only participants who complete Stage 2. Attrition differs slightly across arms; Stage 1 accuracy is balanced among completers, though we cannot rule out selection on unobservables (Appendix C.6.3).

[65] We also conduct several other model checks. First, Stage 2 accuracy is positively correlated with model-predicted accuracy, indicating the model is good at predicting who performs well—even if it overstates personalization benefits (Appendix C.6.5). Second, the model passes a placebo test (Appendix C.6.6) randomly reshuffling treatment assignments—under which no gains are predicted—suggesting its predictions are not simply an artifact of noise. Third, re-running the model using observed $\hat{r}$ values in Stage 2 (rather than treating r as latent) yields similar results (Appendix C.6.7), indicating that misprediction of r is not the primary issue.

Stage 1 data from the same participants who are called back in Stage 2, and predicts only treatments it has already seen in Stage 1. Thus, the overprediction of gains must stem either from (i) model misspecification, or (ii) changes in participant behavior between Stage 1 and Stage 2 due to forces not captured in the model.[66] Since the model fits Stage 1 well, we next examine whether shifts in participant behavior between stages explain why personalization delivers smaller gains than expected.

6.2.2 Why is personalization delivering smaller gains than predicted?

Participant behavior can shift between Stage 1 and Stage 2 due to factors not captured by the model—-such as fatigue, forgetting, or context shifts. To test this, we re-estimate the decision model in Section 5.2.2 using only Stage 2 data and compare each participant's adherence curves across stages. We find substantial deviations (Appendix Figure C.25): for some individuals the shift is large—a 20 percentage point change implies, for example, that a participant who in Stage 1 approved whenever their own human signal was above 50% now only approves when it is above 70% in Stage 2. What matters is the wide dispersion, with many participants far from zero. This instability may partly reflect the participant population: all were sourced through Prolific and had fewer years of loan experience than the field-recruited professionals in our main experiment.

Suggestively, personalization performs better where decision strategies remained stable across stages. Specifically, we examine heterogeneity in treatment effects by the magnitude of change in adherence curve estimates (i.e., average absolute differences in r_0, r_1, and r_2, averaging over the boundaries defined under each participant's policy—r_0, r_1 for binary and r_0, r_1, r_2 for trinary coarsenings). Splitting at the median adherence shift, participants with below-median shifts show positive gains of about 2 percentage points (though underpowered for significance), whereas those with larger shifts see no benefit (Appendix Table C.35); an interaction specification points the same way, with the effect of personalization positive and significant at the 10% level when adherence differences are near zero (Appendix Table C.36). [67]

In summary, the predicted gains were modest to begin with—below what the experiment could

[66] A natural third candidate is a winner's curse: selecting each participant's best-looking policy from noisy estimates would mechanically inflate predicted gains. Our procedure guards against this mechanical form: assignments maximize posterior-mean accuracy from the hierarchical model—so individual estimates are partially pooled toward group means—and predicted gains are posterior means of the *selected* policy evaluated out of sample, which are unbiased under the model. Selection optimism therefore enters only insofar as the posteriors themselves are miscalibrated, a possibility subsumed in (i).

[67] Because the adherence shift is estimated from the same Stage 2 decisions as the outcome, we read this split as suggestive.

detect—suggesting that personalization offers only incremental improvements over the universal AI-only binary coarsening in this setting; moreover, behavioral instability between estimation and deployment further limits what any policy tailored to past data can deliver.In environments where such shifts are common, or where decision-makers are unfamiliar with the task, a well-calibrated universal coarsening rule captures much of the available benefit and provides a simple and effective alternative. While personalization may prove more valuable in institutional contexts with long-term behavioral data, our findings suggest that universal calibration is a strong and practical default.

7 Conclusion and Discussion

We show in this paper—both theoretically and empirically—that coarsening AI signals at calibrated thresholds can improve decision-making outcomes, including in high-stakes contexts like loan approvals. Our study offers one of the first empirical applications of information design to AI-assisted decision-making: calibrated coarsening moves professional loan specialists' decisions toward the Bayesian benchmark, showing the important relevance of this approach. The proposed approach is transparent, adaptable, and easy to implement—making it a practical tool firms can use to improve AI-assisted decisions.

A crucial question is whether our results generalize to settings where humans have access to information that the AI was not trained on (e.g., interviews unavailable to the AI). In such settings, personalized coarsening rules may be especially valuable and may offer greater scope for improvement, as they can adapt information presentation to the decision-maker's residual knowledge and context, and the human's unique signals can complement the AI in ways that aggregated policies cannot. A caveat, of course, is that the effectiveness of such personalized rules may be fragile in the face of context shifts or behavioral drift: changes in decision-maker behavior or environment can limit the gains to personalization, as our second experiment suggests.

Calibrated coarsening remains generalizable in a different sense. With rapidly evolving AI technologies—and corresponding information structures and the way humans update from them—coarsening policies can be *recalibrated* (i.e., information re-partitioned) and therefore seamlessly adjusted to these shifts, making it resilient to inevitable technological and organizational change.

Future research could explore dynamic settings with strategic considerations, where the Receiver's objectives diverge from the Sender's due to misaligned incentives or motivated biases (e.g., by gender or race in lending). While our framework can accommodate such incentive mis-

alignments, richer extensions—such as dynamic policies without commitment—introduce new challenges, including strategic behavior by the Receiver that could influence the Sender's information design. Another natural direction is to move beyond binary classification. While our current experiments focus on approval decisions, in practice loan officers may also determine terms such as the loan amount. Different loan amounts can be interpreted as inducing different cost of errors, which can be accommodated in our current framework by assigning case-specific thresholds and aggregating. However, beyond this, a natural extension is to treat loan amounts themselves as part of the action space and extend to any (finite) number of actions to capture richer decision environments.

Additionally, developing dynamic, adaptive systems that adjust information presentation in real time may hold significant promise. Such systems would not only account for how people respond under different information policies (e.g., adhering vs. non-adhering types), but also address the dynamic questions of information acquisition: which loans are approved directly shapes the data available for retraining and improving the AI model. Because information presentation influences approval decisions, information design is not a one-shot exercise—especially in the presence of the selective labels problem—but rather an evolving process in which both human behavior, ground-truth data collection, and model retraining jointly determine long-run performance.

Finally, even in contexts where humans may not hold ultimate decision rights, their residual knowledge and contextual judgment can still enhance AI outcomes. Developing systematic methods to elicit and incorporate such complementary inputs into hybrid decision systems remains a challenge. More broadly, we believe advances in the empirical application of information design will be instrumental in shaping the future of work and decision-making across countless domains.

References

Agarwal, N., Moehring, A., Rajpurkar, P., and Salz, T. (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working Paper 31422, National Bureau of Economic Research.

Agarwal, N., Moehring, A., and Wolitzky, A. (2025). Designing human-AI collaboration: A sufficient-statistic approach.

Agrawal, A., Gans, J., and Goldfarb, A. (2019). *The economics of artificial intelligence: an agenda*. University of Chicago Press.

Agrawal, A., Gans, J., and Goldfarb, A. (2022). *Power and prediction: The Disruptive Economics of Artificial Intelligence*. Harvard Business Review Press.

Alonso, R. and Câmara, O. (2016). Bayesian persuasion with heterogeneous priors. *Journal of Economic Theory*, 165:672–706.

Angelova, V., Dobbie, W. S., and Yang, C. S. (2026). Algorithmic recommendations and human discretion. *Review of Economic Studies*, 93(4):2250–2283.

Anthony, C., Bechky, B. A., and Fayard, A.-L. (2023). "Collaborating" with AI: Taking a system view to explore the future of work. *Organization Science*, 34(5):1672–1694.

Athey, S. C., Bryan, K. A., and Gans, J. S. (2020). The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings*, volume 110, pages 80–84. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Avoyan, A., Khubulashvili, R., and Mekerishvili, G. (2026). Optimal disclosure on crowdfunding platforms. *Review of Industrial Organization*. Online first.

Aybas, Y. C. and Turkel, E. (2019). Persuasion with coarse communication. *arXiv preprint arXiv:1910.13547*.

Bai, B., Dai, H., Zhang, D., Zhang, F., and Hu, H. (2021). The impacts of algorithmic work assignment on fairness perceptions and productivity. In *Academy of Management Proceedings*, volume 2021, page 12335. Academy of Management Briarcliff Manor, NY 10510.

Balakrishnan, M., Ferreira, K., and Tong, J. (2022). Improving human-algorithm collaboration: Causes and mitigation of over-and under-adherence. *Available at SSRN 4298669*.

Banker, S. and Khetani, S. (2019). Algorithm overdependence: How the use of algorithmic recommendation systems can increase risks to consumer well-being. *Journal of Public Policy & Marketing*, 38(4):500–515.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021). Is the most accurate AI the best teammate? optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414.

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11.

Barberis, N. and Shleifer, A. (2003). Style investing. *Journal of Financial Economics*, 68(2):161–199.

Bastani, H., Bastani, O., and Sinchaisri, W. P. (2026). Improving human sequential decision making with reinforcement learning. *Management Science*, 72(1):733–755.

Bell, J. J., Pescher, C., Tellis, G. J., and Füller, J. (2024). Can AI help in ideation? a theory-based model for idea screening in crowdsourcing contests. *Marketing Science*, 43(1):54–72.

Ben-Michael, E., Greiner, D. J., Huang, M., Imai, K., Jiang, Z., and Shin, S. (2025). 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.

Benjamin, D. J. (2018). Errors in probabilistic reasoning and judgment biases. *Handbook of Behavioral Economics*, 2. Available at SSRN: https://ssrn.com/abstract=3269112.

Benjamin, D. J. (2019). Chapter 2 - errors in probabilistic reasoning and judgment biases. In Bernheim, B. D., DellaVigna, S., and Laibson, D., editors, *Handbook of Behavioral Economics - Foundations and Applications 2*, volume 2 of *Handbook of Behavioral Economics: Applications and Foundations 1*, pages 69–186. North-Holland.

Bick, A., Blandin, A., and Deming, D. J. (2026). The rapid adoption of generative AI. *Management Science*. Articles in Advance.

Biermann, J., Horton, J. J., and Walter, J. (2022). Algorithmic advice as a credence good. *ZEW-Centre for European Economic Research Discussion Paper*, (22-071).

Bloomberg (2024). Wells fargo racial disparity case heads to class action decision. *Bloomberg*. Accessed May 6, 2025.

Boleslavsky, R. and Cotton, C. (2015). Grading standards and education quality. *American Economic Journal: Microeconomics*, 7(2):248–279.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2012). Salience theory of choice under risk. *The Quarterly Journal of Economics*, 127(3):1243–1285.

Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, attention, and choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.

Boyacı, T., Canyakmaz, C., and de Véricourt, F. (2024). Human and machine: The impact of machine input on decision making under cognitive limitations. *Management Science*, 70(2):1258–1275.

Braghieri, L. (2026). Paternalistic information design. CEPR Discussion Paper DP21484, Centre for Economic Policy Research.

Brocklehurst, P., Field, D., Greene, K., Juszczak, E., Keith, R., Kenyon, S., Linsell, L., Mabey, C., Newburn, M., Plachcinski, R., et al. (2017). Computerised interpretation of fetal heart rate during labour (infant): a randomised controlled trial. *The Lancet*, 389(10080):1719–1729.

Brynjolfsson, E., Li, D., and Raymond, L. (2025). Generative ai at work*. *The Quarterly Journal of Economics*, 140(2):889–942.

Brynjolfsson, E. and McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). 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.

Bundorf, M. K., Polyakova, M., and Tai-Seale, M. (2024). How do consumers interact with digital expert advice? experimental evidence from health insurance. *Management Science*, 70(11):7617–7643.

BusinessWire (2021). Lendsmart integrates with freddie mac loan product advisor to expedite the underwriting process. https://www.businesswire.com/news/home/20210713005014/en/Lendsmart-Integrates-With-Freddie-Mac-Loan-Product-Advisor-to-Expedite-th Accessed May 6, 2025.

Caplin, A., Deming, D. J., Li, S., Martin, D. J., Marx, P., Weidmann, B., and Ye, K. J. (2026). The ABCs of who benefits from working with AI: Ability, beliefs, and calibration. *Management Science*, 72(7):5843–5852.

Carlile, P. R. (2004). Transferring, translating, and transforming: An integrative framework for managing knowledge across boundaries. *Organization science*, 15(5):555–568.

Caro, F. and de Tejada Cuenca, A. S. (2023). Believing in analytics: Managers' adherence to price recommendations from a dss. *Manufacturing & Service Operations Management*, 25(2):524–542.

Carvajal, D., Franco, C., and Isaksson, S. (2024). Will artificial intelligence get in the way of achieving gender equality? *NHH Dept. of Economics Discussion Paper*, (03).

Cestonaro, C., Delicati, A., Marcante, B., Caenazzo, L., and Tozzo, P. (2023). Defining medical liability when artificial intelligence is applied on diagnostic algorithms: a systematic review. *Frontiers in Medicine*, 10:1305756.

CFPB (2025). Supervisory highlights: Advanced technologies. Technical report, Consumer Financial Protection Bureau. Last accessed March 2025.

Chakraborty, I., Chiong, K., Dover, H., and Sudhir, K. (2025). Can AI and AI-hybrids detect persuasion skills? Salesforce hiring with conversational video interviews. *Marketing Science*, 44(1):30–53.

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., and Wadman, K. (2025). How people use chatgpt. Technical report, National Bureau of Economic Research Working paper.

Chen, Y., Lin, T., Tang, W., and Tucker-Foltz, J. (2025). Explainable information design. *arXiv preprint arXiv:2508.14196*.

Chernozhukov, V., Fernandez-Val, I., and Galichon, A. (2009). Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575.

Choudhary, V., Marchetti, A., Shrestha, Y. R., and Puranam, P. (2025). Human-AI ensembles: When can they work? *Journal of Management*, 51(2):536–569.

Christin, A. (2020). The ethnographer and the algorithm: beyond the black box. *Theory and Society*, 49(5):897–918.

Cockburn, I. M., Henderson, R., and Stern, S. (2018). *The Impact of Artificial Intelligence on Innovation*, volume 24449. National Bureau of Economic Research.

Corradini, V. (2024). Information and access in school choice systems: Evidence from New York City. Working Paper (this version: October 2, 2024).

Dai, T. and Singh, S. (2025). Using artificial intelligence as gatekeeper or second opinion: Designing patient pathways for artificial intelligence augmented healthcare. *Production and Operations Management*. OnlineFirst.

Daugherty, P. R. and Wilson, H. J. (2018). *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press.

Davenport, T. H. and Kirby, J. (2016). *Only humans need apply: Winners and losers in the age of smart machines*. Harper Business New York.

De-Arteaga, M., Fogliato, R., and Chouldechova, A. (2020). A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12.

de Clippel, G. and Zhang, X. (2022). Non-bayesian persuasion. *Journal of Political Economy*, 130(10):2594–2642.

Decker, C. (2022). What's in an airbnb five-star rating? an empirical model of bayesian persuasion. *Working Paper, Job Market Paper*.

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013).

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3):1155–1170.

Enke, B. and Graeber, T. (2023). Cognitive uncertainty. *The Quarterly Journal of Economics*, 138(4):2021–2067.

Enke, B., Graeber, T., and Oprea, R. (2023). Confidence, self-selection, and bias in the aggregate. *American Economic Review*, 113(7):1933–1966.

Enke, B. and Zimmermann, F. (2019). Correlation neglect in belief formation. *Review of Economic Studies*, 86:313–332.

European Commission (2024). EU Artificial Intelligence Act: Article 14. https://artificialintelligenceact.eu/ai-act-explorer/.

Exley, C. L. and Kessler, J. B. (2023). Information avoidance and image concerns. *The Economic Journal*, 133(656):3153–3168.

Faraj, S., Pachidi, S., and Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. *Information and Organization*, 28(1):62–70.

FDIC (2024). Small Business Lending Survey 2024. https://www.fdic.gov/publications/small-business-lending-survey-2024. Accessed May 6, 2025.

FICO (2024). FICO Score History. https://www.fico.com/en/products/fico-score. Accessed May 6, 2025.

Fong, G. T., Krantz, D. H., and Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive psychology*, 18(3):253–292.

Fong, G. T. and Nisbett, R. E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General*, 120(1):34.

Freddie Mac (2025). Loan Product Advisor (LPA). https://sf.freddiemac.com/tools-learning/technology-tools/our-solutions/loan-product-advisor. Accessed May 6, 2025.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2021). Will humans-in-the-loop become borgs? merits and pitfalls of working with ai. *Management Information Systems Quarterly (MISQ)-Vol*, 45.

Gabaix, X. (2014). A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics*, 129(4):1661–1710.

Garrett, B. L. and Monahan, J. (2020). Judging risk. *Calif. L. Rev.*, 108:439.

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., Coughlin, J. F., Guttag, J. V., Colak, E., and Ghassemi, M. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine*, 4(1):31.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Goel, A. M. and Thakor, A. V. (2015). Information reliability and welfare: A theory of coarse credit ratings. *Journal of Financial Economics*, 115(3):541–557.

Green, B. and Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.

Greenes, R. (2007). *Clinical decision support: the road ahead*. Elsevier.

Grether, D. M. (1992). Testing bayes rule and the representativeness heuristic: Some experimental evidence. *Journal of Economic Behavior & Organization*, 17(1):31–57.

Grimon, M.-P. and Mills, C. (2025). Better together? a field experiment on human-algorithm interaction in child protection. *arXiv preprint arXiv:2502.08501*.

Guan, M., Oprea, R., and Yuksel, S. (2023). Too much information. Working paper, UCSB.

Guan, M., Oprea, R., and Yuksel, S. (2025). Beyond instrumental value: How complexity shapes information demand. *Working Paper*.

Guo, Y. and Shmaya, E. (2019). The interval structure of optimal disclosure. *Econometrica*, 87(2):653–675.

Guo, Z., Wu, Y., Hartline, J., and Hullman, J. (2025). The value of information in human-AI decision-making. *arXiv preprint arXiv:2502.06152*.

Guo, Z., Wu, Y., Hartline, J. D., and Hullman, J. (2024). A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Harbaugh, R. and Rasmusen, E. (2018). Coarse grades: Informing the public by withholding information. *American Economic Journal: Microeconomics*, 10(1):210–235.

Hardy, C., Lawrence, T. B., and Grant, D. (2005). Discourse and collaboration: The role of conversations and collective identity. *Academy of management review*, 30(1):58–77.

Hossain, T. and Okui, R. (2013). The binarized scoring rule. *Review of Economic Studies*, 80(3):984–1001.

Humlum, A. and Vestergaard, E. (2024). The adoption of chatgpt. Technical report, IZA Discussion Papers.

Ibrahim, R., Kim, S.-H., and Tong, J. (2021). Eliciting human judgment for prediction algorithms. *Management Science*, 67(4):2314–2325.

Imai, K., Jiang, Z., Greiner, D. J., Halen, R., and Shin, S. (2023). Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 186(2):167–189.

Jabbour, S., Fouhey, D., Shepard, S., Valley, T. S., Kazerooni, E. A., Banovic, N., Wiens, J., and Sjoding, M. W. (2023). Measuring the impact of AI in the diagnosis of hospitalized patients: a randomized clinical vignette survey study. *Jama*, 330(23):2275–2284.

Kahneman, D., Slovic, P., and Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kamenica, E. (2019). Bayesian persuasion and information design. *Annual Review of Economics*, 11(1):249–272.

Kamenica, E. and Gentzkow, M. (2011). Bayesian persuasion. *American Economic Review*, 101(6):2590–2615.

Karlinsky-Shichor, Y. and Netzer, O. (2024). Automating the B2B salesperson pricing decisions: A human-machine hybrid approach. *Marketing Science*, 43(1):138–157.

Kassir, S., Baker, L., Dolphin, J., and Polli, F. (2023). AI for hiring in context: A perspective on overcoming the unique challenges of employment research to mitigate disparate impact. *AI and Ethics*, 3(3):845–868.

Kelley, S., Ovchinnikov, A., Hardoon, D. R., and Heinrich, A. (2022). Antidiscrimination laws, artificial intelligence, and gender bias: A case study in nonmortgage fintech lending. *Manufacturing & Service Operations Management*, 24(6):3039–3059.

Kellogg, K. C., Valentine, M. A., and Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of management annals*, 14(1):366–410.

Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C. P., Ball, R. L., Montine, T. J., et al. (2020). Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ digital medicine*, 3(1):23.

Kim, H., Glaeser, E. L., Hillis, A., Kominers, S. D., and Luca, M. (2024). Decision authority and the returns to algorithms. *Strategic Management Journal*, 45(4):619–648.

Kolotilin, A. (2018). Optimal information disclosure: A linear programming approach. *Theoretical Economics*, 13(2):607–635.

Kolotilin, A., Mylovanov, T., Zapechelnyuk, A., and Li, M. (2017). Persuasion of a privately informed receiver. *Econometrica*, 85(6):1949–1964.

Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. (2021). Towards a science of human-AIdecision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471*.

Lakkaraju, H. and Bastani, O. (2020). How do I fool you? Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85.

Lakkaraju, H. and Farronato, C. (2022). When algorithms explain themselves: AI adoption and accuracy of experts' decisions. Technical report, National Bureau of Economic Research.

Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization science*, 33(1):126–148.

Levina, N. (2005). Collaborating on multiparty information systems development projects: A collective reflection-in-action view. *Information systems research*, 16(2):109–130.

Levy, G., Moreno de Barreda, I., and Razin, R. (2022). Persuasion with correlation neglect: A full manipulation result. *American Economic Review: Insights*, 4(1):123–38.

Li, H., Li, J., Luo, Y., and Zhang, X. (2024). AI persuasion, Bayesian attribution, and career concerns of doctors. Working Paper, arXiv:2410.01114.

Li, H. and Shi, X. (2017). Discriminatory information disclosure. *American Economic Review*, 107(11):3363–3385.

Liu, Y. and Wang, H. (2024). *Who on Earth Is Using Generative AI?* World Bank.

Logg, J. and Schlund, R. (2024). A simple explanation reconciles "algorithm aversion" and "algorithm appreciation": Hypotheticals vs. real judgments. *Real Judgments (January 8, 2024)*.

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

Mahmud, H., Islam, A. N., Ahmed, S. I., and Smolander, K. (2022). What influences algorithmic decision-making? a systematic literature review on algorithm aversion. *Technological Forecasting and Social Change*, 175:121390.

Maron, R. C., Utikal, J. S., Hekler, A., Hauschild, A., Sattler, E., Sondermann, W., Haferkamp, S., Schilling, B., Heppt, M. V., Jansen, P., et al. (2020). Artificial intelligence and its effect on dermatologists' accuracy in dermoscopic melanoma image classification: Web-based survey study. *Journal of Medical Internet Research*, 22(9):e18091.

McGrath, S., Mehta, P., Zytek, A., Lage, I., and Lakkaraju, H. (2020). When does uncertainty matter?: Understanding the impact of predictive uncertainty in ml assisted decision making. *arXiv preprint arXiv:2011.06167*.

McLaughlin, B. and Spiess, J. (2022). Algorithmic assistance with recommendation-dependent preferences. *arXiv preprint arXiv:2208.07626*.

McLaughlin, B. and Spiess, J. (2024). Designing algorithmic recommendations to achieve human-AI complementarity. *arXiv preprint arXiv:2405.01484*.

Mezrich, J. L. (2022). Demystifying medico-legal challenges of artificial intelligence applications in molecular imaging and therapy. *PET clinics*, 17(1):41–49.

Milkman, K. L., Chugh, D., and Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on psychological science*, 4(4):379–383.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., and Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, 2(1):129–140.

Mozannar, H. and Sontag, D. (2020). Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, pages 7076–7087. PMLR.

Mullainathan, S. (2002). Thinking through categories. MIT Mimeo.

Mullainathan, S., Schwartzstein, J., and Shleifer, A. (2008). Coarse thinking and persuasion. *The Quarterly Journal of Economics*, 123(2):577–619.

Noti, G. and Chen, Y. (2023). Learning when to advise human decision makers. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*.

Nunes, I., Ayres-de Campos, D., Ugwumadu, A., Amin, P., Banfield, P., Nicoll, A., Cunningham, S., Sousa, P., Costa-Santos, C., Bernardes, J., et al. (2017). Central fetal monitoring with and without computer analysis: a randomized controlled trial. *Obstetrics & Gynecology*, 129(1):83–90.

Otis, N., Clarke, R., Delecourt, S., Holtz, D., and Koning, R. (2024). The uneven impact of generative AI on entrepreneurial performance. *Working Paper*.

Pachidi, S., Berends, H., Faraj, S., and Huysman, M. (2021). Make way for the algorithms: Symbolic actions and change in a regime of knowing. *Organization Science*, 32(1):18–41.

Piccione, M. and Rubinstein, A. (2026). Failing to correctly aggregate signals. *Games and Economic Behavior*, 158:122–131.

PRNewswire (2024). Tidalwave collaborates with freddie mac to optimize lender and borrower engagement in real-time. https://www.prnewswire.com/news-releases/tidalwave-collaborates-with-freddie-mac-to-optimize-lender-and-borrower-e html. Accessed: 2025-05-06.

Raghu, M., Blumer, K., Corrado, G., Kleinberg, J., Obermeyer, Z., and Mullainathan, S. (2019). The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*.

Rai, A., Constantinides, P., and Sarker, S. (2019). Next generation digital platforms: Toward human-AI hybrids. *Mis Quarterly*, 43(1):iii–ix.

Roy, M. C. and Lerch, F. J. (1996). Overcoming ineffective mental representations in base-rate problems. *Information Systems Research*, 7(2):233–247.

Sah, S. and Loewenstein, G. (2015). Conflicted advice and second opinions: Benefits, but unintended consequences. *Organizational Behavior and Human Decision Processes*, 130:89–107.

Sellier, A.-L. and Dahl, D. W. (2011). Focus! creative success is enjoyed through restricted choice. *Journal of Marketing Research*, 48(6):996–1007.

Sellier, A.-L., Scopelliti, I., and Morewedge, C. K. (2019). Debiasing training improves decision making in the field. *Psychological science*, 30(9):1371–1379.

Snyder, C., Keppler, S., and Leider, S. (2026). Algorithm reliance: Fast and slow. *Management Science*, 72(1):368–385.

Soll, J. B. and Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of experimental psychology: Learning, memory, and cognition*, 35(3):780.

Staradub, V. L., Messenger, K. A., Hao, N., Wiley, E. L., and Morrow, M. (2002). Changes in breast cancer therapy because of pathology second opinions. *Annals of surgical oncology*, 9:982–987.

Stevenson, M. T. and Doleac, J. L. (2024). Algorithmic risk assessment in the hands of humans. *American Economic Journal: Economic Policy*, 16(4).

Sun, J., Zhang, D. J., Hu, H., and Van Mieghem, J. A. (2022). Predicting human discretion to adjust algorithmic prescription: A large-scale field experiment in warehouse operations. *Management Science*, 68(2):846–865.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Taudien, A., Fügener, A., Gupta, A., and Ketter, W. (2022). The effect of AI advice on human confidence in decision-making.

Taylor, P. and Potts, H. W. (2008). Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate. *European Journal of Cancer*, 44(6):798–807.

Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

Von Krogh, G. (2018). Artificial intelligence in organizations: New opportunities for phenomenon-based theorizing. *Academy of Management Discoveries*, 4(4):404–409.

Xiang, J. (2026). Physicians as persuaders: Evidence from hospitals in china. *The RAND Journal of Economics*, 57(2):451–468.

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, 69(3):237–249.

Yaniv, I. (2004). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, 93(1):1–13.

Yaniv, I. and Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, 83(2):260–281.

Ye, W., Bullo, F., Friedkin, N., and Singh, A. K. (2022). Modeling human-AI team decision making. *arXiv preprint arXiv:2201.02759*.

Zhang, W. and Misra, S. (2024). Coarse personalization. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 1206–1208.

Zitzmann, S., Lüdtke, O., Robitzsch, A., and Hecht, M. (2021). On the performance of bayesian approaches in small samples: A comment on Smid, McNeish, Miocevic, and van de Schoot (2020). *Structural Equation Modeling: A Multidisciplinary Journal*, 28(1):40–50.

Tables and Figures

Table I: Participant Demographics by Recruitment Channel

	Industry Channels	Contract Work	Overall
N (participants)	96*	53	149*
Age	46	40	44
Female	60.42%	54.72%	58.39%
Bachelor's Degree or Above	51.04%	79.25%	60.67%
Income	113,047	98,396	107,836
Loan Specialist Experience	100.00%	100.00%	100.00%
Loan Specialist Years	15	7	12
Bank / Credit Union Experience	95.83%	96.23%	95.97%
Applied to Loans Before	95.83%	98.11%	96.64%
Loan Amount (USD$)	222,311	143,398	193,615
Used Supervised AI Before	21.88%	18.87%	20.81%
Never Used any AI	25.00%	22.64%	24.16%

Industry Channels: LinkedIn, Indeed, NAMU/NAMP. Contract Work: Upwork, Prolific. *One participant completed all experimental blocks except the demographics and bias elicitations, and is excluded from this table.

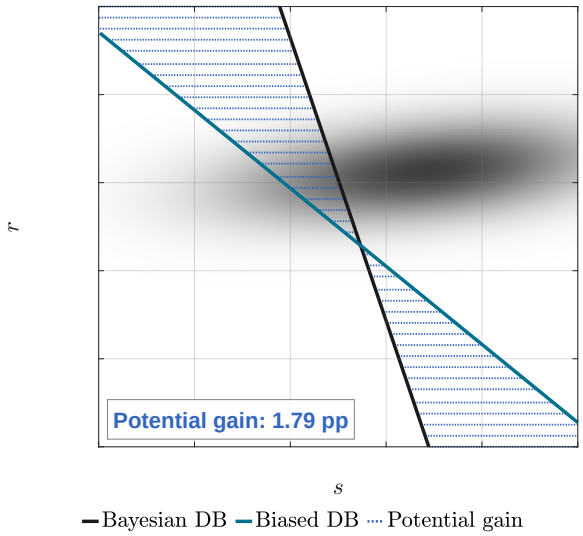
Table II: Decision-making outcomes under different treatment conditions

	Dependent variable: Share of Decisions Correct		
	(1) No FE	(2) Ind.+Case+Block	(3) Ind.+Case
Constant	0.738 [0.012]	—	—
Human-Only	−0.060 [0.016]	−0.059 [0.015]	−0.058 [0.015]
RI p-value (2-sided, left)	(<0.001, <0.001)	(<0.001, <0.001)	(<0.001, <0.001)
Probability	−0.040 [0.018]	−0.030 [0.016]	−0.029 [0.016]
RI p-value (2-sided, left)	(0.023, 0.011)	(0.059, 0.030)	(0.063, 0.031)
Binary High (70)	−0.072 [0.017]	−0.058 [0.016]	−0.058 [0.016]
RI p-value (2-sided, left)	(<0.001, <0.001)	(<0.001, <0.001)	(<0.001, <0.001)
Binary Low (30)	−0.049 [0.016]	−0.039 [0.014]	−0.039 [0.013]
RI p-value (2-sided, left)	(0.005, 0.003)	(0.010, 0.005)	(0.010, 0.005)
Observations	7,500	7,500	7,500
R^2	0.003	0.306	0.306
Omnibus F-test (RI p)	5.34 (<0.001)	5.78 (<0.001)	5.73 (<0.001)

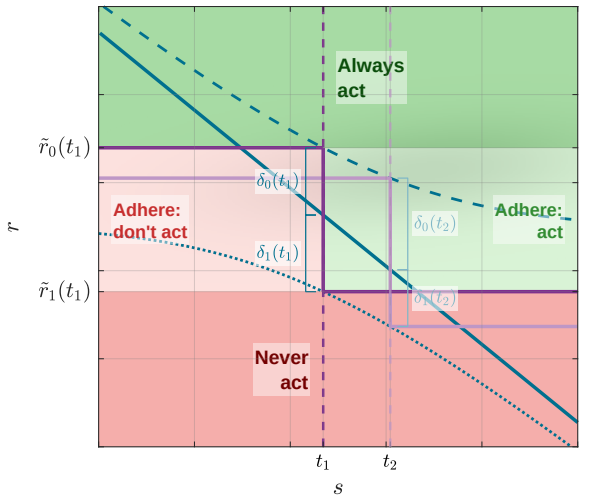
Note: Regression estimates per Equation 5 of loan approval accuracy by treatment condition, with *Binary Medium (50)* as the omitted reference group. Column (1) reports estimates without fixed effects; Column (2) includes individual, case, and block-order fixed effects; Column (3) includes individual and case fixed effects. Asymptotic standard errors clustered at the individual level in brackets. Randomization-inference (RI) p-values (two-sided, one-sided left) are reported in parentheses beneath each coefficient, based on 10,000 design-consistent permutations of treatment–block assignments within participants. One-sided (left) p-values test the theory's directional prediction that *Binary Medium* outperforms each alternative; two-sided values are reported alongside.

Figure I: AI underweighting ($\alpha_s = 0.5$): Decision Boundaries and Binary Coarsening

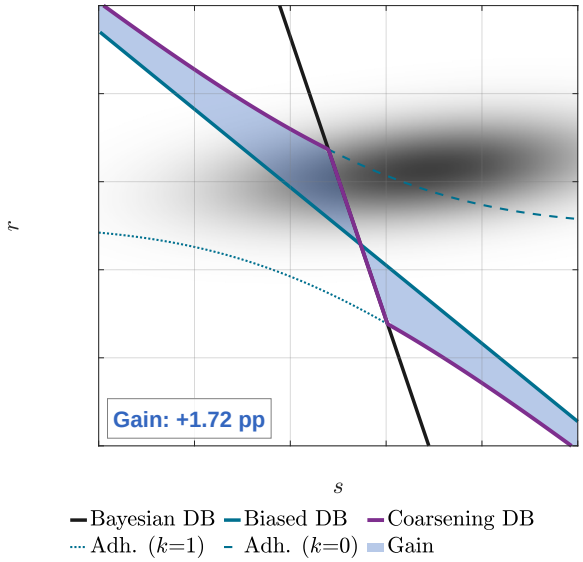
(a) Full revelation: Bayesian vs. biased DBs



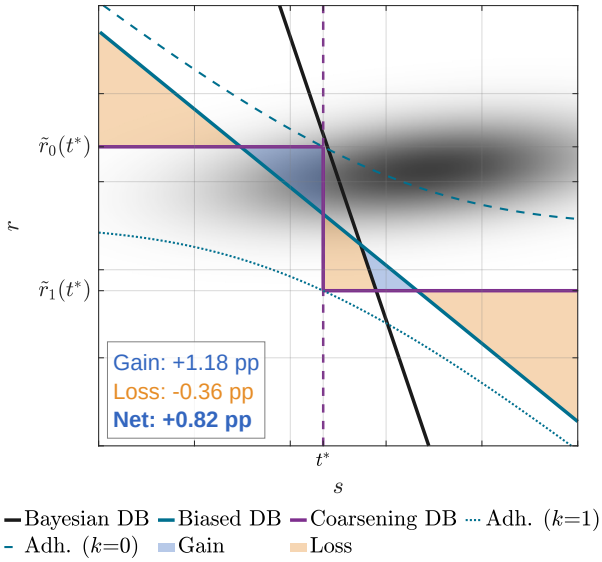
(b) Binary coarsening: Adherence curves + DBs



(c) Known r: gains from coarsening

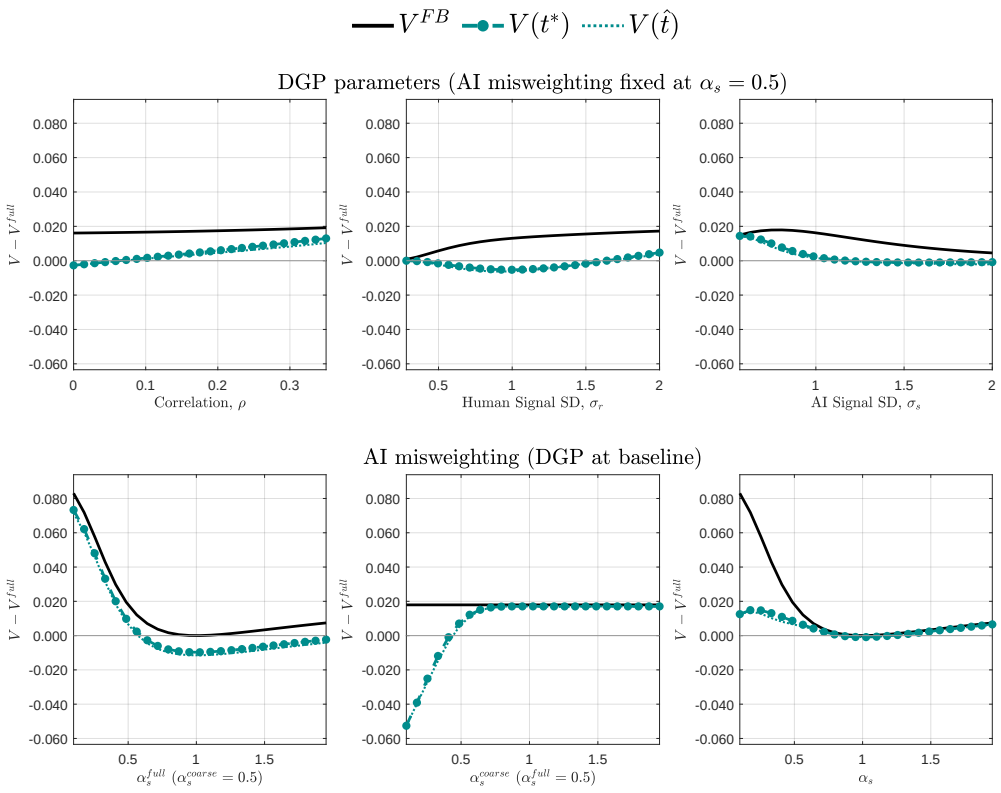


(d) Unknown r: gains/losses from coarsening



Note: Illustration of AI underweighting ($\alpha_s = 0.5$) Receiver's behavior under different disclosure regimes. Shading shows the density of (joint normal) signals. Parameters: $\hat{\sigma}_r = 2.27$, $\hat{\sigma}_s = 0.79$, $\hat{\rho} = 0.25$, prior $p = 2/3$ estimated via Maximum Likelihood from our experimental data (Appendix B.1), and $c_{FP} = c_{TP} = 1$. Panel a: Bayesian DB (black) and biased DB (teal) separating the s-r plane into $a = 1$ (above) and $a = 0$ (below). The dotted blue region marks where the DBs disagree; the potential gain (bottom left) is the accuracy gain from disagreement correction. Panel b: Adherence curves $\tilde{r}_1$ (dotted) and $\tilde{r}_0$ (dashed) in teal. The solid purple line is the DB induced by threshold t_1, with associated choices $a = 1$ (green) and $a = 0$ (red); faded purple shows how the DB shifts as the threshold moves to t_2. Panel c: Accuracy gains from binary coarsening when t can be conditioned on r (blue shaded region), with their expectation at bottom left; solid purple line is the (inverted) r-dependent optimal threshold $t^*(r)$. Panel d: Accuracy gains (blue) and losses (orange) from binary coarsening when r is unknown, with expected net effect at bottom left; solid purple line is the DB induced by the optimal threshold t^*.

Figure II: Accuracy gap relative to biased full revelation across parameters



Note: Each panel plots the expected welfare gap relative to biased full revelation (zero line) for three benchmarks: coarsening at the optimal threshold t^* (dashed teal with circles), coarsening at the AI-only threshold $\hat{t}$ (dotted teal), and Bayesian first-best (solid black). Each panel varies one parameter, holding all others at the baseline of Figure I. **Top row:** each panel varies one DGP parameter, setting $\alpha_s = 0.5$ under both disclosure regimes. **Bottom row** (DGP at baseline): left panel varies misweighting of the fully revealed signal, α_s^{full}, holding $\alpha_s^{\text{coarse}} = 0.5$; middle panel varies α_s^{coarse} holding $\alpha_s^{\text{full}} = 0.5$; right panel varies a common misweighting parameter α_s. All panels set $c_{FP} = c_{TP} = 1$, so welfare is measured in expected accuracy. Parameter ranges restricted so that $\chi \geq 0$ and $\tilde{\gamma}_1 \leq 0$.

Figure III: Main experiment survey flow

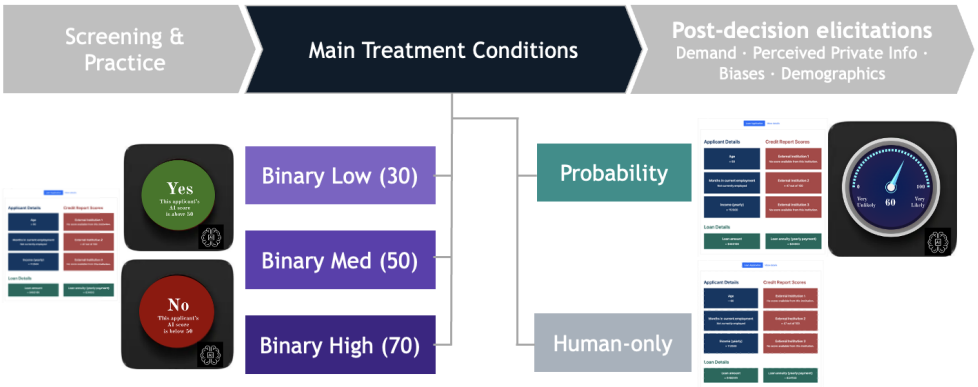
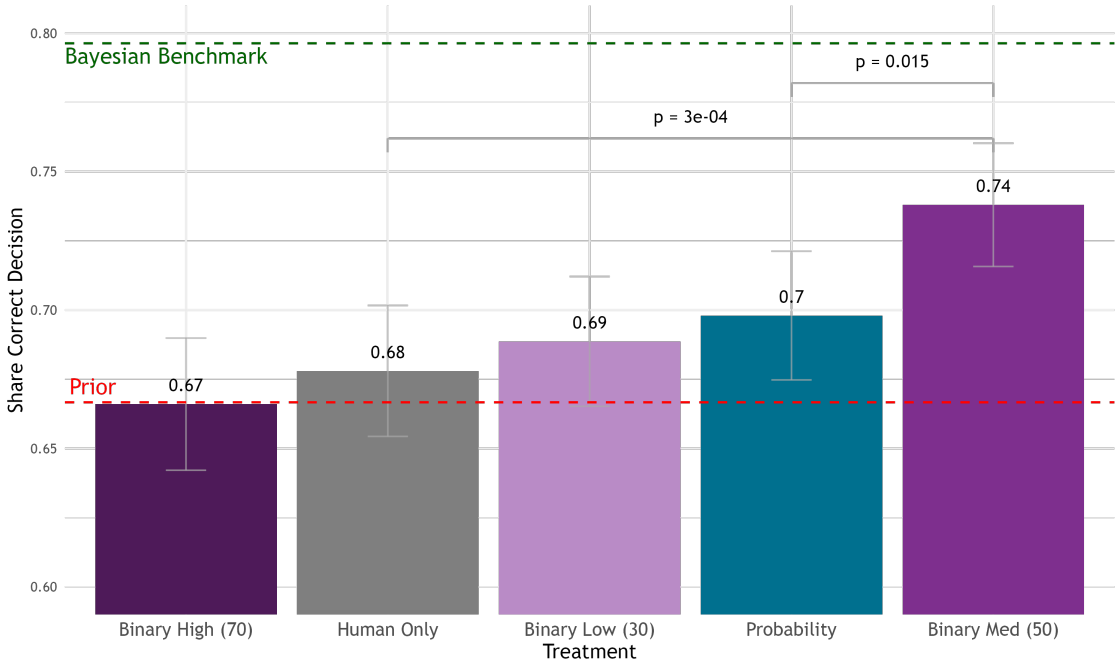
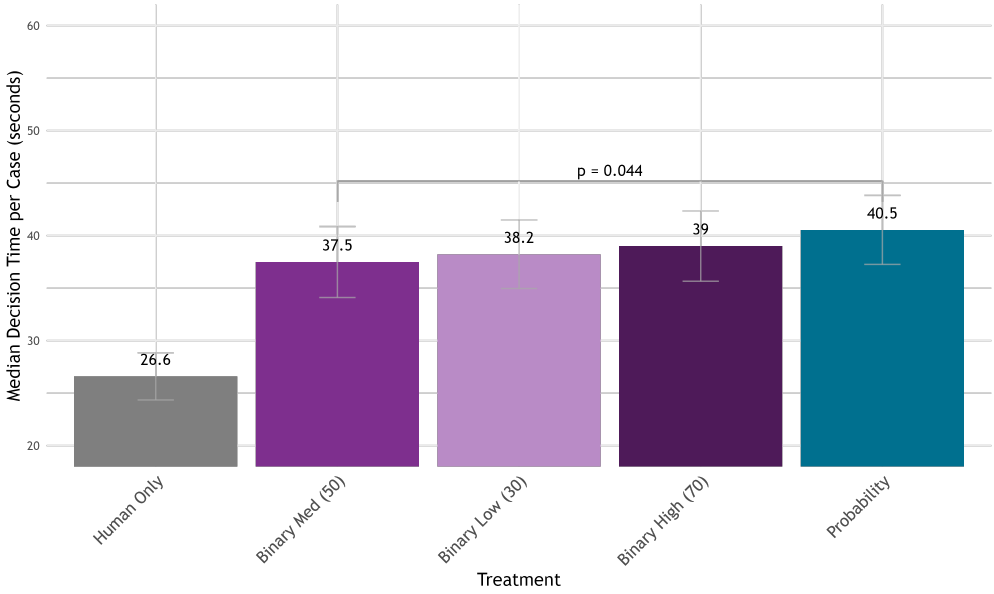


Figure IV: Decision-making outcomes under different treatment conditions



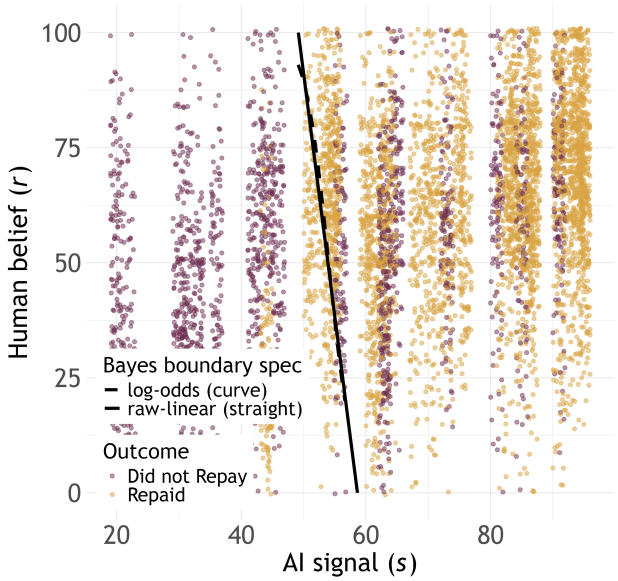
Note: Bars represent the share of correct decisions under each treatment condition. Dashed red line indicates the prior, i.e., the baseline repayment rate; dashed green line marks the Bayesian benchmark, estimated from the empirical joint distribution. p-values are from pairwise t-tests comparing *Binary Medium* to *Probability* and *Human* only; these are descriptive, and our preferred inference is the randomization-inference p-values of Table II. Error bars represent 95% confidence intervals. Specialists: $N = 150$; Total decisions: $N = 7,500$ (1,500 per block).

Figure V: Median per-case decision time across treatment conditions

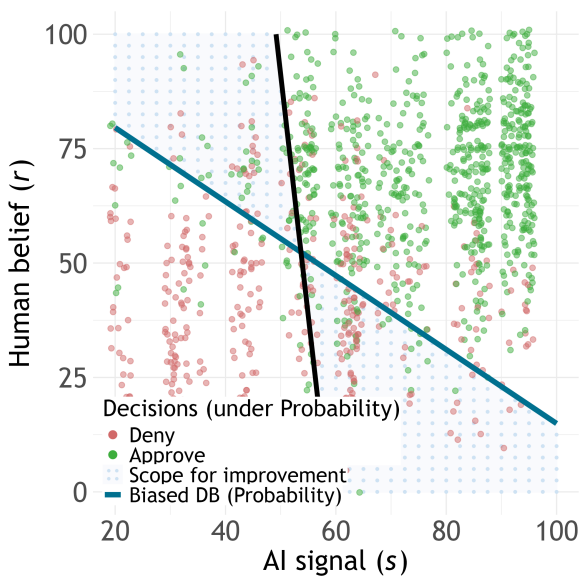


Note: Per-case decision time is measured as the elapsed time between the onset of consecutive cases within a block (nine intervals per ten-case block); the within-subject median is used to summarize each specialist ($N = 150$), and bars report the across-specialist mean of these medians. Error bars indicate 95% confidence intervals across specialists. The p-value is from a within-subject (paired) t-test comparing each specialist's median decision time in *Binary Medium* and *Probability*.

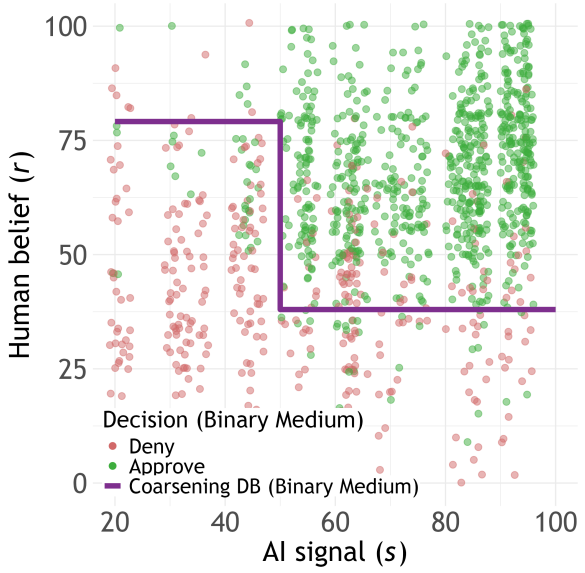
Figure VI: Decision Boundaries, Estimated from Experiment Data (Pooled)



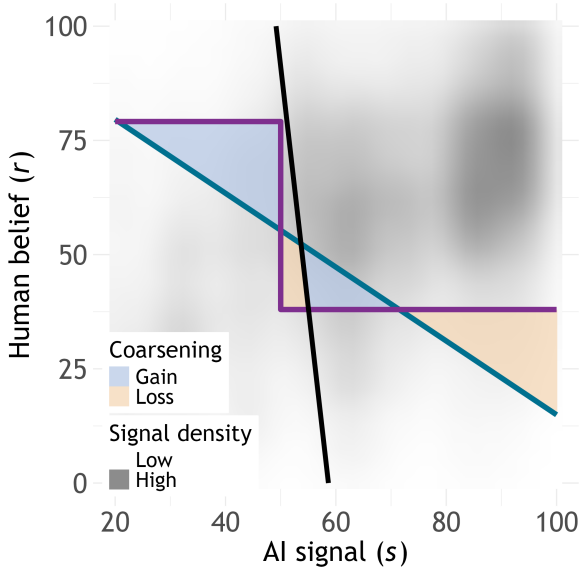
(a) Bayesian Decision Boundary
(All cases with elicited r, pooled)



(b) Biased DB and Scope for Improvement
(*Probability* Treatment, Pooled)



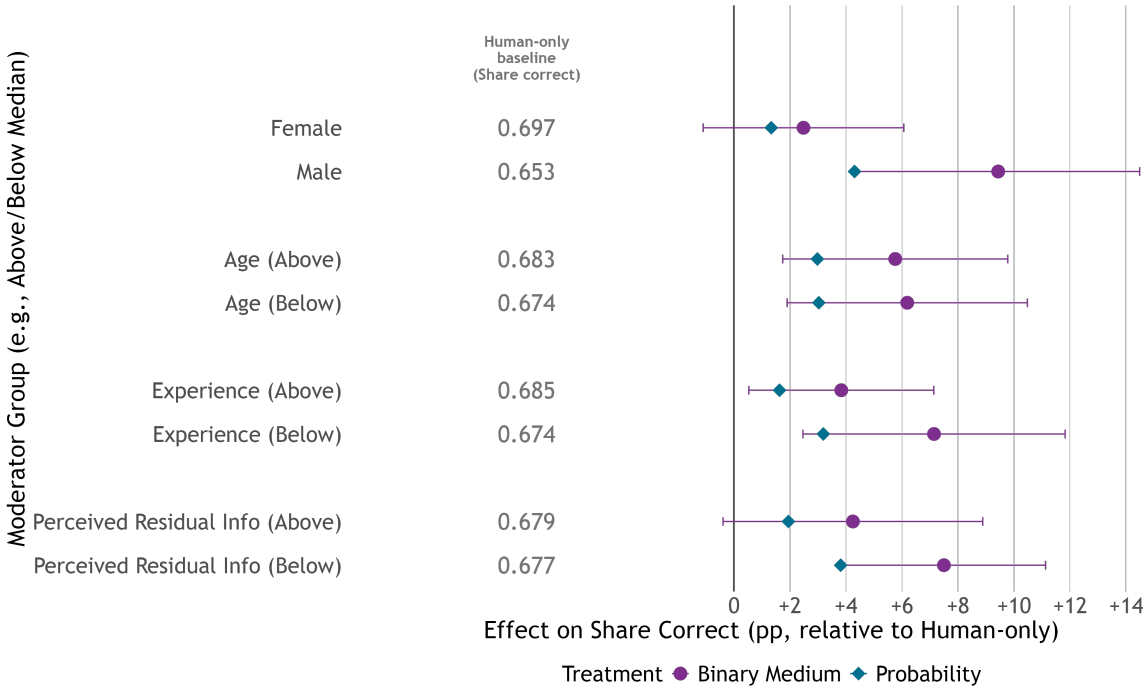
(c) Coarsening DB
(*Binary Medium* , Pooled)



(d) Coarsening Gains and Losses
(All cases with elicited r, pooled)

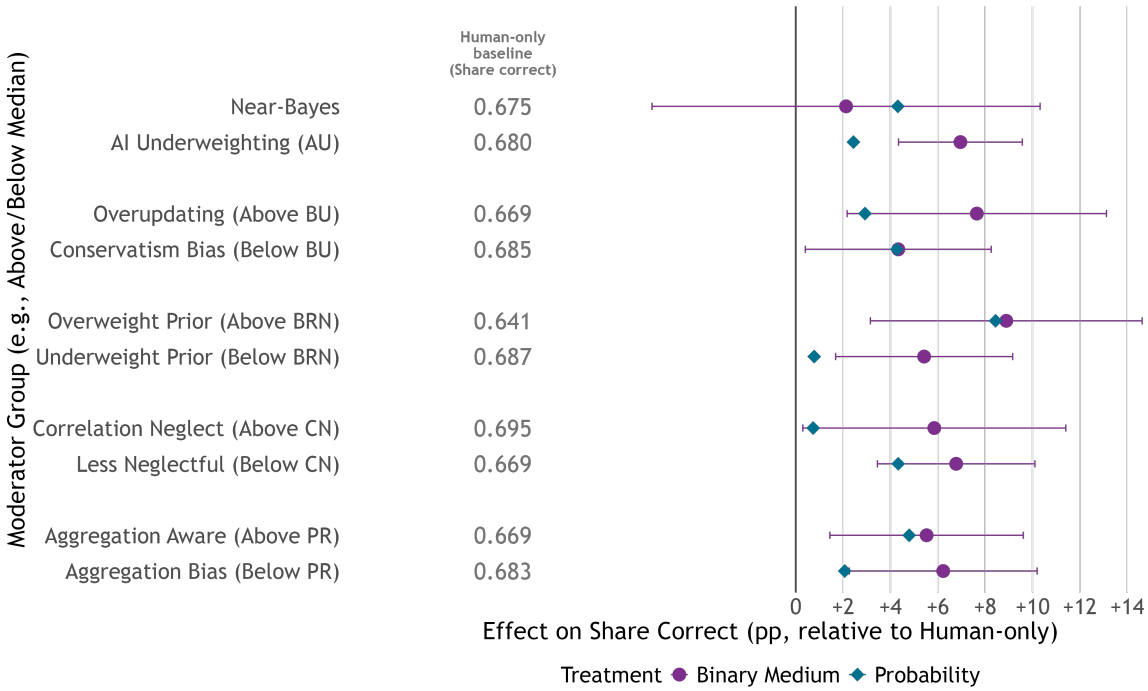
Note: All panels plot the AI signal s (Sender's type) against the loan specialist's private signal r (Receiver's type) expressed as 0–100 repayment probability. Panel (a) shows the Bayesian decision boundary, estimated by a logistic regression of the repayment outcome on s and r (dashed: log-odds; solid: raw-linear). Points are all individual cases with elicited r, jittered, colored by realized outcome (repaid / did not repay). Panels (b) and (c) plot the posterior median of the HLM-estimated distribution of loan-specialist decision boundaries, with points showing all specialist decisions under the relevant treatment. Panel (b) plots *Probability* (teal); the shaded dotted region marks the scope for improvement relative to the Bayesian boundary. Panel (c) plots *Binary Medium* (purple step). Appendix Figure C.15 has the analogous plots for *Binary Low* and *Binary High*. Panel (d) overlays all three boundaries and shades, over the grayscale signal density (all cases with elicited r), the regions where coarsening yields accuracy gains (blue) or losses (orange) relative to full-revelation.

Figure VII: Heterogeneous Treatment Effects by Demographics



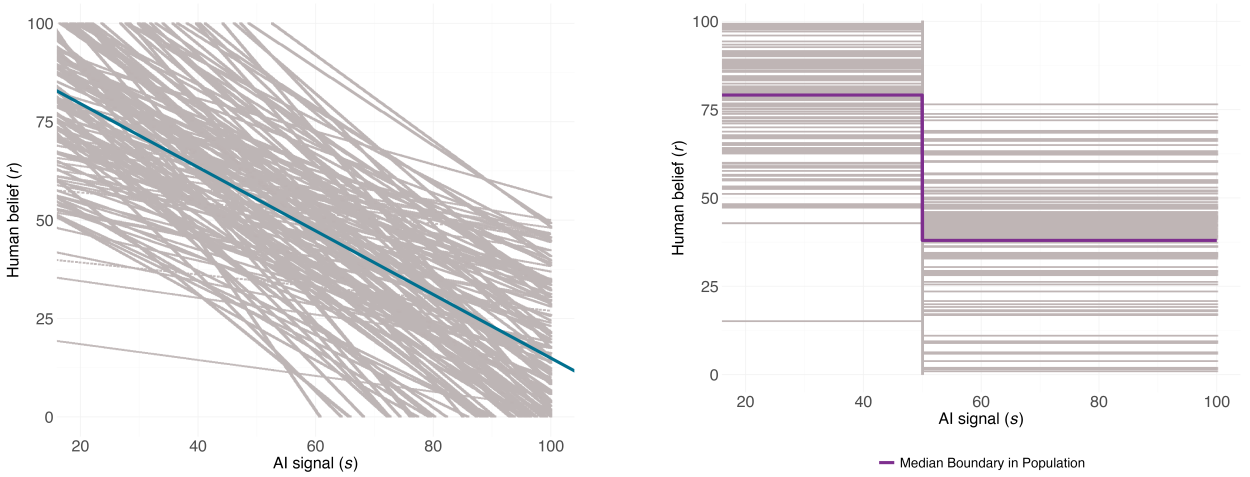
Note: Each point is the treatment effect on the share of correct decisions—in percentage points, relative to that subgroup's own *Human-only* baseline (the vertical line at zero)—for *Binary Medium* (purple circles) and *Probability* (teal diamonds). Effects are treatment coefficients (relative to *Human-only*) from the same fixed-effects specification as the subgroup regressions in Appendix C.2.2 — case, individual, and block fixed effects, with standard errors clustered at the individual level. The right-hand column reports the *Human-only* baseline accuracy (share correct) for each subgroup. Non-gender subgroups are defined by median splits, where 'Above'/'Below' refer to participants above/below the median for each variable. Error bars represent 95% confidence intervals clustered at the individual level, displayed only for *Binary Medium* for simplicity. Formal tests of differences in treatment effects across subgroups are reported in Appendix C.2.2.

Figure VIII: Heterogeneous Treatment Effects by Bias Elicitations



Note: Each point is the treatment effect on the share of correct decisions—in percentage points, relative to that subgroup's own *Human-only* baseline (the vertical line at zero)—for *Binary Medium* (purple circles) and *Probability* (teal diamonds). Effects are treatment coefficients (relative to *Human-only*) from the same fixed-effects specification as the subgroup regressions in Appendix C.2.2 — case, individual, and block fixed effects, with standard errors clustered at the individual level. The right-hand column reports the *Human-only* baseline accuracy (share correct) for each subgroup. Subgroups are defined by median splits, where 'Above'/'Below' indicate participants above/below the median. Error bars represent 95% confidence intervals clustered at the individual level, displayed only for *Binary Medium* for simplicity. Formal tests of differences in treatment effects across subgroups are reported in Appendix C.2.2.

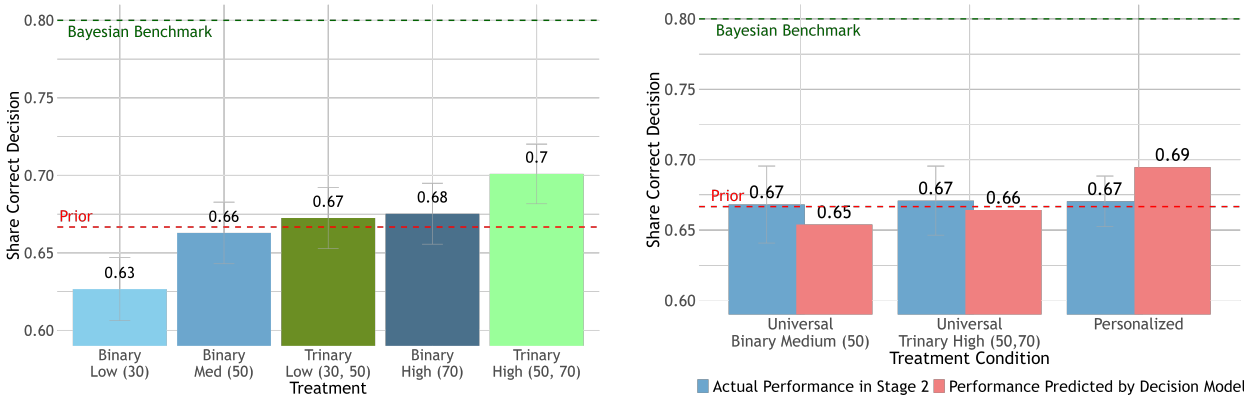
Figure IX: Heterogeneity in decision boundaries: *Probability* and *Binary Medium*



(a) Probability (b) Binary Medium

Note: Estimated decision boundary for each individual specialist under *Probability* and *Binary Medium*, based on posterior samples from the HLM. Gray lines represent individual decision boundaries (median across posterior samples). The colored lines shows the median boundary across specialists—for *Probability* (teal), this uses the median slope and intercept; for *Binary Medium* (purple), the median of each relevant point of the adherence curve (e.g., $\tilde{r}_0(t), \tilde{r}_1(t)$).

Figure X: Stage 1 and Stage 2 results, personalization experiment

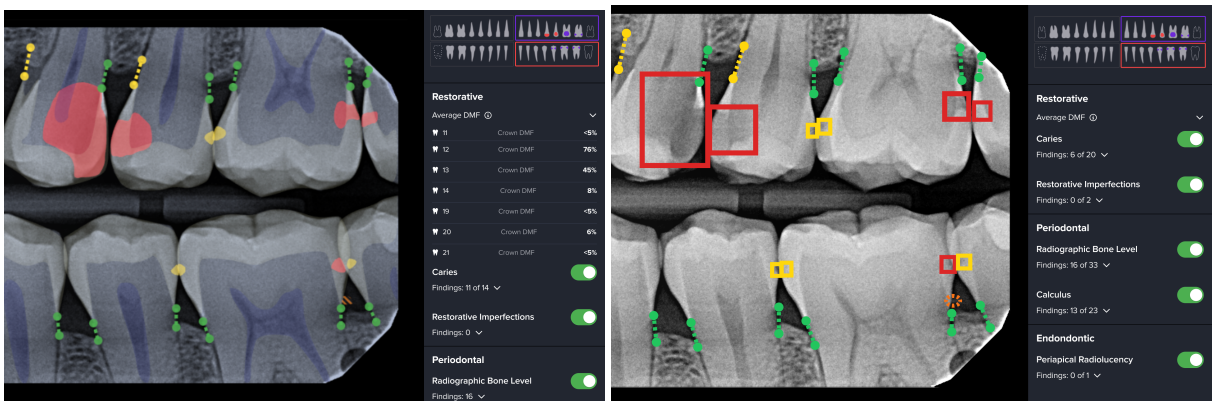


(a) Stage 1 (b) Stage 2

Note: Panel (a): Bars represent the share of correct decisions under each treatment condition. Dashed red line indicates prior; dashed green line marks Bayesian benchmark. Error bars represent 95% CIs. Panel (b): Blue bars represent the empirical share of correct decisions, and red bars indicate the performance *predicted* by the decision model in Section 5.2.2, estimated on Stage 1 data.

A Real-world Examples of Coarsened AI signals

Figure A.1: VideaAI Dental Interface (left: probability, right: binary)



B Theory: Proofs and Additional Results

B.1 Calibration of the signal environment

Sample. We calibrate the signal environment of Section 2 to the Experiment 1 data. The estimation sample pools all decision rounds with a non-missing belief elicitation and AI signal ($N = 6{,}379$ participant-rounds from 150 specialists); the state $\omega \in \{0, 1\}$ is the realized repayment label, with $\hat{p} = \Pr(\omega = 1)$ equal to the design prior $2/3$.

Transform and likelihood. Elicited beliefs and AI signals are reported on a $[0, 100]$ probability scale, while the model's signals are unbounded bivariate normals. We map (r, s) to log-odds space: $x \mapsto \text{logit}\big(\min\{\max\{x/100,\, \varepsilon\},\, 1-\varepsilon\}\big)$ with $\varepsilon = 0.005$. Within each state we assume the transformed pair is bivariate normal with state-specific means $(\mu_{r\omega}, \mu_{s\omega})$ and a common covariance matrix Σ, as in the model. The MLE is available in closed form: state-specific sample means and the pooled within-state scatter matrix.

Normalization. The model normalizes conditional means to $\mu_0 = 0$ and $\mu_1 = 1$. We apply the affine map $x \mapsto (x - \hat{\mu}_{x,0})/(\hat{\mu}_{x,1} - \hat{\mu}_{x,0})$ separately to each signal. Correlations are invariant to affine maps and standard deviations scale by the same factor, giving the calibrated parameters used

throughout the paper:

$$\hat{\sigma}_r = 2.27 \ \ (\text{SE } 0.14), \qquad \hat{\sigma}_s = 0.79 \ \ (\text{SE } 0.007), \qquad \hat{\rho} = 0.25 \ \ (\text{SE } 0.015),$$

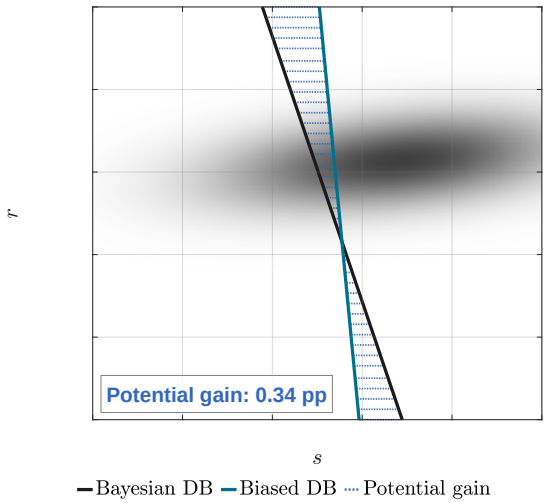
with standard errors and confidence intervals from a participant-clustered nonparametric bootstrap ($B = 2{,}000$). These values parameterize all theory-based figures and quantitative statements (together with $p = 2/3$ and $\mu_0 = 0$, $\mu_1 = 1$), including the implied residual-information coefficient $\hat{\chi} = 0.10$ used in Section 2.

Robustness. As an alternative to the log-odds transform, we recalibrate r into posterior space using the same kernel-density (KDE) calibration used to construct the AI signal s, and renormalize as above. This yields $\hat{\sigma}_r = 1.90$, $\hat{\sigma}_s = 0.74$, $\hat{\rho} = 0.29$; all qualitative conclusions are unchanged. (Figure notes in earlier drafts quoted this alternative triple; all figures are generated with the log-odds calibration above.) The two transforms differ mainly in how they treat the mass of extreme beliefs (r near 0 or 100), to which σ_r is most sensitive.

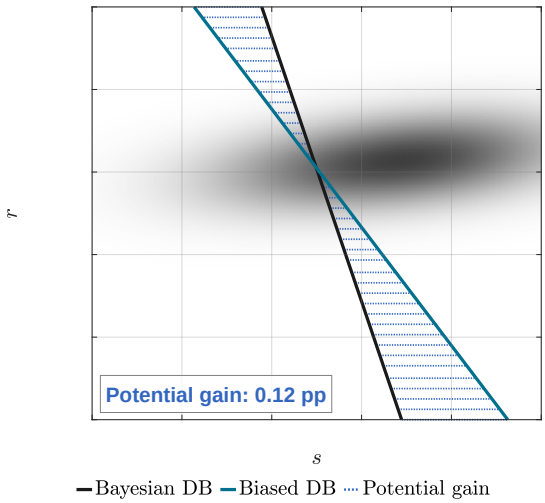
B.2 Illustrations under other biases

See Figures B.1 to B.4 for illustrations under other biases.

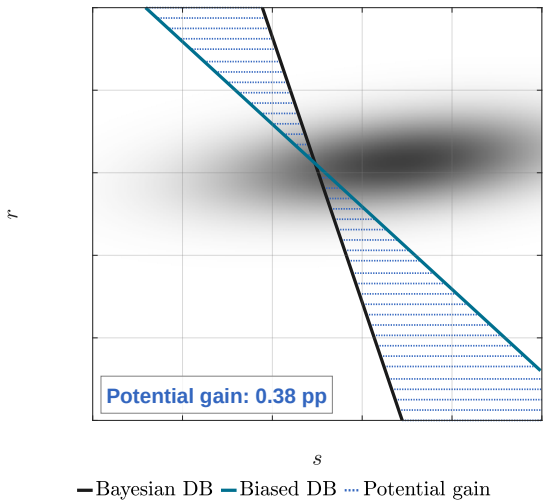
Figure B.1: Full revelation: DBs under different biases



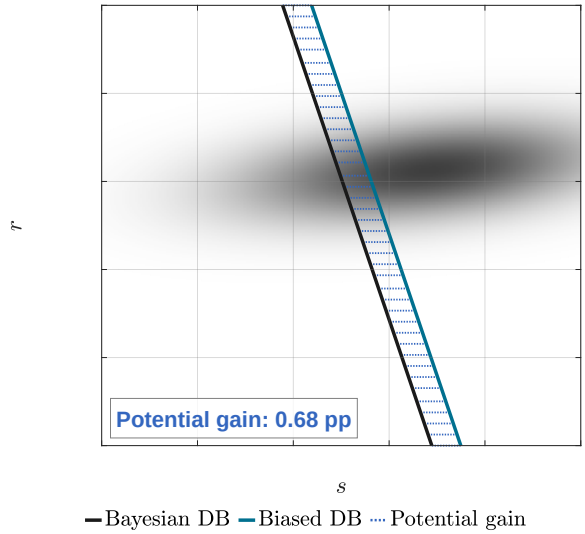
(a) AI overweighting



(b) Overconfidence



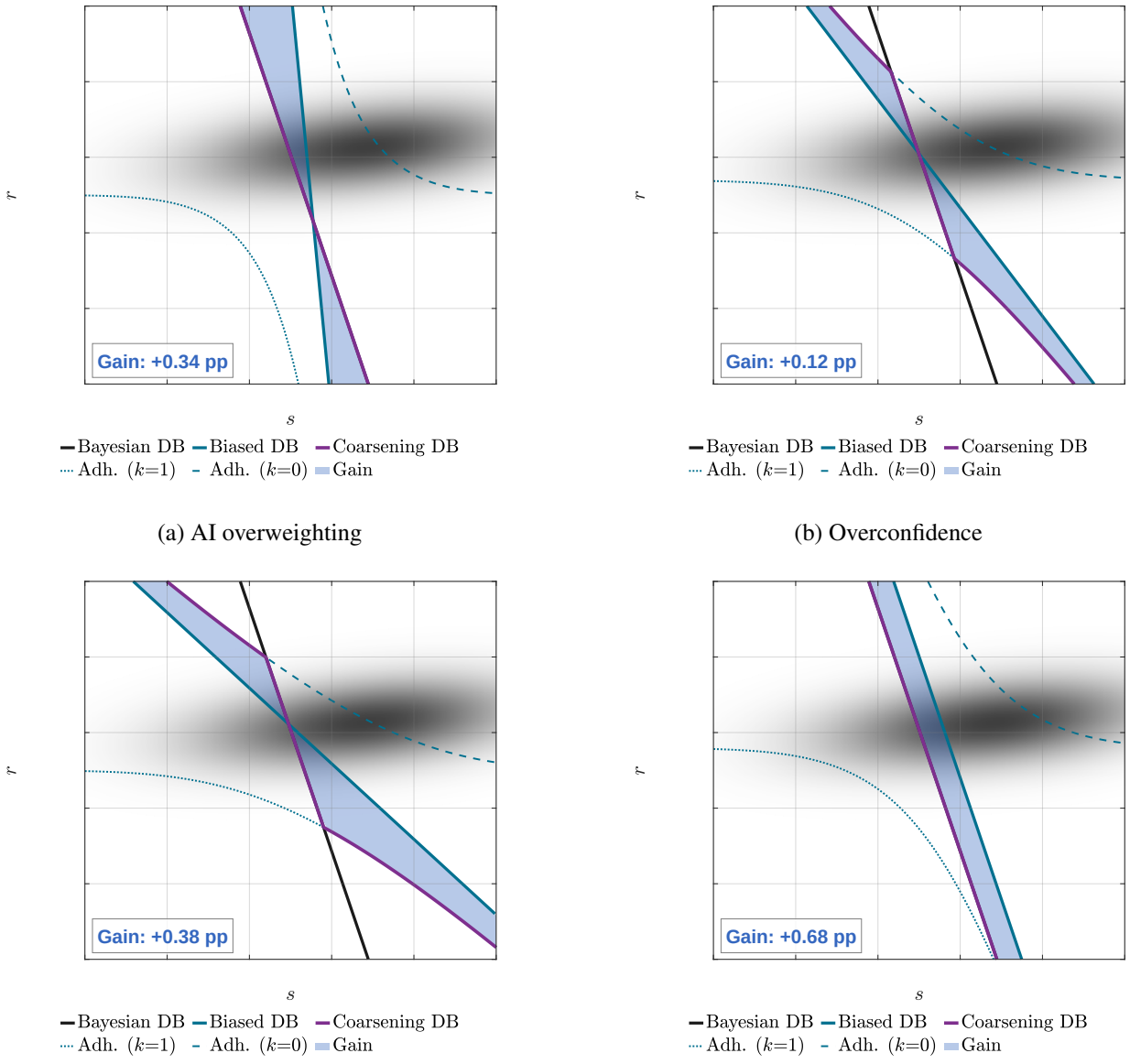
(c) Correlation Neglect



(d) Base-rate neglect

Note: Illustration of DBs under different biases. Shading shows the density of joint-normal signals. Parameters are estimated by maximum likelihood from our experimental data (Appendix B.1): $\hat{\sigma}_r = 2.27$, $\hat{\sigma}_s = 0.79$, $\hat{\rho} = 0.25$, and prior $p = 2/3$. The Bayesian DB (black) and biased DB (teal) separate the s-r plane into $a = 1$ (above) and $a = 0$ (below). The dotted blue region marks where the DBs disagree; potential gain (bottom left) is the accuracy gain from correcting disagreement-region decisions. Biases are as follows: Panel a: AI overweighting, $\alpha_s = 1.5$. Panel b: overconfidence, perceiving the variance of r as σ_r^2/α_r, with $\alpha_r = 1.5$. Panel c: correlation neglect, ignoring the correlation between s and r by perceiving $\tilde{\rho} = \theta\rho$ with $\theta = 0$. Panel d: base-rate neglect, where the prior odds enter the posterior map as $(p/(1-p))^{\kappa}$, with $\kappa = 0.5$.

Figure B.2: Known r: gains from coarsening



(a) AI overweighting

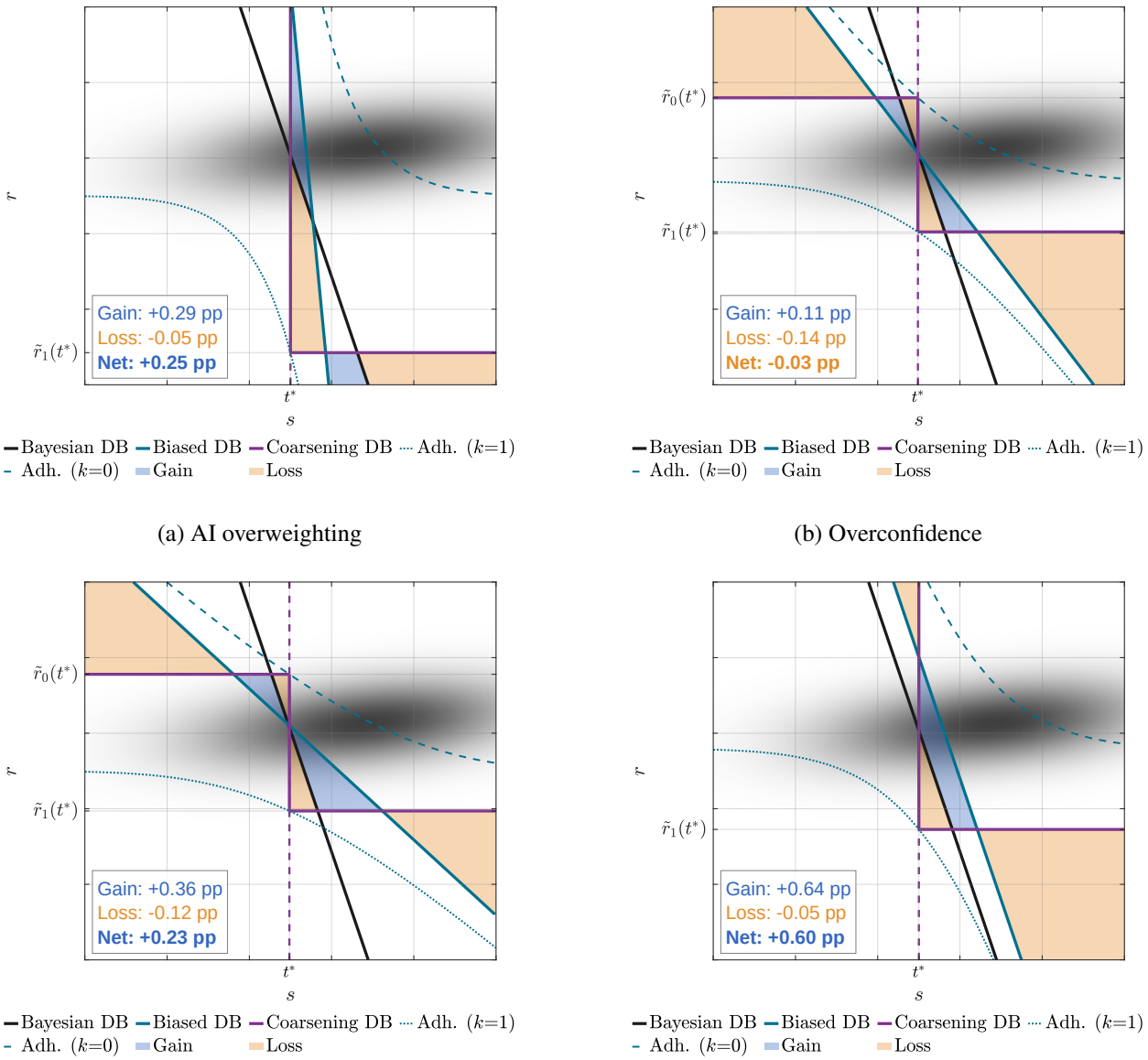
(b) Overconfidence

(c) Correlation Neglect

(d) Base-rate neglect

Note: Illustration of binary coarsening with known r under different biases. Shading shows the density of joint-normal signals. Parameters are estimated by maximum likelihood from our experimental data (Appendix B.1): $\hat{\sigma}_r = 2.27$, $\hat{\sigma}_s = 0.79$, $\hat{\rho} = 0.25$, and prior $p = 2/3$. Accuracy gains from binary coarsening when t can be conditioned on r (blue shaded region), with their expectation at bottom left; solid purple line is the (inverted) r-dependent optimal threshold $t^*(r)$. Biases are as follows: Panel a: AI overweighting, $\alpha_s = 1.5$. Panel b: overconfidence, perceiving the variance of r as σ_r^2/α_r, with $\alpha_r = 1.5$. Panel c: correlation neglect, ignoring the correlation between s and r by perceiving $\tilde{\rho} = \theta\rho$ with $\theta = 0$. Panel d: base-rate neglect, where the prior odds enter the posterior map as $(p/(1-p))^\kappa$, with $\kappa = 0.5$.

Figure B.3: Unknown r: gains and losses from coarsening



(a) AI overweighting

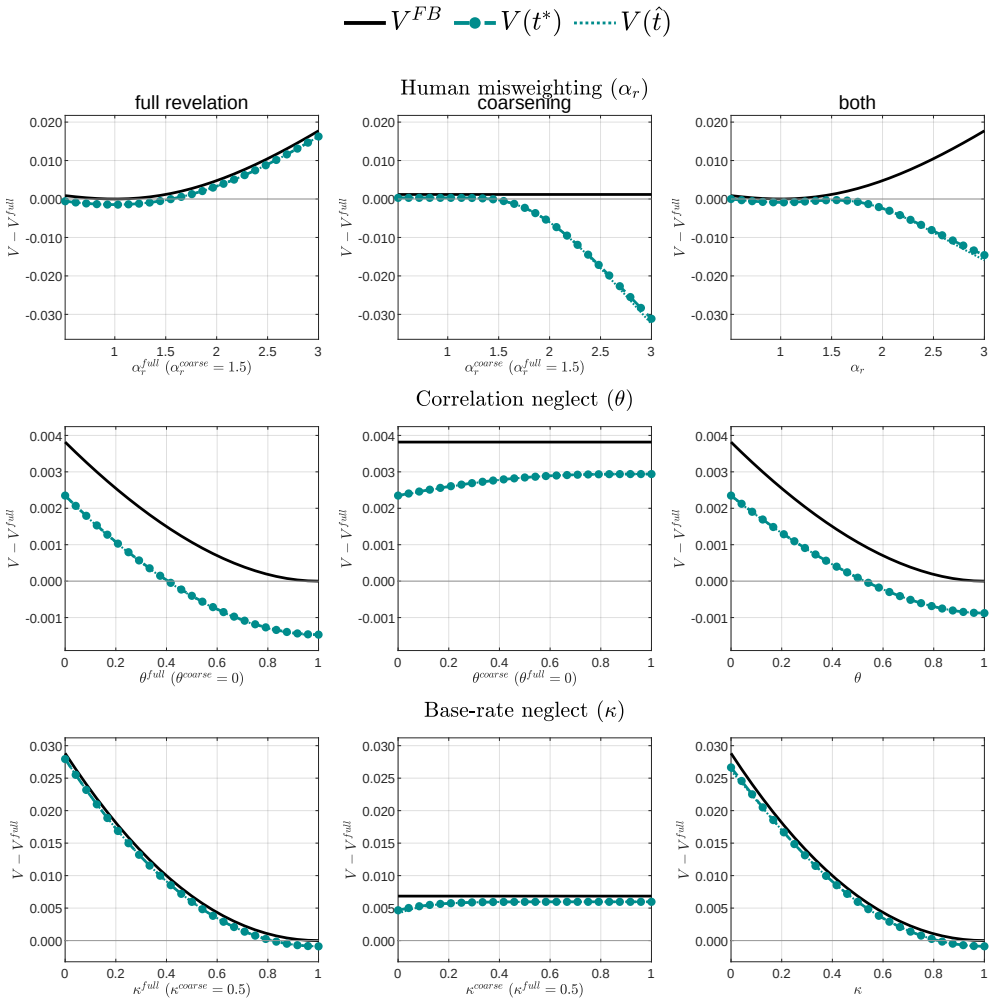
(b) Overconfidence

(c) Correlation Neglect

(d) Base-rate neglect

Note: Illustration of binary coarsening with unknown r under different biases. Shading shows the density of joint-normal signals. Parameters are estimated by maximum likelihood from our experimental data (Appendix B.1): $\hat{\sigma}_r = 2.27$, $\hat{\sigma}_s = 0.79$, $\hat{\rho} = 0.25$, and prior $p = 2/3$. Accuracy gains (blue) and losses (orange) from binary coarsening when r is unknown, with expected net effect at bottom left; solid purple line is the DB induced by the optimal threshold t^*. Biases are as follows: Panel a: AI overweighting, $\alpha_s = 1.5$. Panel b: overconfidence, perceiving the variance of r as σ_r^2/α_r, with $\alpha_r = 1.5$. Panel c: correlation neglect, ignoring the correlation between s and r by perceiving $\tilde{\rho} = \theta\rho$ with $\theta = 0$. Panel d: base-rate neglect, where the prior odds enter the posterior map as $(p/(1-p))^\kappa$, with $\kappa = 0.5$.

Figure B.4: Accuracy gap relative to biased full revelation across parameters: Other biases



Note: Each panel plots the expected welfare gap relative to biased full revelation (zero line) for three benchmarks: coarsening at the optimal threshold t^* (dashed teal with circles), coarsening at the AI-only threshold $\hat{t}$ (dotted teal), and the Bayesian first-best (solid black). See note to Figure II for further detail. Each row replicates Figure II's bottom row for a different bias: the left panel varies the bias under full revelation, holding the bias under binary coarsening fixed, the middle panel does the opposite, and the right panel varies a common bias under both regimes. Biases are as follows. **Top row:** Human misweighting, perceiving the variance of r as σ_r^2/α_r. **Middle row:** correlation neglect, ignoring the correlation between s and r by perceiving $\tilde{\rho} = \theta\rho$. **Bottom row:** base-rate neglect, where the prior odds enter the posterior map as $(p/(1-p))^{\kappa}$.

B.3 Known-r disclosure

We show that when disclosure can be conditioned on the human signal r, binary coarsening is optimal among deterministic monotone disclosure rules, and we give conditions under which the Sender fully implements the Bayesian decision boundary.

Disclosure and decision rules. A *deterministic monotone disclosure rule* Φ maps the AI signal s into messages whose preimages are ordered (possibly degenerate) intervals. We identify each message with the endpoint pair (ℓ, u), $-\infty \leq \ell \leq u \leq \infty$, of its preimage's interval hull; whether endpoints are included never matters below, since behavior at a single signal realization has zero probability. Full revelation is $\Phi(s) = s$, with endpoints (s, s); binary coarsening at t is $\Phi(s) = \mathbf{1}\{s > t\}$, with a "No" message $(-\infty, t)$ and a "Yes" message (t, ∞); full censorship sends the single message $(-\infty, \infty)$ regardless of s. We say a message (ℓ, u) has *weakly lower endpoints* than (ℓ', u') if $\ell \leq \ell'$ and $u \leq u'$: shifting a message's endpoints rightward means it indicates (weakly) higher AI signals.

A *decision rule* $\tilde{a}(r, \cdot) \in \{0, 1\}$ specifies the action a type-r Receiver takes after each message. This is the primitive of this appendix: the DB $\tilde{r}$ and the adherence curves $\tilde{r}_k$ of the main text are its restrictions to full revelation and to binary messages, respectively. Consistent with the main text, define

$$\tilde{r}_1(t) = \inf\{r : \tilde{a}(r, (t, \infty)) = 1\}, \qquad \tilde{r}_0(t) = \sup\{r : \tilde{a}(r, (-\infty, t)) = 0\},$$

the lowest type that acts after a "Yes" and the highest type that does not act after a "No."

Assumption 1. *The decision rule satisfies the following regularity conditions:*

1. Message isotonicity. *Messages indicating higher signals push toward action: if (ℓ, u) has weakly lower endpoints than (ℓ', u'), then $\tilde{a}(r, (\ell, u)) \leq \tilde{a}(r, (\ell', u'))$ for every r.*

2. Full-revelation consistency. *Under full revelation the Receiver acts according to the biased DB (3), with $\tilde{\gamma}_1 < 0$: $\tilde{a}(r, (s, s)) = 1$ iff $r > \tilde{r}(s)$, so the inverse DB $\tilde{s}(r) = (r - \tilde{\gamma}_0)/\tilde{\gamma}_1$ is well defined and finite.*

3. Type monotonicity. *Higher types are weakly more inclined to act: $\tilde{a}(r, m)$ is weakly increasing in r for every message m.*

4. Regular adherence. *The adherence curves $\tilde{r}_1(t)$ and $\tilde{r}_0(t)$ are continuous, weakly decreasing in t, and satisfy the sandwich condition $\tilde{r}_1(t) < \tilde{r}(t) < \tilde{r}_0(t)$ for all t.*

Recall the welfare contribution of acting at (r, s), $Q(r, s)$, defined in (4). Since the Bayesian posterior is strictly increasing in s, $Q(r, \cdot)$ crosses zero exactly once, from below; let $s^*(r)$ denote this crossing point, the Bayesian cutoff in s-space. When $\chi > 0$ it coincides with the inverse of the Bayesian DB, $s^*(r) = (r - \gamma_0^*)/\gamma_1^*$; when the Bayesian DB is vertical ($\chi = 0$), it is constant in r.

For fixed r, let

$$\mathcal{T}(r) = \{t : \tilde{r}_1(t) \leq r \leq \tilde{r}_0(t)\}$$

be the set of thresholds at which type r *adheres*: acts after "Yes" and not after "No."[68] By Assumption 1.4, $\mathcal{T}(r)$ is a closed (possibly unbounded) interval, and since $r = \tilde{r}(\tilde{s}(r))$, the sandwich condition gives $\tilde{r}_1(\tilde{s}(r)) < r < \tilde{r}_0(\tilde{s}(r))$: the band contains a neighborhood of the full-revelation cutoff $\tilde{s}(r)$. Write $\underline{t}(r) = \inf \mathcal{T}(r)$ and $\bar{t}(r) = \sup \mathcal{T}(r)$.

Proposition 2. *Suppose Assumption 1 holds and fix r with $s^*(r) \neq \tilde{s}(r)$. Binary coarsening at the threshold*

$$t^*(r) = \arg\min_{t \in \mathcal{T}(r)} |t - s^*(r)|$$

(i) strictly increases the Sender's r-conditional expected utility relative to full revelation;

(ii) is optimal among deterministic monotone disclosure rules;

(iii) fully implements the Bayesian DB, $t^*(r) = s^*(r)$, whenever the bias and the default action point in opposite directions:

$$s^*(r) > \tilde{s}(r) \text{ and } \tilde{a}(r, (-\infty, \infty)) = 0, \qquad \text{or} \qquad s^*(r) < \tilde{s}(r) \text{ and } \tilde{a}(r, (-\infty, \infty)) = 1.$$

Applying $t^*(\cdot)$ pointwise therefore yields an optimal r-conditional disclosure rule.

Proof. Suppose $\tilde{s}(r) < s^*(r)$; the opposite case is symmetric. Write $\tilde{s} = \tilde{s}(r)$ and $s^* = s^*(r)$.

Step 1: the value of a cutoff. If type r acts exactly on $\{s > \tau\}$, the Sender's r-conditional payoff is, up to the positive marginal density of r,

$$V_r(\tau) = \int_\tau^\infty Q(r, s)\, ds,$$

which is finite since Q carries the signal density. Since $V_r'(\tau) = -Q(r, \tau)$ is positive for $\tau < s^*$ and negative for $\tau > s^*$, V_r is strictly single-peaked with peak s^*: any move of the cutoff toward the Bayesian one strictly raises welfare.

Step 2: which cutoffs coarsening can implement. Binary coarsening at t makes type r act exactly on $\{s > t\}$ if and only if $t \in \mathcal{T}(r)$. Maximizing the single-peaked V_r over the closed

[68] Under type monotonicity, adherence at t is equivalent to $r \in [\tilde{r}_1(t), \tilde{r}_0(t)]$; ties at the boundary types are immaterial throughout.

interval $\mathcal{T}(r)$ yields the projection of the peak onto the band,

$$t^*(r) = \min\{\max\{s^*, \underline{t}(r)\}, \bar{t}(r)\} = \min\{s^*, \bar{t}(r)\} \in (\tilde{s},\, s^*],$$

where the second equality uses $\underline{t}(r) < \tilde{s} < s^*$.

Step 3: strict improvement (part i). Full revelation implements the cutoff $\tilde{s}$. Since $Q(r,\cdot) < 0$ on $(\tilde{s}, t^*(r)) \subset (-\infty, s^*)$,

$$V_r(t^*(r)) - V_r(\tilde{s}) = -\int_{\tilde{s}}^{t^*(r)} Q(r,s)\, ds > 0 :$$

coarsening drags the Receiver's effective cutoff strictly toward the Bayesian one, and every step of the way raises welfare.

Step 4: no monotone rule does better (part ii). Let Φ be any deterministic monotone rule. Along s, its messages have weakly increasing endpoints, so by message isotonicity the induced action is weakly increasing in s: Φ implements some cutoff τ. We claim $\tau \le \bar{t}(r)$—no monotone rule can push the Receiver's cutoff above the adherence band. Suppose instead $\tau > \bar{t}(r)$ and pick any $x \in (\bar{t}(r), \tau)$. Then r would in fact adhere to binary coarsening at x: (a) the message $\Phi(x)$ induces no action (as $x < \tau$) and has weakly higher endpoints than the "No" message $(-\infty, x)$, so "No" at x induces no action either; (b) $x > \tilde{s}$, so full revelation at x induces action, and the "Yes" message (x, ∞) has weakly higher endpoints than (x, x), so it induces action too. Hence $x \in \mathcal{T}(r)$, contradicting $x > \bar{t}(r)$. Every attainable cutoff thus satisfies $\tau \le \bar{t}(r)$, and by single-peakedness $V_r(\tau) \le V_r(\min\{s^*, \bar{t}(r)\}) = V_r(t^*(r))$.

Step 5: full implementation (part iii). Suppose $\tilde{a}(r, (-\infty, \infty)) = 0$: the default action is inaction. The "No" message at any threshold has weakly lower endpoints than $(-\infty, \infty)$, so it never induces action; and, as in Step 4(b), the "Yes" message at any $t > \tilde{s}$ induces action. Hence every $t > \tilde{s}$ lies in $\mathcal{T}(r)$: the band is unbounded above, $\bar{t}(r) = \infty$, and $t^*(r) = \min\{s^*, \infty\} = s^*$. The case $s^* < \tilde{s}$ with default action is symmetric. □

Step 5 shows that when the default is inaction the adherence band is $\mathcal{T}(r) = [\underline{t}(r), \infty)$: a "No" can never backfire, so the only binding constraint on the Sender is whether "Yes" is persuasive enough. Full implementation then holds if and only if $\underline{t}(r) \le s^*(r)$—automatic when the Receiver is "too eager to act" ($\tilde{s}(r) < s^*(r)$, part (iii)), and otherwise a restriction on the strength of the bias, as the following illustration shows.

Illustration: AI misweighting. Fix a type r whose default action—absent any information about s—is inaction, $\tilde{a}(r, (-\infty, \infty)) = 0$ (the default-action case is symmetric), and index the bias by α_s. Given this default, an underweighting Receiver ($\alpha_s < 1$) requires a stronger AI signal to act than a Bayesian, so $\tilde{s}(r; \alpha_s) > s^*(r)$; overweighting ($\alpha_s > 1$) reverses the inequality. Figure B.5 works in the (α_s, t)-plane, with thresholds in AI-signal units on the vertical axis.

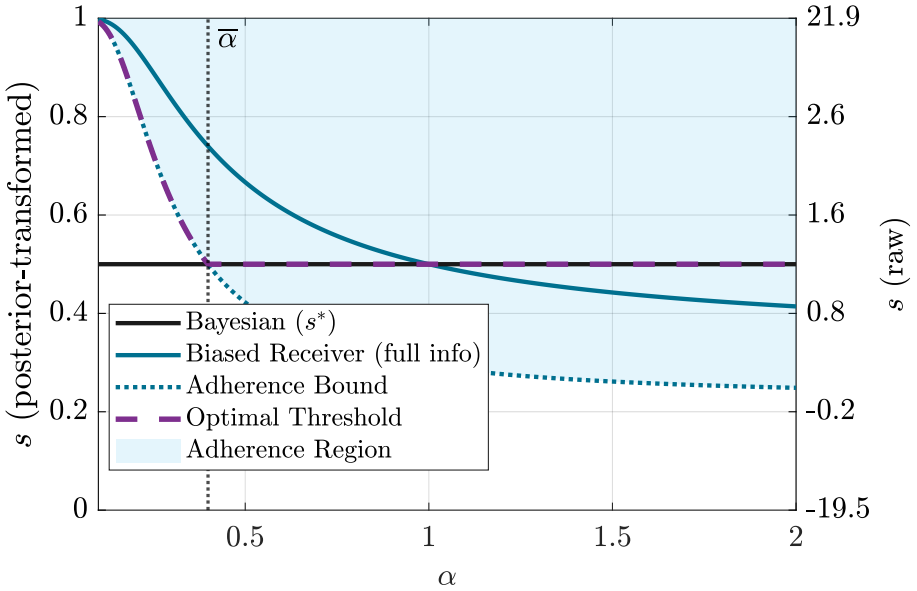
Three curves organize the analysis. The Bayesian cutoff $s^*(r)$ (solid black) is flat: it does not depend on the bias. The full-revelation cutoff $\tilde{s}(r; \alpha_s)$ (solid teal) is decreasing in α_s—the more weight the Receiver places on the AI signal, the weaker the signal needed to induce action—and crosses the Bayesian cutoff exactly at $\alpha_s = 1$. The *adherence bound* (dotted teal) is the lowest threshold whose "Yes" induces type r to act; it lies below the teal curve by the sandwich condition and is likewise decreasing. Because the default is inaction, a "No" never induces action, so the adherence set is everything above the frontier (shaded).

The optimal threshold is the projection of $s^*(r)$ onto this set,

$$t^*(r; \alpha_s) = \max\{s^*(r),\, \underline{t}(r; \alpha_s)\},$$

plotted as the dashed purple line: the upper envelope of the black cutoff and the blue frontier. Two regimes emerge, separated by $\overline{\alpha} < 1$, where the frontier and $s^*(r)$ (vertical dotted line) cross. For $\alpha_s \geq \overline{\alpha}$, the Receiver adheres to the Bayesian cutoff and the Sender fully implements the first-best; for $\alpha_s > 1$ this is part (iii) of Proposition 2, and it extends to mild underweighting by continuity of the frontier. For $\alpha_s < \overline{\alpha}$, a "Yes" at $s^*(r)$ is not persuasive enough, so the Sender raises the threshold onto the frontier; by part (i), this still strictly improves on full revelation for every $\alpha_s \neq 1$.

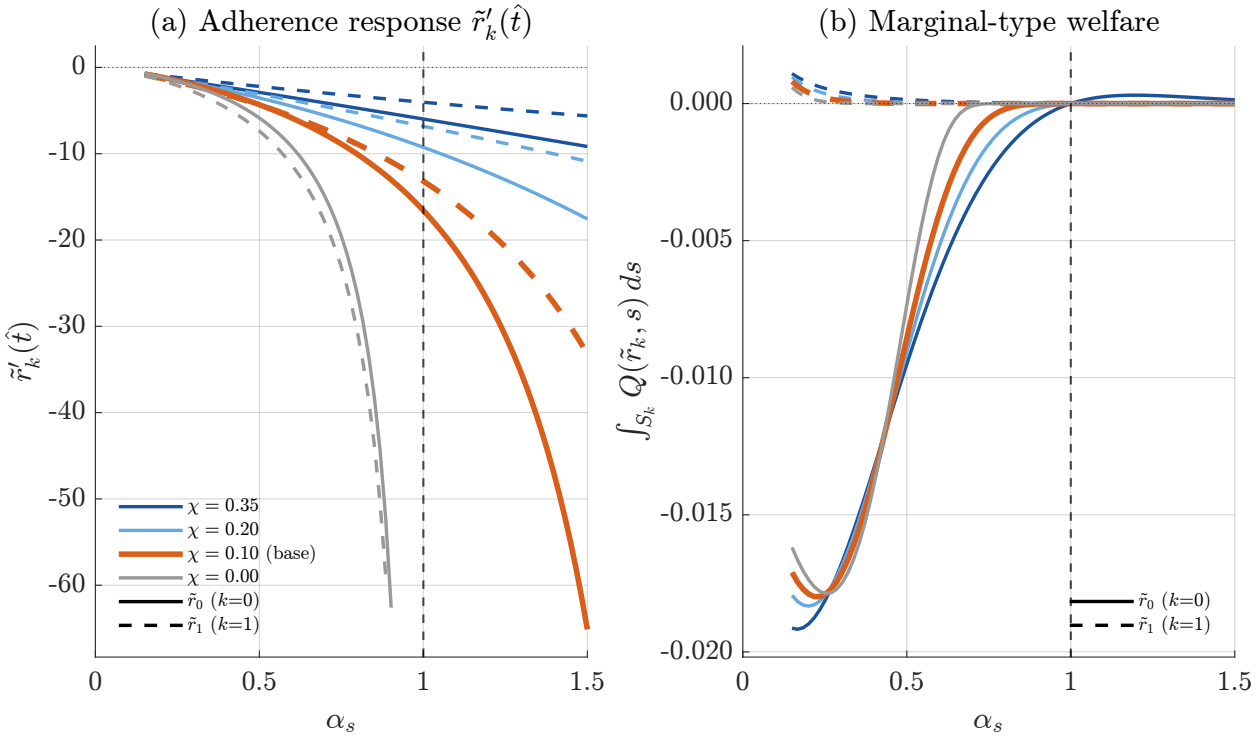
Figure B.5: AI misweighting when r is known: graphical illustration



Note: Boundaries, adherence, and the optimal threshold with known r, under different values of AI misweighting α_s. For transparency we use a stripped-down parametrization: $p = \frac{1}{3}$, $s \mid \omega \sim N(\omega, 1)$, and r held fixed (and hence suppressed); the qualitative picture is identical under the calibrated parameters of Figure I. The y-axis is labeled both in raw AI-signal units (right) and in Bayesian posterior units (left). Solid black: Bayesian cutoff $s^*(r)$. Solid teal: full-revelation cutoff $\tilde{s}(r; \alpha_s)$. Dotted teal: adherence bound $\underline{t}(r; \alpha_s)$; the shaded region above it is the adherence set. Dashed purple: optimal threshold $t^*(r; \alpha_s)$. Vertical dotted line: $\overline{\alpha}$, below which full implementation fails.

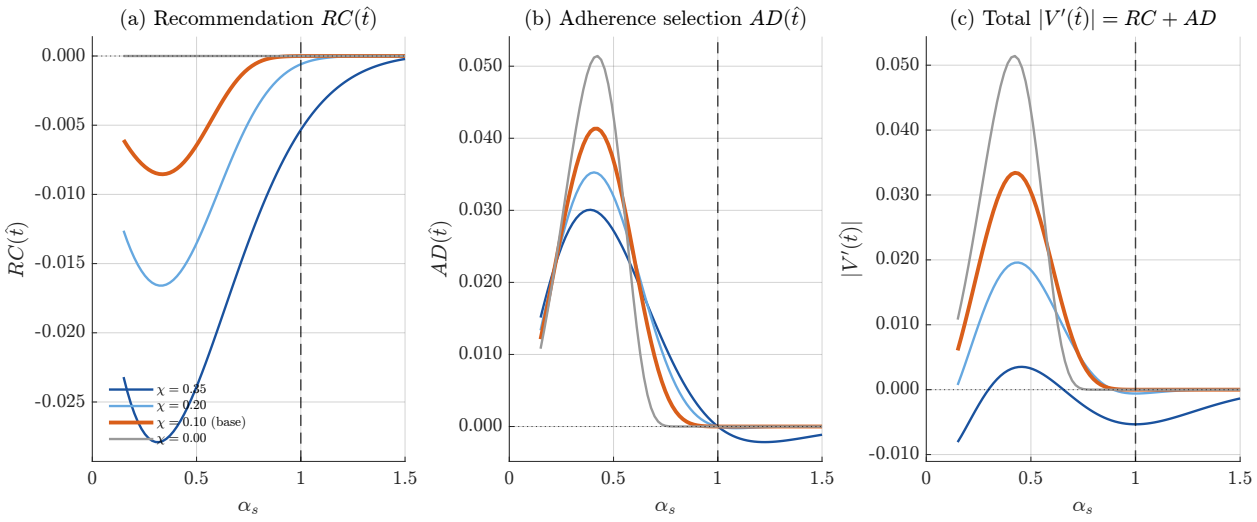
B.4 First-order condition decomposition at the AI-only threshold $\hat{t}$

Figure B.6: AD term as a function of AI-misweighting α_s



Note: The two factors of adherence-selection term evaluated at $t = \hat{t}$ as functions of the AI-misweighting parameter α_s, each shown for a family of residual-information levels χ (colors). Their product, $AD(\hat{t}) = \sum_k \tilde{r}'_k(\hat{t}) \int_{S_k} Q(\tilde{r}_k, s), ds$, is plotted in Figure B.7. Panel (a) plots the derivative of the adherence curves $\tilde{r}'_0(\hat{t})$ (solid) and $\tilde{r}'_1(\hat{t})$ (dashed). Panel (b) plots the marginal-type welfare terms $\int_{S_k} Q(\tilde{r}_k, s), ds$ for $k \in 0, 1$ (solid k=0, dashed k=1), with $S_0 = (-\infty, \hat{t})$ and $S_1 = (\hat{t}, \infty)$. We use our baseline DGP parameters estimated from our experimental data (Appendix B.1): $\hat{\sigma}_r = 2.27$, $\hat{\sigma}_s = 0.79$, $\hat{\rho} = 0.25$ ($\hat{\chi} = 0.10$, bold); the χ family is traced by varying $\rho \in 0, 0.15, 0.25, 0.35$, giving $\chi \in 0.35, 0.20, 0.10, 0.00$.

Figure B.7: FOC decomposition as a function of AI-misweighting



Note: First-order condition evaluated at $t = \hat{t}$ as a function of α_s, shown for a family of residual-information levels χ (colors). Panel (a) plots the recommendation term $RC(\hat{t})$, panel (b) the adherence-selection term $AD(\hat{t})$, and panel (c) the total FOC $|V'(\hat{t})|$. See figure B.6 for details on the baseline χ and its variation.

B.5 Proof of Proposition 1

Notation and parameterization. Write Q_χ and $h_{\omega,\chi}$ when dependence on the residual weight χ matters, and let $f_\omega(s)$ denote the marginal density of s given ω (which does not depend on χ). Throughout the proof, χ is the free parameter: we hold fixed the AI precision σ_s, the conditional dispersion $\tau := \sigma_{r|s} = \sigma_r\sqrt{1-\rho^2}$, the prior and payoff parameters (p, c_{FP}, c_{TP}), and the behavioral primitives $(\hat{r}, g, \delta)$ defined below. Equivalently, (ρ, σ_r) co-move so that

$$\beta := \frac{\rho\sigma_r}{\sigma_s} = 1 - \frac{\chi\tau}{\sigma_s},$$

and every $\chi \in [0, \sigma_s/\tau]$ corresponds to a valid model: $\chi \geq 0$ is the maintained assumption $\sigma_s \geq \rho\sigma_r$, and $\chi \leq \sigma_s/\tau$ is $\rho \geq 0$ (note however, that the precision ordering between s and r can reverse for larger values of χ). The distribution of r given (s, ω) is

$$r \mid s, \omega \sim \mathcal{N}\big(\beta s + (1-\beta)\mu^\omega,\ \tau^2\big)\,;$$

at $\chi = 0$ the conditional mean is simply s, so r is a garbling of s.

As in the main text, integrating the welfare contribution Q over all Receiver types gives the marginal welfare of acting at s,

$$m(s) = \int_{\mathbb{R}} Q_\chi(r, s)\, dr = p\, c_{TP} f_1(s) - (1-p)\, c_{FP} f_0(s),$$

which does not depend on χ. The AI-only threshold solves $m(\hat{t}) = 0$,[69] and since the likelihood ratio f_1/f_0 is strictly increasing in s, m crosses zero only once:

$$\operatorname{sign} m(s) = \operatorname{sign}(s - \hat{t}), \tag{6}$$

a sign fact used repeatedly below. Acting is welfare-improving on average above the AI-only threshold, and welfare-reducing below it.

At the threshold $\hat{t}$, write

$$\hat{r} = \tilde{r}(\hat{t}) = \tilde{\gamma}_0 + \tilde{\gamma}_1\hat{t}, \qquad g = -\tilde{\gamma}_1 \geq 0,$$

[69] Recall $\hat{t} = \frac{1}{2} + \sigma_s^2 \log\frac{(1-p)c_{FP}}{p\, c_{TP}}$.

so the full-revelation DB is $\tilde{r}(s) = \hat{r} - g(s - \hat{t})$, and write the adherence bandwidths at $\hat{t}$ as

$$\delta_1 := \delta_1(\hat{t}) = \hat{r} - \tilde{r}_1(\hat{t}), \qquad \delta_0 := \delta_0(\hat{t}) = \tilde{r}_0(\hat{t}) - \hat{r}.$$

The main text imposes strictly positive bandwidths; here we allow $\delta_i = 0$ only to state boundary cases such as $\Gamma(0,0) = 0$. Finally, let $\bar{c} := p\, c_{TP} + (1-p) c_{FP}$.

The proof turns on two welfare increments:

$$J(\delta) := \int_{\hat{t}}^{\infty} m(s) \left[\Phi\left(\frac{s - \hat{r} + \delta_1}{\tau} \right) - \Phi\left(\frac{s - \hat{r}}{\tau} \right) \right] ds + \int_{-\infty}^{\hat{t}} m(s) \left[\Phi\left(\frac{s - \hat{r} - \delta_0}{\tau} \right) - \Phi\left(\frac{s - \hat{r}}{\tau} \right) \right] ds, \tag{7}$$

$$S(g) := \int_{\mathbb{R}} m(s) \left[\Phi\left(\frac{s - \hat{r} + g(s - \hat{t})}{\tau} \right) - \Phi\left(\frac{s - \hat{r}}{\tau} \right) \right] ds. \tag{8}$$

As Step 1 shows, both are welfare gains at $\chi = 0$ relative to a flat DB at $\hat{r}$ (a Receiver who ignores the AI signal): $J(\delta)$ is the gain from binary coarsening at $\hat{t}$, and $S(g)$ the gain from full revelation. Define

$$\bar{\chi}(g, \delta) = \frac{\pi}{\bar{c}} \big(J(\delta) - S(g) \big), \qquad J_\infty := \int_{\hat{t}}^{\infty} m(s) \left[1 - \Phi\left(\frac{s - \hat{r}}{\tau} \right) \right] ds - \int_{-\infty}^{\hat{t}} m(s)\, \Phi\left(\frac{s - \hat{r}}{\tau} \right) ds.$$

As shown in Step 4, S maps $[0, \infty)$ continuously and strictly increasingly onto $[0, J_\infty)$ while $J(\delta) \in [0, J_\infty)$, so the slope tolerance

$$\Gamma(\delta) := S^{-1}(J(\delta))$$

is well defined.

Step 1: welfare generated by a DB. For a measurable decision boundary $b : \mathbb{R} \to \mathbb{R}$, let

$$W_\chi[b] = \int_{\mathbb{R}} \int_{b(s)}^{\infty} Q_\chi(r, s)\, dr\, ds$$

denote the expected welfare when each type r acts iff $r > b(s)$.[70] Since $Q_\chi(r,s) = p\, c_{TP} h_{1,\chi}(r,s) -$

[70] If $b_t(s)$ is the step DB induced by binary coarsening at t, then the main-text objective is $V(t) = W_\chi[b_t]$.

$(1-p)c_{FP}h_{0,\chi}(r,s)$, integrating r out using its conditional normal distribution given (s,ω) yields

$$W_\chi[b] = p\,c_{TP}\int f_1(s)\,\Phi\Big(\tfrac{\beta s+(1-\beta)\mu^1-b(s)}{\tau}\Big)\,ds \;-\; (1-p)c_{FP}\int f_0(s)\,\Phi\Big(\tfrac{\beta s+(1-\beta)\mu^0-b(s)}{\tau}\Big)\,ds. \quad (9)$$

At $\chi=0$ we have $\beta=1$: the two Φ terms coincide, and (9) collapses to

$$W_0[b] = \int_{\mathbb{R}} m(s)\,\Phi\left(\frac{s-b(s)}{\tau}\right)ds. \quad (10)$$

We compare three DBs: the flat DB $b_H(s)\equiv\hat r$ (ignoring the AI signal), the full-revelation DB $b_{FR}(s)=\hat r-g(s-\hat t)$, and the step DB induced by binary coarsening at $\hat t$,

$$b_{BC}(s)=\begin{cases}\hat r+\delta_0, & s<\hat t,\\ \hat r-\delta_1, & s>\hat t,\end{cases}$$

whose two levels are determined by the adherence bandwidths. Comparing (10) with (7)–(8),

$$W_0[b_{BC}]-W_0[b_H]=J(\delta),\qquad W_0[b_{FR}]-W_0[b_H]=S(g).$$

Hence, writing $\Delta(\chi)=W_\chi[b_{BC}]-W_\chi[b_{FR}]$ for the welfare gain of binary coarsening at $\hat t$ over full revelation,

$$\Delta(0)=J(\delta)-S(g). \quad (11)$$

Step 2: a uniform Lipschitz bound in χ. For a fixed DB b, χ enters (9) only through the conditional mean $\beta s+(1-\beta)\mu^\omega = s-\frac{\chi\tau}{\sigma_s}(s-\mu^\omega)$. Differentiating under the integral,[71]

$$\frac{dW_\chi[b]}{d\chi} = -\frac{1}{\sigma_s}\Bigg[p\,c_{TP}\int f_1(s)(s-\mu^1)\,\varphi\Big(\tfrac{\beta s+(1-\beta)\mu^1-b(s)}{\tau}\Big)\,ds \\ -(1-p)c_{FP}\int f_0(s)(s-\mu^0)\,\varphi\Big(\tfrac{\beta s+(1-\beta)\mu^0-b(s)}{\tau}\Big)\,ds\Bigg].$$

Applying this to b_{BC} and b_{FR} and differencing, $\Delta'(\chi)$ has the same form, with each $\varphi(\cdot)$ replaced by the difference of the two values of φ at the DBs $b_{BC}(s)$ and $b_{FR}(s)$. Since φ takes values in $(0,\frac{1}{\sqrt{2\pi}}]$, each such difference is at most $\frac{1}{\sqrt{2\pi}}$ in absolute value. Next, observe that since

[71] Differentiation under the integral is justified by dominated convergence, with dominating function $\frac{1}{\sqrt{2\pi}\sigma_s}|s-\mu^\omega|f_\omega(s)$.

f_ω is normal, the mean absolute deviation is given by: $\int |s-\mu^\omega| f_\omega(s)\,ds = \sigma_s\sqrt{2/\pi}$ for each ω. Therefore,

$$|\Delta'(\chi)| \le \frac{p\,c_{TP}+(1-p)c_{FP}}{\sigma_s}\cdot\frac{1}{\sqrt{2\pi}}\cdot\sigma_s\sqrt{\frac{2}{\pi}} = \frac{\bar{c}}{\pi}. \tag{12}$$

This bound is chosen for transparency; tighter bounds exist (e.g., exploiting the proximity of the two DBs) but are less transparent.

Step 3: dominance below $\bar{\chi}$. By (11), (12), and the definition of $\bar{\chi}$,

$$\Delta(\chi) = \Delta(0) + \int_0^\chi \Delta'(x)\,dx \ge J(\delta) - S(g) - \frac{\bar{c}}{\pi}\chi = \frac{\bar{c}}{\pi}\big(\bar{\chi}(g,\delta)-\chi\big).$$

Hence $\chi < \bar{\chi}(g,\delta)$ implies $\Delta(\chi) > 0$: binary coarsening at $\hat{t}$ strictly dominates full revelation. This proves part (i).

Step 4: properties of J and S. By inspection, $J(0,0) = S(0) = 0$. Differentiating (7)–(8) (justified as in Step 2),

$$\frac{\partial J}{\partial \delta_1} = \frac{1}{\tau}\int_{\hat{t}}^{\infty} m(s)\,\varphi\!\left(\frac{s-\hat{r}+\delta_1}{\tau}\right)ds, \qquad \frac{\partial J}{\partial \delta_0} = -\frac{1}{\tau}\int_{-\infty}^{\hat{t}} m(s)\,\varphi\!\left(\frac{s-\hat{r}-\delta_0}{\tau}\right)ds,$$
$$S'(g) = \frac{1}{\tau}\int_{\mathbb{R}} m(s)(s-\hat{t})\,\varphi\!\left(\frac{s-\hat{r}+g(s-\hat{t})}{\tau}\right)ds.$$

By the sign fact (6), the integrand of $\partial J/\partial\delta_1$ is positive ($m>0$ above $\hat{t}$), that of $\partial J/\partial\delta_0$ is negative ($m<0$ below $\hat{t}$), and $m(s)(s-\hat{t})>0$ for almost every s. Hence all three derivatives are strictly positive: J is continuous and strictly increasing in each bandwidth, and S is continuous and strictly increasing; in particular $J(\delta)>0$ iff $\delta \neq (0,0)$. Intuitively, widening either bandwidth pulls the step DB further toward the AI-only rule—acting above $\hat{t}$, not acting below—and steepening the full-revelation DB does the same, which at $\chi=0$ can only raise welfare.

For the limits, as $\min\{\delta_1,\delta_0\}\to\infty$ we have $\Phi\big(\frac{s-\hat{r}+\delta_1}{\tau}\big)\to 1$ for $s>\hat{t}$ and $\Phi\big(\frac{s-\hat{r}-\delta_0}{\tau}\big)\to 0$ for $s<\hat{t}$; and as $g\to\infty$, $\Phi\big(\frac{s-\hat{r}+g(s-\hat{t})}{\tau}\big)\to \mathbf{1}\{s>\hat{t}\}$ for every $s\neq\hat{t}$. Since the bracketed terms in (7)–(8) are bounded by one and $|m|$ is integrable, dominated convergence gives $J(\delta)\to J_\infty$ and $S(g)\to J_\infty$: both approach the welfare gain of the AI-only rule itself. The limits are not attained: for finite δ,

$$J_\infty - J(\delta) = \int_{\hat{t}}^{\infty} m(s)\left[1-\Phi\!\left(\frac{s-\hat{r}+\delta_1}{\tau}\right)\right]ds - \int_{-\infty}^{\hat{t}} m(s)\,\Phi\!\left(\frac{s-\hat{r}-\delta_0}{\tau}\right)ds,$$

and both terms are strictly positive by (6) ($m > 0$ against $1 - \Phi > 0$ above $\hat{t}$; $m < 0$ against $\Phi > 0$ below). Replacing the arguments of Φ by $\frac{s-\hat{r}+g(s-\hat{t})}{\tau}$ gives the identical expression for $J_\infty - S(g)$, and the same argument applies. Hence $J(\delta) < J_\infty$ and $S(g) < J_\infty$ for all finite arguments, and S maps $[0, \infty)$ continuously and strictly increasingly onto $[0, J_\infty)$.

Step 5: parts (ii) and (iii). Since $\bar{\chi}(g, \delta) = \frac{\pi}{\bar{c}}\big(J(\delta) - S(g)\big)$, Step 4 immediately gives part (ii): $\bar{\chi}$ is continuous, strictly increasing in δ_1 and in δ_0, and strictly decreasing in g.

By Step 4, $J(\delta) \in [0, J_\infty)$ for every finite δ and S^{-1} exists on $[0, J_\infty)$, so $\Gamma(\delta) = S^{-1}(J(\delta))$ is well defined and finite. By strict monotonicity of S,

$$\bar{\chi}(g, \delta) > 0 \iff J(\delta) > S(g) \iff g < \Gamma(\delta),$$

the main equivalence in part (iii): coarsening at $\hat{t}$ dominates for some $\chi > 0$ exactly when the adherence it commands is worth more, at $\chi = 0$, than the Receiver's own response to the fully revealed signal. The remaining properties of Γ follow by composition: J and S^{-1} are continuous and strictly increasing, so Γ is continuous and strictly increasing in each bandwidth, with $\Gamma(0, 0) = S^{-1}(0) = 0$ and $\Gamma(\delta) < \infty$ for every finite δ. Since $J(d, d) \uparrow J_\infty$ as $d \to \infty$ while $S(g) \uparrow J_\infty$ only as $g \to \infty$ (so $S^{-1}(x) \to \infty$ as $x \uparrow J_\infty$), we get $\Gamma(d, d) \uparrow \infty$.

The final claims follow. If $g = 0$, then $\bar{\chi}(0, \delta) > 0 \Leftrightarrow J(\delta) > S(0) = 0 \Leftrightarrow \delta \neq (0, 0)$. And for any fixed g, choose d with $\Gamma(d, d) > g$; by monotonicity of Γ, $g < \Gamma(\delta)$ whenever both bandwidths exceed d. ■

C Main Experiment Design

C.1 Main Experiment Instructions

Figure C.1: Instructions to participants to highlight prior and AI details

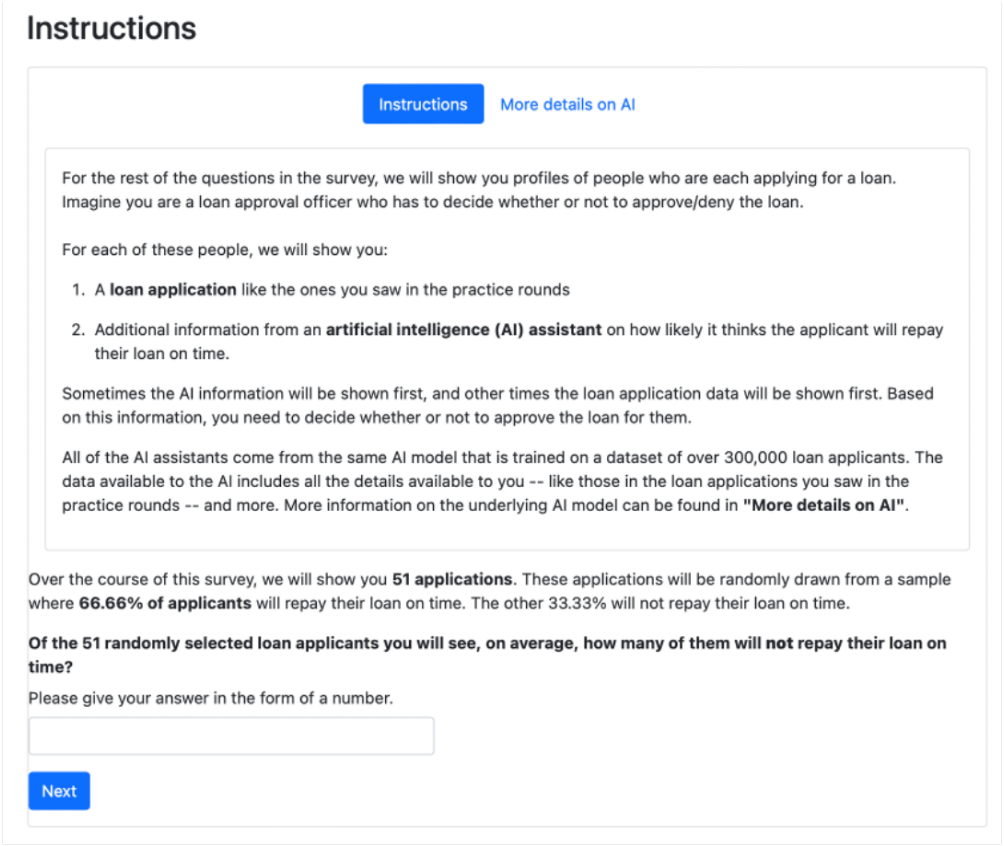


Figure C.2: Detailed instructions give "More details on AI"



Figure C.3: Instructions for each treatment condition highlight AI performance (*Binary Medium*) condition shown here as example

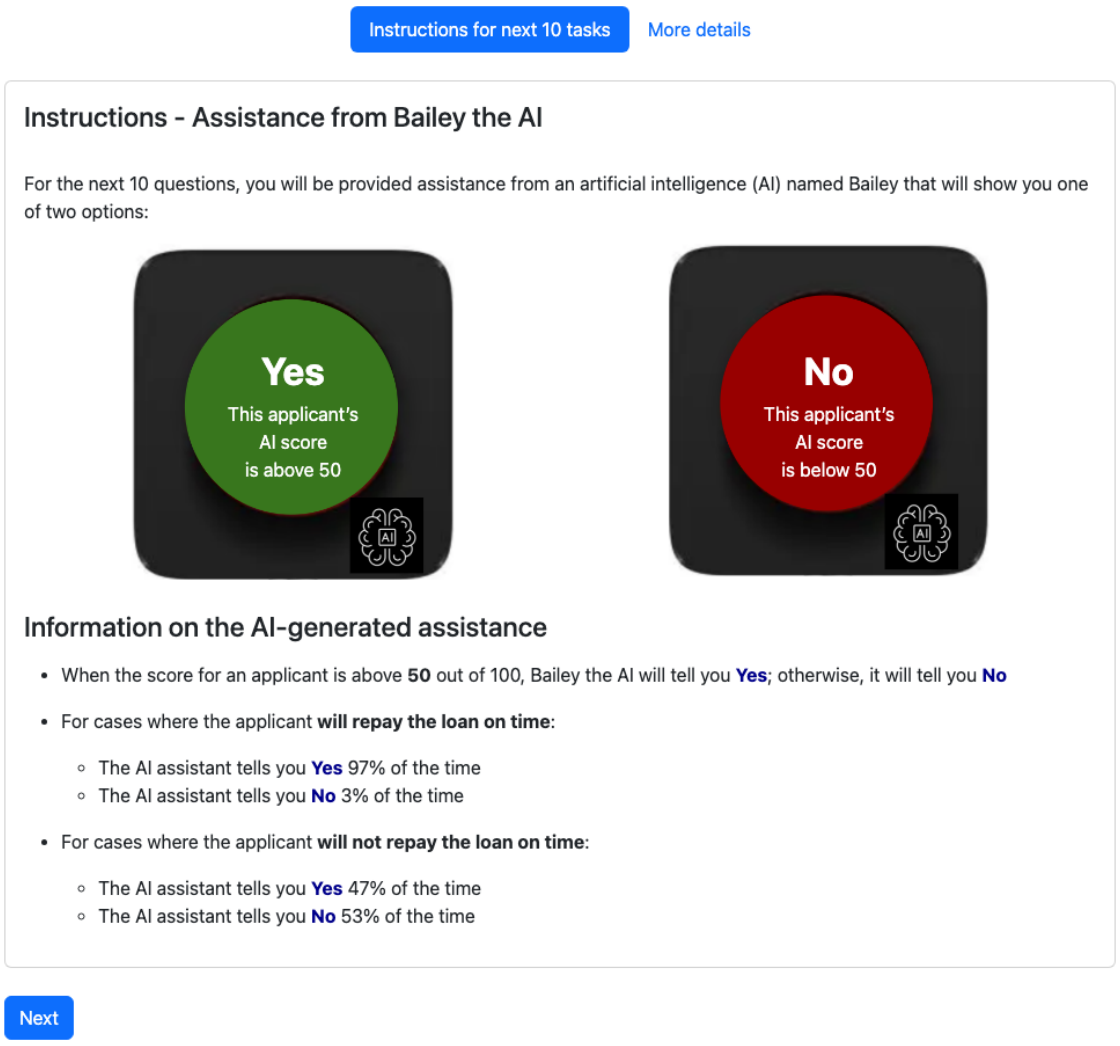


Figure C.4: Instructions for the *Probability* condition

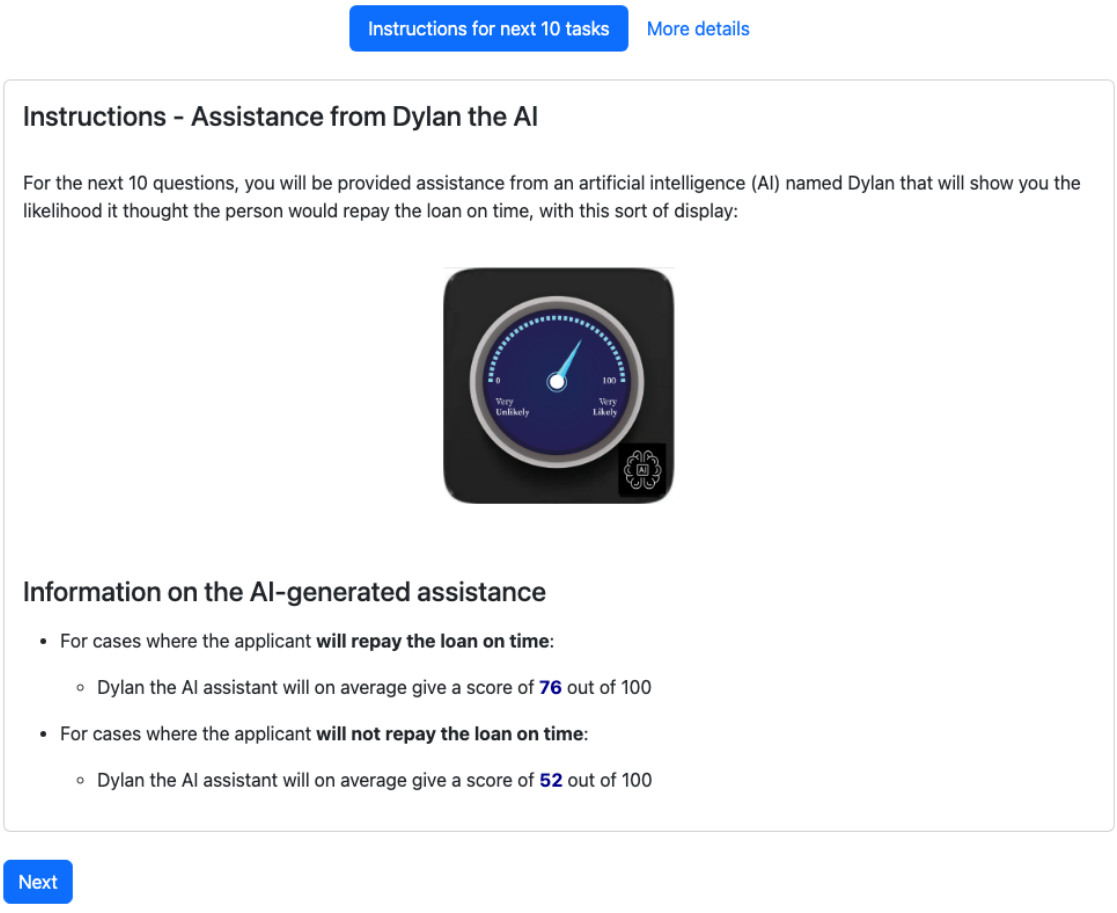


Figure C.5: Detailed instructions for each treatment condition (*Binary Medium*) condition shown here as example

Instructions for next 2 tasks More details

Information on the loan applicants

- On average, about **66.66% of loan applicants** will repay their loan on time

Information on the AI-generated assistance

- When the score for an applicant is above **50** out of 100, Bailey the AI will tell you **Yes**; otherwise, it will tell you **No**

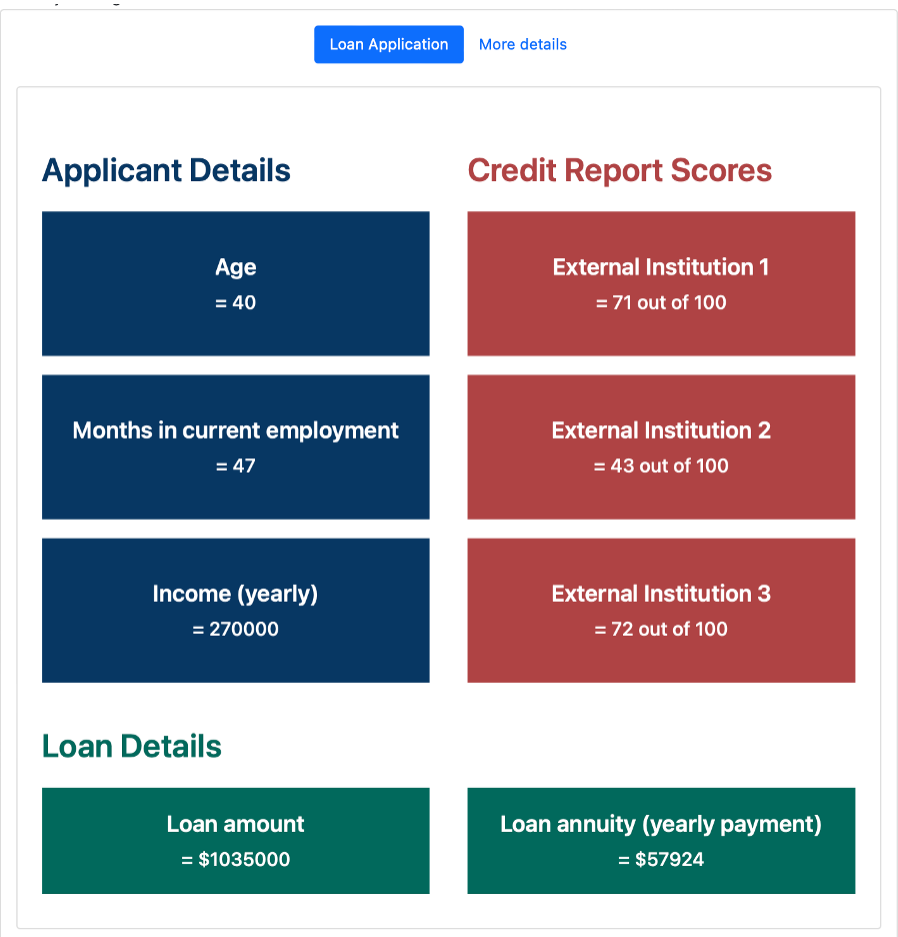


Information on performance in practice rounds

- During the practice rounds, you decided on 10 applications, where 66.66% applicants would go on to repay their loans. Here is how you performed, as well as how Bailey the AI would've performed if it had assisted you:
  - Of the 6 applicants who went on to **repay the loan on time**:
    - You correctly approved the loan for 1 out of 6 of them
    - Bailey the AI would've told you **Yes** for 6 out of 6 of them
  - Of the 4 applicants who would go on to **not repay the loan on time**:
    - You correctly denied the loan for 0 out of 4 of them
    - Bailey the AI would've told you **No** for 2 out of 4 of them

Next

Figure C.6: Example of first signal (in this case, loan application)



- **Loan amount:** The amount of credit that will be granted to the applicant if the loan is approved.
- **Loan annuity:** The yearly payment that the borrower will have to make on a regular basis (to cover both the principal and interest).
- **Age:** The applicant's current age.
- **Months in current employment:** The number of months the applicant has been in their current employment.
- **Income (yearly):** The yearly income the applicant currently earns.
- **External Institution 1-3:** Normalized credit score from external data source #1-3 (e.g., FICO, VantageScore), ranging from 0-100, with 100 being the most credit-worthy.

Figure C.7: Example of first "posterior" elicitation



Figure C.8: Example of second signal (in this case, AI with *Binary Medium* condition shown here as example)

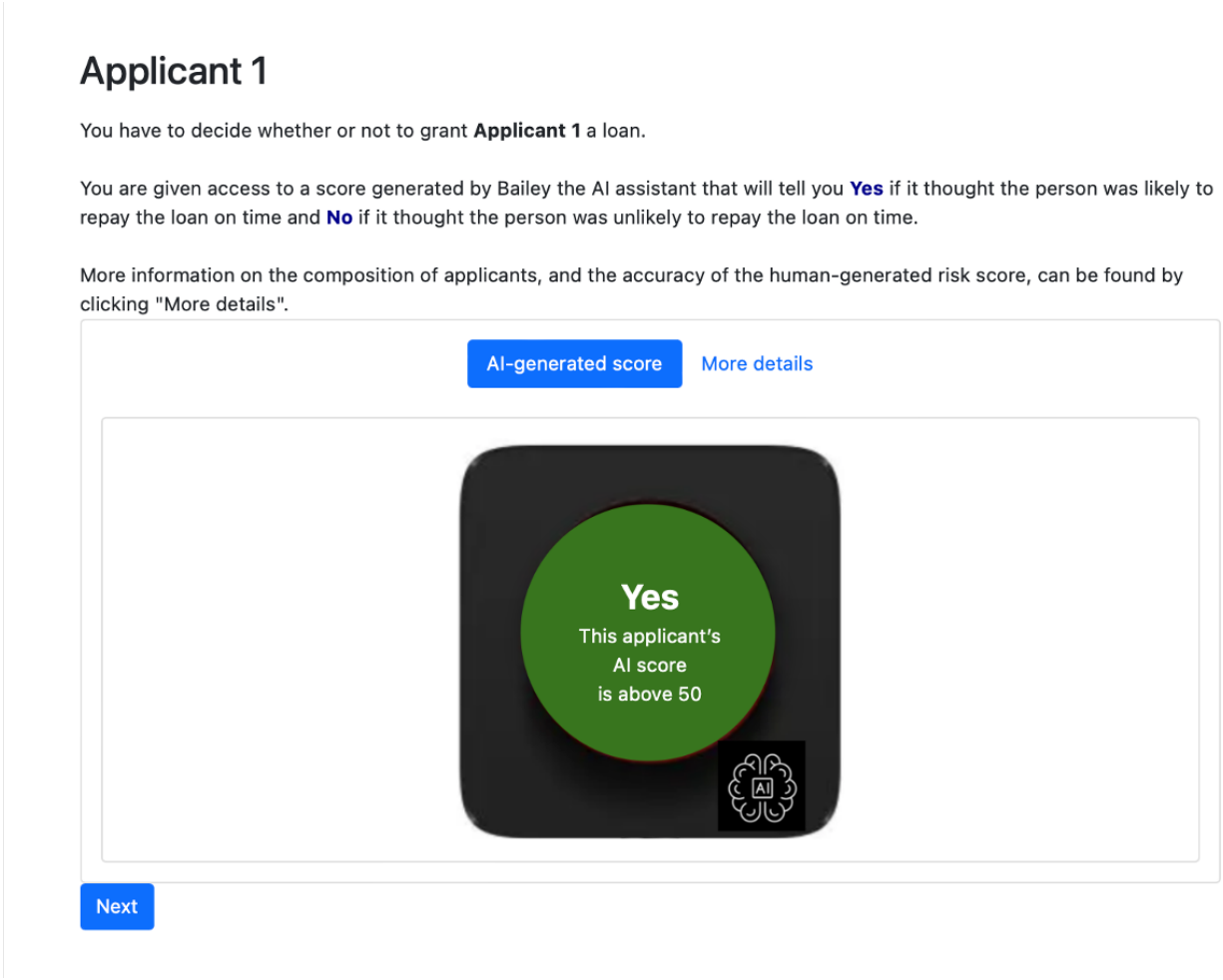
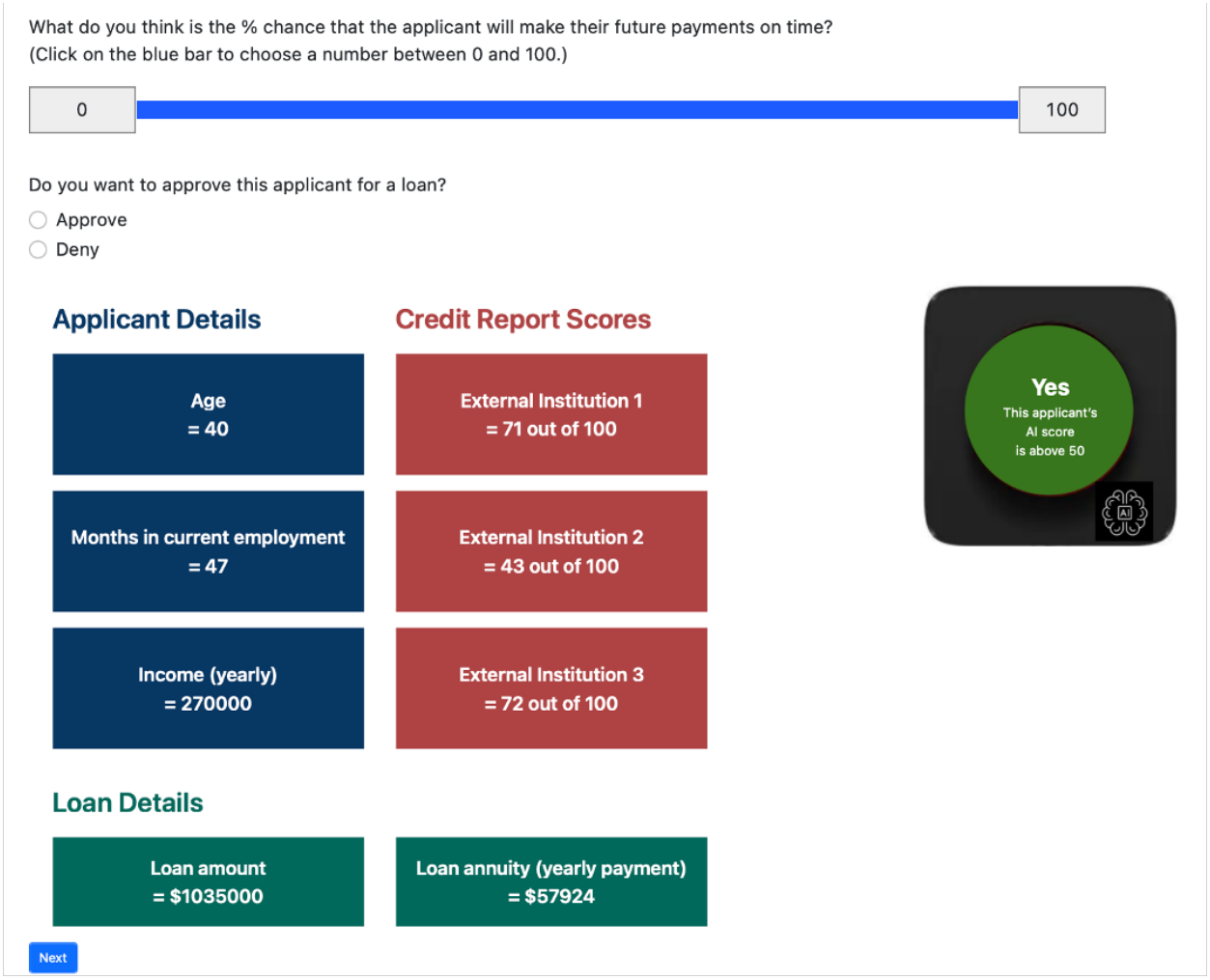


Figure C.9: Approve/deny screen displays both signals again for easy decision-making



C.1.1 Additional survey questions

Demand for AI – Participants first set a threshold for the binary ("Yes/No") assistant—*"What threshold would you like to set? (Choose a number between 0 and 100.)"*—and then choose among the available options:

> Recall that you can choose to get help from two types of AI assistant. **Yes/No**, who tells you *Yes* if it thinks the person is likely to repay the loan on time and *No* otherwise (telling you *Yes* when it thinks the applicant's score is above your chosen threshold, and *No* below); or **Score**, who shows the likelihood it thinks the person will repay, e.g., 67 out of 100. For the next few loan applications, you get to choose which type of AI you would prefer to help you.

We then elicit willingness to pay for each form through an incentivized multiple price list (MPL):

> In the next questions you make a series of decisions by completing a list; each row presents two options. **Option A** always involves using the AI assistant for some amount of money per loan application; **Option B** involves not using any AI assistance. The price increases from -10 to 0 to $+10$ cents down the rows. When the price is *negative* (e.g., -8 cents), you are *paid* 8 cents per application for using the AI; when it is *positive* (e.g., 8 cents), you *pay* 8 cents per application to use it.
>
> *For the next cases, I prefer to...* [Option A: use the AI's assistance for the price on the left] / [Option B: do not use any AI assistance], for each price on the list.
>
> One row is randomly selected and implemented for the corresponding decisions, where you earn a bonus of 10 cents for each loan you correctly grant (applicant repays) or correctly deny (applicant does not repay), plus the condition stated on the selected row.

The highest price on the list (10 cents) equals the bonus for a correct decision. Choices reveal willingness to pay in units of the per-case reward.

Perceived residual information

> You just saw 10 loan applications as well as Dylan the AI's score for them. Imagine that you will go through 100 more loan applications.
>
> If you **will get both the loan application information and Dylan the AI's score [only had access to the Dylan the AI's score/only had access to the loan application itself]**, how many questions do you **think you will get right out of the 100 applications**?

Demographics – We collect standard demographic questions (gender, education, age, race/ethnicity, household income), as well as previous employment in financial institutions, work/certification

experience in loan underwriting and processing, historical experience applying for loans, and familiarity with AI.

Biases – Participants answer a series of cognitive questions documented below to establish a baseline measure of susceptibility to behavioral biases. It is important to note that all these bias measures were deliberately elicited outside the loan context to avoid priming participants. While this design choice reduces the risk of demand effects, it may also attenuate the observed influence of certain biases compared to what would be seen in the loan context. The extent to which these biases manifest in real-world decision-making likely depends on how closely the elicitation tasks align with the specific structure of the decision environment.

C.1.2 Bias elicitations

In all of the bias elicitation questions below, we additionally tell participants how their decision is incentivized (1 point = 1 cent).

> We will pay you more points the closer your decision is to the statistically-correct percentage chance given the information we provide.
>
> - Specifically, we will pay you 100 points if your decision corresponds to this correct answer. We subtract 3 points for every percentage point you are away from the correct answer.
> - You cannot make losses, meaning you always earn at least 0 points.

Base-rate Neglect We utilize a question from the base-rate neglect cognitive task in Enke et al. (2023), which is a simplified variant of the taxi-cab problem designed to capture responses that neglect base rates, as introduced in Kahneman et al. (1982). In this task, subjects estimate the probability that a bike is actually defective, given that the base rate for defects is 10% and that a quality control machine classifies the bike as defective. A common incorrect response, which neglects the base rate, is 75%, whereas the statistically correct answer is 25%.

The base-rate parameter (BRN) is estimated by solving the following equation for BRN:

$$\frac{X}{100}\left(0.75 \times 0.10^{BRN} + 0.25 \times 0.9^{BRN}\right) - \left(0.75 \times 0.10^{BRN}\right) = 0$$

where X is the subject's stated probability (in percentage form) that a bike is actually defective.

The resulting BRN parameter indicates how much the subject either underweights or overweights the prior probability of a defective bike. An $BRN = 1$ indicates no base-rate neglect. A value between 0 and 1 indicates underweighting of the prior (base-rate neglect), whereas a value greater than 1 suggests overweighting of the prior. Since a negative parameter (any value of X exceeding 75) lacks meaningful interpretation, we treat these observations as NA.

> Assume that, on average, out of every 100 bicycles produced by a bike manufacturer, 90 are good and 10 are defective.
>
> There is a *human quality control worker* who inspects bicycles at the end of the production line, just as the AI does. However, the worker's classification can vary from time to time. On average, the worker correctly identifies a bicycle (as good or defective) 75 out of 100 times, but misidentifies it 25 out of 100 times.
>
> Now a bicycle produced by the manufacturer has randomly been selected. Next, this specific bicycle was inspected by the quality control worker, and you have been told about the worker's classification below. Based on this classification, your task is to state the likelihood (percentage chance) that this specific bicycle is actually defective.
>
> You learn that the randomly selected bicycle has been classified as defective by the quality control worker. **What do you think is the likelihood (percentage chance) that it is actually defective? (Round to the nearest integer.)**

AI Misweighting We adapt this question from the base-rate neglect cognitive task in Enke et al. (2023). We measure AI misweighting by the ratio of the response to the AI question (see below) to the response to the base-rate neglect question (see above):

$$\text{AI misweighting} = \frac{\text{AIresponse}}{\text{baseRateNeglect}}$$

A value of 1 indicates Bayesian weighting of the AI signal; a value below 1 indicates AI underweighting, and above 1 AI overweighting. Because the human signal carries little residual information in our setting, the empirically relevant deviation is underweighting: in the heterogeneity analysis we accordingly split participants at the median into an *AI Underweighting (AU)* group (below median, who underweight the AI more heavily) and a *Near-Bayes* group (above median), rather than interpreting overweighting separately.

Assume that, on average, out of every 100 bicycles produced by a bike manufacturer, 90 are good and 10 are defective.

There is a *quality control machine* powered by *artificial intelligence (AI)* that labels whether bicycles are good or defective at the end of the production process. The machine's classification can vary from time to time. On average, the machine correctly classifies a bicycle (as good or defective) 75 out of 100 times, but incorrectly classifies it 25 out of 100 times.

Now a bicycle produced by the manufacturer has randomly been selected. Next, this specific bicycle was run through the quality control machine, and you have been told about the machine's classification below. Based on this classification, your task is to state the likelihood (percentage chance) that the specific bicycle is actually defective.

You learn that the randomly selected bicycle has been classified as *defective* by the AI quality control machine. **What do you think is the likelihood (percentage chance) that it is actually defective? (Round to the nearest integer.)**

Correlation Neglect We utilize a question from the correlation neglect cognitive task in Enke et al. (2023), in which subjects are asked to estimate the weight of a bucket based on provided estimates from Ann and Charlie. A common incorrect response is to compute the average of 40 and 70, yielding 55, whereas the statistically correct answer is 40. We define the correlation neglect parameter (CN) as

$$CN = \frac{\text{correlation_neglect}}{40}.$$

A value of CN closer to 1 is desirable (no neglect), and a value above 1 indicates correlation neglect.

Three people: Ann, Bob, and Charlie. Each of them is interested in estimating the weight of a water bucket in pounds.

Ann and Bob both get to take a peek at the bucket. They are equally good at estimating. Each tends to get their weight estimates right on average, but sometimes they make random mistakes. Ann and Bob are equally likely to make mistakes in any given estimate they make.

Ann and Bob both share their estimates with Charlie, who never sees the bucket himself; Charlie's estimate is the average of Ann's and Bob's estimates. You do not see the bucket either, but you talk to Ann and Charlie, who share their best estimates of the weight of the bucket with you:

- Ann's estimate: 70
- Charlie's estimate: 40

> Your task is to estimate the weight of the bucket. **What is your best estimate of the weight of the bucket?** (Round to the nearest integer.)

Balls and Urns We utilize a question from the balls-and-urns belief updating task in Enke et al. (2023). In this task, subjects are asked to indicate the percentage chance that the selected bag is the one that contains more red chips, given that the drawn chip is red. Often, subjects exhibit conservatism bias by providing posterior estimates strictly between 50% and 70%, even though the Bayesian answer is 70%. We define the belief updating parameter BU as the ratio of the subject's response to 70:

$$BU = \frac{\text{balls_urns}}{70}.$$

A BU value closer to 1 is desirable. In particular, BU values between $\frac{50}{70}$ and 1 indicate conservatism bias, while BU values greater than 1 suggest over-updating. Responses yielding BU values below $\frac{50}{70}$ or exactly $\frac{100}{70}$ are considered anomalous and are treated as NA.

> There are two bags. One bag contains 70 red chips and 30 blue chips. The other one contains 30 red chips and 70 blue chips.
>
> We secretly flipped a (fair) coin. If it came up **HEADS**, we chose the bag with more red chips. If it came up **TAILS**, we chose the bag with more blue chips. Therefore, you do not know which bag was selected.
>
> Next, we drew one chip at random from the bag selected by the coin toss. You will learn the color of this randomly-drawn chip below, then you need to figure out (in percent) which bag was selected.
>
> You are told that one **red chip** has randomly been drawn from the secretly selected bag. **What do you think is the likelihood (percentage chance) that the selected bag is the one with more red chips?** (Round to the nearest integer.)

Aggregation bias We adapt a variant of the leading example of Piccione and Rubinstein (2026) in order to capture a range of failure to correctly (Bayesian) aggregate signals. The correct answer here would be 98%, but common failures include taking the minimum (20%), the average (45%), the maximum (70%), the product (14%), or the probability of "at least one of the two events" happening under independence, i.e., the complement of the product of the complements (76%).

$$PR = \frac{\text{aggregation_bias}}{70}.$$

As subjects rarely answer anything remotely close to correct (98%), we define this term (PR for Piccione-Rubinstein) to be > 1 for anyone who correctly realizes that their posterior should be higher than the maximum of the two signals, and < 1 if they fail to update in the right direction.

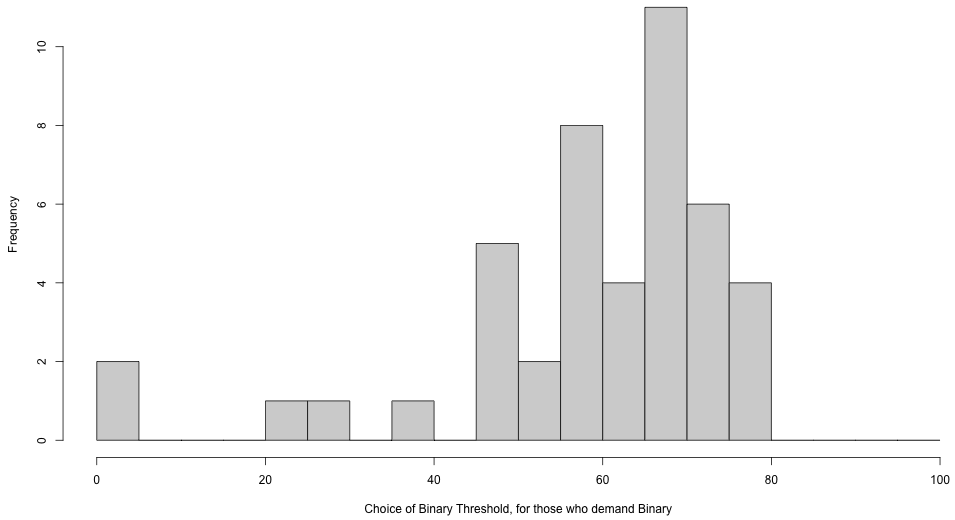
> The proportion of newborns in a country with a specific genetic trait is 1%.
>
> Two screening tests, A and B, are used to identify this trait in newborns. However, the tests are not perfect.
>
> A study has found that:
>
> - 70% of the newborns who are found to be positive according to test A have the genetic trait.
> - 20% of the newborns who are found to be positive according to test B have the genetic trait.
>
> The two tests are conditionally independent, meaning:
>
> - When a newborn has the genetic trait, a positive result in one test does not affect the likelihood of a positive result in the other.
> - When a newborn does not have the genetic trait, a positive result in one test does not affect the likelihood of a positive result in the other.
>
> Suppose that a newborn is found to be positive according to both tests. **What is your estimate of the likelihood (in %) that this newborn has the trait?**

C.2 Empirical Evidence—Main Experiment

C.2.1 Demand

Before asking participants to choose between the three AI assistance types, we elicited their ideal binary threshold—the score above which they would want the model to output a "Yes, this applicant's score is above [threshold]" recommendation. Figure C.10 shows the distribution: choices are skewed to the right of 50, with many participants selecting thresholds above the best-performing decision threshold. This pattern may reflect a tendency to emulate a "cherry-picking" strategy

Figure C.10: Chosen binary threshold among participants who opted for a binary AI signal



Note: Distribution of the binary threshold values, selected by participants who opted for a binary AI signal.

that loan specialists are often trained under, prioritising identification of applicants highly likely to repay over eliminating lemons.

Table C.1 reports the full breakdown of AI-assistance choices. "No AI" can be viewed as an extreme coarsening that pools all signals into one uninformative message—in practice, this choice could reflect algorithm aversion and/or an associated cognitive cost to receiving or processing AI information, the latter of which is not captured in our theoretical framework.

Table C.1: Willingness to Pay and Highest Payoff Treatment, by Choice of AI Assistant

		Highest Payoff Treatment				
Choice of AI	Total (#)	Human Only (% of column)	Binary (% of column)	Prob (% of column)	Mean WTP Binary (cents/case)	Mean WTP Prob (cents/case)
No AI	35	57	19	13	0.8 (4.48)	-0.2 (3.86)
Binary	45	22	31	33	3.8 (4.10)	2.4 (4.04)
Probability	70	22	50	54	3.2 (4.11)	2.5 (3.74)

Note: Summary of participants' choice of AI assistant, their highest payoff treatment, and mean willingness to pay (WTP), expressed in cents per case. The *Highest Payoff Treatment* columns report column percentages—the distribution of AI-assistant choices within each highest-payoff-treatment group—so each column sums to 100; *Total* is the number of participants making each choice.

Participants appear to select into the types of signals most effective for them, even though they received no feedback on how they did in each treatment. The majority of participants for whom the

highest-payoff treatment was *Human-only* chose not to use any AI,[72] while the majority selected probability scores when they did best with them. Those who did best in one of the three binary treatments were more mixed, with about half preferring a probability risk score instead of a binary.

We also explore how demographic characteristics relate to AI-assistant choice (Table C.2). Participants who chose no AI were older on average (47 years vs. 42 and 41 for the probability and binary groups, respectively), and more likely to report never having used any AI before (40%, compared to 17% and 22%). This suggests that familiarity with AI—and perhaps comfort with technology more broadly—plays a role in adoption decisions.

Table C.2: Participant Demographics by AI Assistant Choice

	Binary	No AI	Probability
Age (years)	41.51	47.31	43.80
Female (%)	64.44	62.86	52.17
Bank/Credit Union Experience (%)	100.00	91.43	95.65
Applied to Loans Before (%)	97.78	97.14	95.65
Loan Amount (USD$)	179364	189439	205203
Loan Specialist Experience (years)	12.09	13.77	11.42
Used Supervised AI Before(%)	24.44	11.43	23.19
Never Used any AI (%)	22.22	40.00	17.39

Note: Summary statistics on participant characteristics, such as age, gender, bank/credit union experience, loan history, loan amount, loan specialist experience, and previous AI usage, broken down by the type of AI assistant selected: Binary, No AI, or Probability. Units in parentheses.

C.2.2 Heterogeneous treatment effects

For the following HTE regression tables, Column (1) shows coefficients from a regression of decision accuracy on treatment indicators and fixed effects (individual, case, and block order) for the below median group. Column (2) shows coefficients for the above median group. Column (3) shows the coefficients on the interaction terms from a regression that includes treatment indicators, above/below median indicators, treatment × above/below median interactions, and case fixed effects, where only the interaction coefficients are presented, representing the difference in treatment effects between the above and below median groups.

Note that any discrepancies between the coefficients in the regression tables (e.g., Table C.5) and the "Overall Accuracy" columns of these true/false positive and negative summary tables (e.g.,

[72]This may be mechanically correlated if people who choose not to use AI are the ones who are reluctant to follow it when forced to use it, leading to lower payoffs under the AI treatments and relatively higher payoffs under the *Human-only* treatment.

Table C.6) arise because the former incorporate fixed effects, while the latter report simple means. Throughout these tables, positive = approve (loan repaid), and the TP/TN/FP/FN shares are percentages of all decisions in each cell.

Gender

Table C.3: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	Male	Female	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.094^{***}	-0.025	0.072^{**}
	(0.026)	(0.018)	(0.030)
Probability	-0.051^{**}	-0.012	0.036
	(0.025)	(0.022)	(0.032)
Binary High (70)	-0.083^{***}	-0.039^{*}	0.037
	(0.021)	(0.023)	(0.030)
Binary Low (30)	-0.056^{**}	-0.028	0.024
	(0.022)	(0.018)	(0.027)
Observations	3,100	4,350	7,450
R^2	0.324	0.304	0.283
Adjusted R^2	0.297	0.280	0.277
Note:	Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table C.4: Share of True/False Positives and Negatives by Treatment and Gender

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
Female						
Human Only	49.89	19.77	14.83	15.52	69.66	870
Binary Med (50)	54.61	20.20	12.59	12.59	74.82	683
Probability	49.72	19.10	16.01	15.17	68.82	712
Male						
Human Only	46.77	18.55	16.13	18.55	65.32	620
Binary Med (50)	57.98	17.11	14.26	10.65	75.10	526
Probability	52.59	18.84	13.25	15.32	71.43	483

Age

Table C.5: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	Age (Below Median)	Age (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.062^{***}	-0.058^{***}	-0.002
	(0.022)	(0.020)	(0.030)
Probability	-0.032	-0.028	-0.001
	(0.024)	(0.023)	(0.032)
Binary High (70)	-0.066^{***}	-0.049^{**}	0.011
	(0.022)	(0.024)	(0.032)
Binary Low (30)	-0.019	-0.062^{***}	-0.040
	(0.019)	(0.020)	(0.027)
Observations	4,000	3,500	7,500
R^2	0.304	0.323	0.283
Adjusted R^2	0.279	0.297	0.277
Note:	Asymptotic standard errors clustered at the individual level in parentheses.		
			*p<0.1; **p<0.05; ***p<0.01

Table C.6: Share of True/False Positives and Negatives by Treatment and Age

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
Age (Above Median)						
Human Only	49.14	19.14	16.29	15.43	68.29	700
Binary Med (50)	59.47	18.41	11.15	10.97	77.88	565
Probability	53.20	17.79	14.95	14.06	71.00	562
Age (Below Median)						
Human Only	48.12	19.25	14.50	18.12	67.38	800
Binary Med (50)	53.22	19.02	15.34	12.42	72.24	652
Probability	48.75	20.00	15.00	16.25	68.75	640

Years of Loan Specialist Experience

Table C.7: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	Experience (Below Median)	Experience (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.071^{***}	-0.038^{**}	0.031
	(0.024)	(0.017)	(0.029)
Probability	-0.040	-0.022	0.017
	(0.024)	(0.022)	(0.032)
Binary High (70)	-0.089^{***}	-0.020	0.069^{**}
	(0.024)	(0.019)	(0.030)
Binary Low (30)	-0.057^{***}	-0.022	0.041
	(0.018)	(0.022)	(0.028)
Observations	4,200	3,250	7,450
R^2	0.300	0.343	0.284
Adjusted R^2	0.276	0.317	0.279

Note: Asymptotic standard errors clustered at the individual level in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table C.8: Share of True/False Positives and Negatives by Treatment and Years of Loan Experience

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
Exp (Above Median)						
Human Only	48.62	19.85	17.08	14.46	68.46	650
Binary Med (50)	58.22	18.57	13.15	10.06	76.79	517
Probability	54.08	18.03	13.28	14.61	72.11	527
Exp (Below Median)						
Human Only	48.57	18.81	14.05	18.57	67.38	840
Binary Med (50)	54.48	19.08	13.44	13.01	73.55	692
Probability	48.35	19.76	16.17	15.72	68.11	668

Perceived Residual Information

Table C.9: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	Perceived (Below Median)	Perceived (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.075^{***}	-0.042^{*}	0.036
	(0.019)	(0.024)	(0.030)
Probability	-0.037^{*}	-0.023	0.019
	(0.020)	(0.026)	(0.033)
Binary High (70)	-0.054^{***}	-0.063^{**}	-0.005
	(0.020)	(0.025)	(0.032)
Binary Low (30)	-0.043^{**}	-0.038^{**}	0.010
	(0.020)	(0.019)	(0.027)
Observations	3,850	3,650	7,500
R^2	0.312	0.309	0.282
Adjusted R^2	0.287	0.284	0.277

Note: Asymptotic standard errors clustered at the individual level in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table C.10: Share of True/False Positives and Negatives by Treatment and Perceived Private Info

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
Perceived (Above Median)						
Human Only	48.22	19.73	15.48	16.58	67.95	730
Binary Med (50)	55.14	18.38	14.17	12.31	73.52	593
Probability	51.45	18.40	14.65	15.50	69.85	587
Perceived (Below Median)						
Human Only	48.96	18.70	15.19	17.14	67.66	770
Binary Med (50)	57.05	19.07	12.66	11.22	76.12	624
Probability	50.24	19.51	15.28	14.96	69.76	615

Baseline Practice Accuracy

Table C.11: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	Practice Acc (Below Median)	Practice Acc (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.065^{***}	-0.046^{*}	0.017
	(0.019)	(0.024)	(0.030)
Probability	-0.033	-0.027	0.013
	(0.021)	(0.027)	(0.033)
Binary High (70)	-0.064^{***}	-0.057^{**}	0.014
	(0.020)	(0.026)	(0.032)
Binary Low (30)	-0.055^{***}	-0.015	0.047^{*}
	(0.017)	(0.022)	(0.027)
Observations	5,150	2,350	7,500
R^2	0.299	0.336	0.282
Adjusted R^2	0.277	0.306	0.276
Note:	Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table C.12: Share of True/False Positives and Negatives by Treatment and Practice Accuracy

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
Prac Acc (Below Median)						
Human Only	46.99	20.39	15.73	16.89	67.38	1030
Binary Med (50)	56.73	18.03	12.74	12.50	74.76	832
Probability	50.80	18.60	15.06	15.54	69.40	817
Prac Acc (Above Median)						
Human Only	52.13	16.60	14.47	16.81	68.72	470
Binary Med (50)	54.81	20.26	14.81	10.13	75.06	385
Probability	50.91	19.74	14.81	14.55	70.65	385

Aggregation bias

Table C.13: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	Rubinstein (Below Median)	Rubinstein (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.062^{***}	-0.055^{***}	0.011
	(0.020)	(0.021)	(0.028)
Probability	-0.042^{*}	-0.007	0.028
	(0.022)	(0.021)	(0.031)
Binary High (70)	-0.074^{***}	-0.022	0.044
	(0.020)	(0.027)	(0.032)
Binary Low (30)	-0.060^{***}	-0.005	0.053^{**}
	(0.017)	(0.021)	(0.027)
Observations	5,050	2,400	7,450
R^2	0.302	0.333	0.283
Adjusted R^2	0.279	0.303	0.277
Note:	Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01		

Table C.14: Share of True/False Positives and Negatives by Treatment and Rubinstein

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
Rubinstein (Below Median)						
Human Only	49.90	18.42	16.24	15.45	68.32	1010
Binary Med (50)	55.88	19.88	12.48	11.76	75.76	825
Probability	51.05	19.11	14.92	14.92	70.16	811
Rubinstein (Above Median)						
Human Only	45.83	21.04	13.54	19.58	66.88	480
Binary Med (50)	56.51	16.67	15.10	11.72	73.18	384
Probability	50.52	18.75	14.84	15.89	69.27	384

Balls and Urns (BU)

Table C.15: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	Balls-Urns (Below Median)	Balls-Urns (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.043^{**}	-0.076^{***}	-0.025
	(0.020)	(0.028)	(0.035)
Probability	-0.001	-0.047	-0.036
	(0.021)	(0.031)	(0.036)
Binary High (70)	-0.052^{**}	-0.079^{***}	-0.013
	(0.023)	(0.027)	(0.035)
Binary Low (30)	-0.042^{*}	-0.031	0.015
	(0.022)	(0.022)	(0.031)
Observations	3,100	3,100	6,200
R^2	0.339	0.307	0.289
Adjusted R^2	0.313	0.280	0.282

Note: Asymptotic standard errors clustered at the individual level in parentheses.
*p<0.1; **p<0.05; ***p<0.01

Table C.16: Share of True/False Positives and Negatives by Treatment and Balls Urns

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
BU (Below Median)						
Human Only	47.10	21.45	13.71	17.74	68.55	620
Binary Med (50)	56.24	19.41	12.48	11.88	75.64	505
Probability	52.00	20.00	12.80	15.20	72.00	500
BU (Above Median)						
Human Only	49.84	17.10	16.45	16.61	66.94	620
Binary Med (50)	56.86	17.69	13.52	11.93	74.55	503
Probability	51.75	18.48	15.81	13.96	70.23	487

AI Underweighting (AU)

Table C.17: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	AU (Below Median)	Near-Bayes (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.070^{***}	-0.021	0.045
	(0.013)	(0.042)	(0.047)
Probability	-0.045^{**}	0.022	0.066^{*}
	(0.019)	(0.029)	(0.038)
Binary High (70)	-0.072^{***}	-0.023	0.058
	(0.018)	(0.034)	(0.040)
Binary Low (30)	-0.040^{***}	-0.040	0.008
	(0.015)	(0.033)	(0.037)
Observations	5,650	1,800	7,450
R^2	0.315	0.316	0.283
Adjusted R^2	0.294	0.279	0.277
Note:	Asymptotic standard errors clustered at the individual level in parentheses.		
			*p<0.1; **p<0.05; ***p<0.01

Note: The AI-misweighting (AU) moderator has a mass point at its median; ties are grouped with the below-median column, yielding 113 vs. 36 participants rather than an even split.

Table C.18: Share of True/False Positives and Negatives by Treatment and AI Underweighting (AU)

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
Near-Bayes (Above Median)						
Human Only	50.83	16.67	17.50	15.00	67.50	360
Binary Med (50)	55.56	18.18	14.14	12.12	73.74	297
Probability	52.25	21.80	13.49	12.46	74.05	289
AU (Below Median)						
Human Only	47.88	20.09	14.69	17.35	67.96	1130
Binary Med (50)	56.25	19.08	13.05	11.62	75.33	912
Probability	50.44	18.10	15.34	16.11	68.54	906

Base-rate Neglect (BRN)

Table C.19: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	BRN (Below Median)	BRN (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.054^{***}	-0.089^{***}	-0.030
	(0.019)	(0.029)	(0.037)
Probability	-0.046^{*}	-0.005	0.042
	(0.024)	(0.023)	(0.033)
Binary High (70)	-0.072^{***}	-0.037	0.036
	(0.023)	(0.028)	(0.035)
Binary Low (30)	-0.029	-0.066^{**}	-0.031
	(0.018)	(0.027)	(0.033)
Observations	4,800	1,850	6,650
R^2	0.318	0.293	0.281
Adjusted R^2	0.295	0.255	0.275
Note:	Asymptotic standard errors clustered at the individual level in parentheses.		
			*p<0.1; **p<0.05; ***p<0.01

Table C.20: Share of True/False Positives and Negatives by Treatment and Base Rate Neglect

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
BRN (Below Median)						
Human Only	49.79	18.96	14.58	16.67	68.75	960
Binary Med (50)	55.10	19.77	12.88	12.24	74.87	784
Probability	50.13	17.79	15.84	16.23	67.92	770
BRN (Above Median)						
Human Only	45.14	18.92	18.38	17.57	64.05	370
Binary Med (50)	56.19	18.39	13.04	12.37	74.58	299
Probability	54.39	21.28	11.15	13.18	75.68	296

Correlation Neglect (CN)

Table C.21: Decision-making outcomes under different treatment conditions

	Dependent variable:		
	Share of Decisions Correct		
	CN (Below Median)	CN (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.068^{***}	-0.059^{**}	0.006
	(0.017)	(0.028)	(0.033)
Probability	-0.024	-0.051^{*}	-0.029
	(0.019)	(0.028)	(0.034)
Binary High (70)	-0.062^{***}	-0.051^{*}	0.011
	(0.020)	(0.030)	(0.035)
Binary Low (30)	-0.044^{**}	-0.046^{**}	-0.003
	(0.017)	(0.023)	(0.027)
Observations	4,750	2,100	6,850
R^2	0.315	0.326	0.290
Adjusted R^2	0.292	0.292	0.284
Note:	Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01		

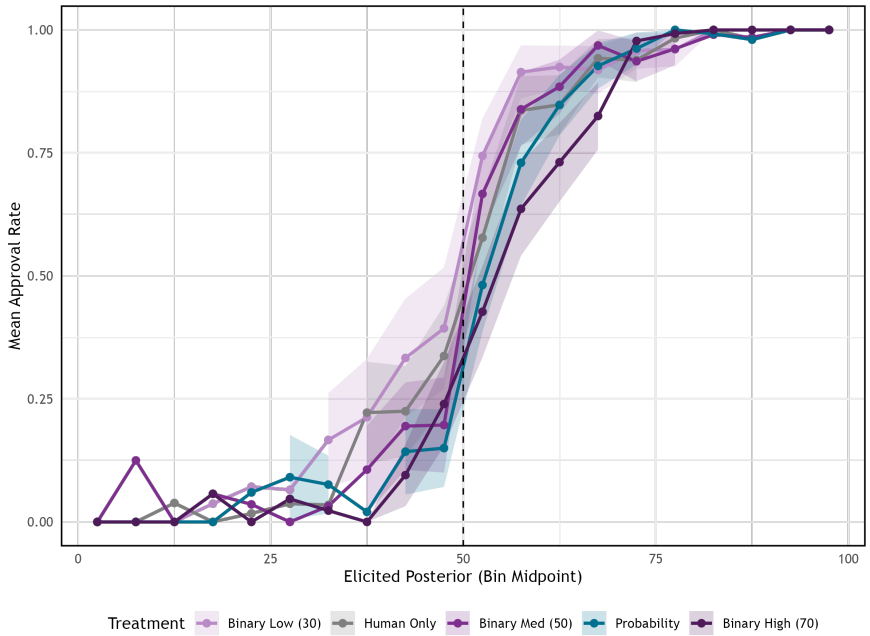
Table C.22: Share of True/False Positives and Negatives by Treatment and Correlation Neglect

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)	Total Count
CN (Below Median)						
Human Only	47.47	19.47	16.11	16.95	66.95	950
Binary Med (50)	55.27	19.15	14.01	11.57	74.42	778
Probability	52.37	18.16	14.21	15.26	70.53	760
CN (Above Median)						
Human Only	51.67	17.86	14.52	15.95	69.52	420
Binary Med (50)	59.76	18.90	11.59	9.76	78.66	328
Probability	47.18	20.47	17.51	14.84	67.66	337

C.2.3 Effective action thresholds

See Figures C.11 to C.12 for the relationship between the elicited posterior belief and the loan decision (by treatment, and by loan specialist respectively).

Figure C.11: Approval Rate vs. Elicited Posterior

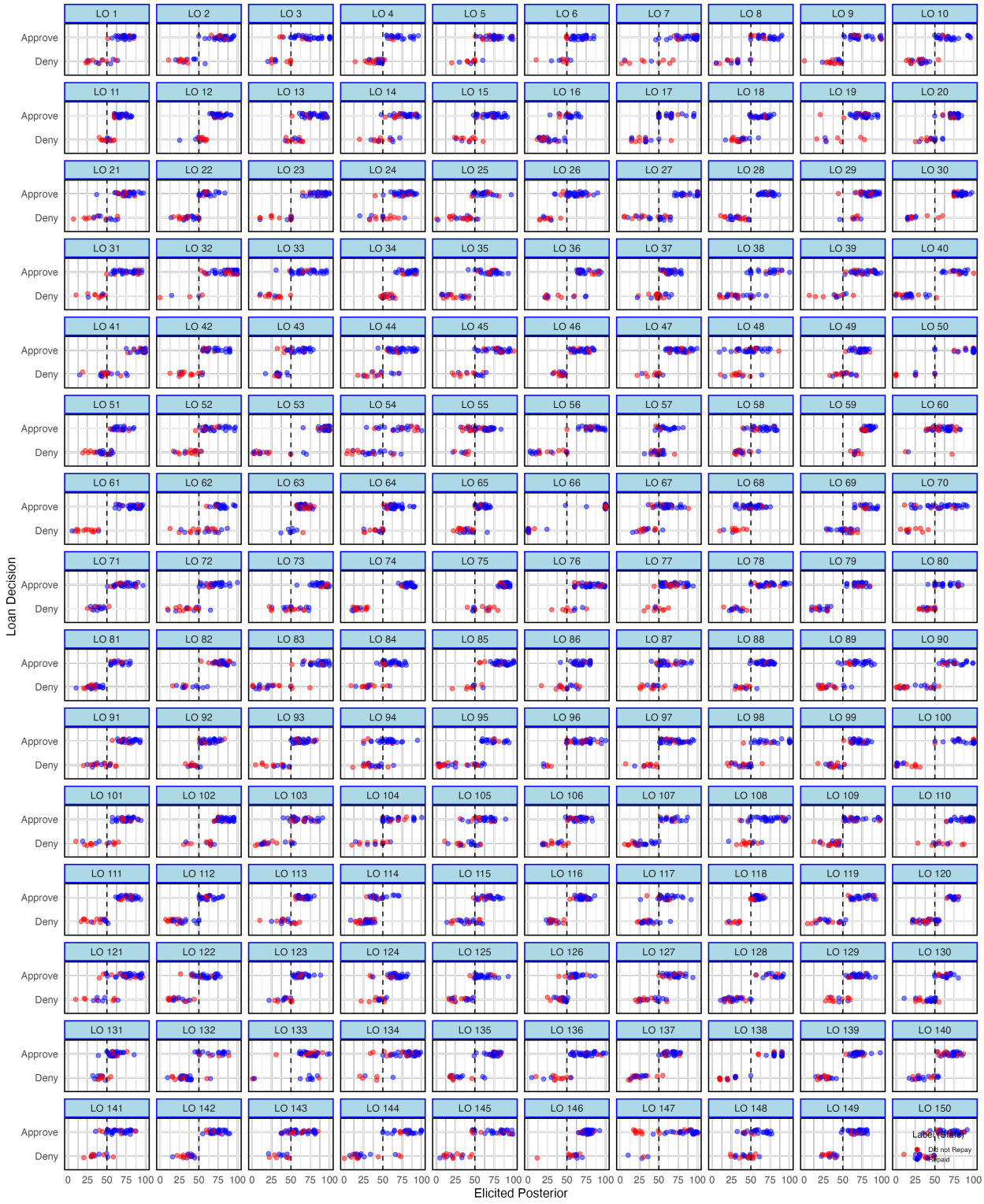


Note: This figure shows the probability of approval as a function of the elicited posterior belief, by treatment. Each line represents a treatment condition, and shaded regions indicate 95% confidence intervals. The dashed vertical line indicates the 50% threshold implied by the incentive structure. Approval rates transition sharply around this threshold across all treatments, suggesting participants apply action thresholds consistent with the experimental incentives.

C.2.4 Algorithmic adherence

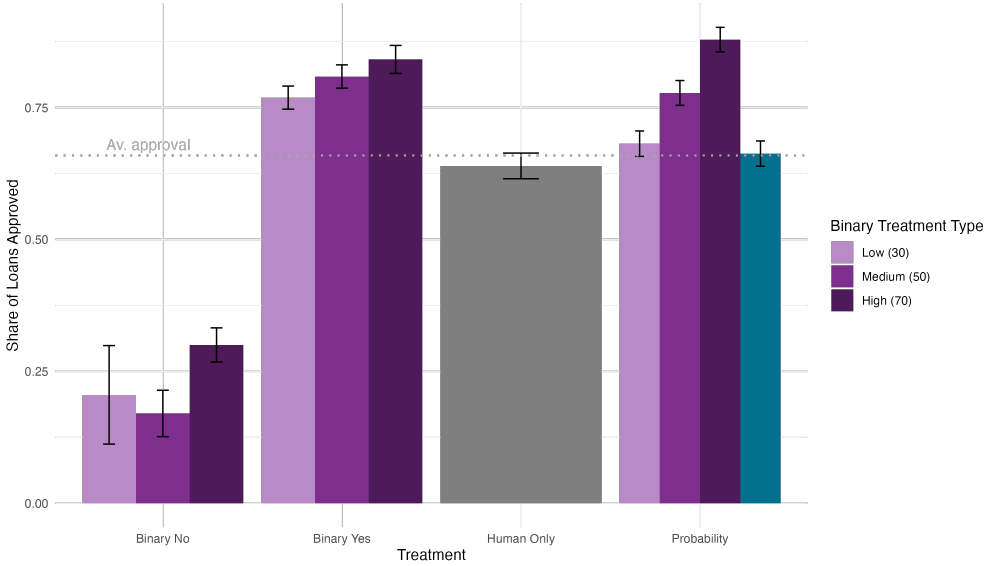
Figure C.13 illustrates share of loans approved across different treatments and thresholds. We would expect a higher threshold—where the AI requires a stronger signal to recommend approval—to lead participants to interpret a "Yes" as a stronger endorsement, increasing approval rates for those cases. It should also make a "No" message weaker, leading to fewer denials. This is evident in the upward-sloping purple bars as we move from lower to higher thresholds. For the *Probability* condition (rightmost cluster), loan approval rates are further broken out by the implied binary recommendation. For instance, under the *Low* threshold, the figure plots the approval rate when the AI score exceeds 30. The teal bar represents the average loan approval rate across the full *Probability* condition. These findings confirm that participants respond to AI-calibrated thresholds but do not blindly follow AI recommendations (since approval rates for when the binary AI recommends "No") and when the binary AI recommends "Yes" are not 0% and 100%, respectively).

Figure C.12: Loan Specialists' Effective Action Thresholds



Note: This figure shows the relationship between the elicited posterior belief and the loan decision (approve/deny) for each loan specialist. The blue dots represent applications that were repaid, and the red dots represent those that were not repaid. For most participants, decisions are clearly separated around the 50% posterior threshold, indicating that loan specialists generally follow the decision rule implied by the incentive structure.

Figure C.13: Loan Approval Rates, by Treatment and (implied) AI Recommendation



Note: Bars show loan approval rates by treatment. *Binary* conditions are split by whether the AI recommends "Yes" or "No" at thresholds of 30%, 50%, or 70%. *Probability* is split by whether the AI score exceeds each threshold; the teal bar shows the overall average. Error bars indicate 95% confidence intervals.

C.2.5 Where loan specialists are performing well

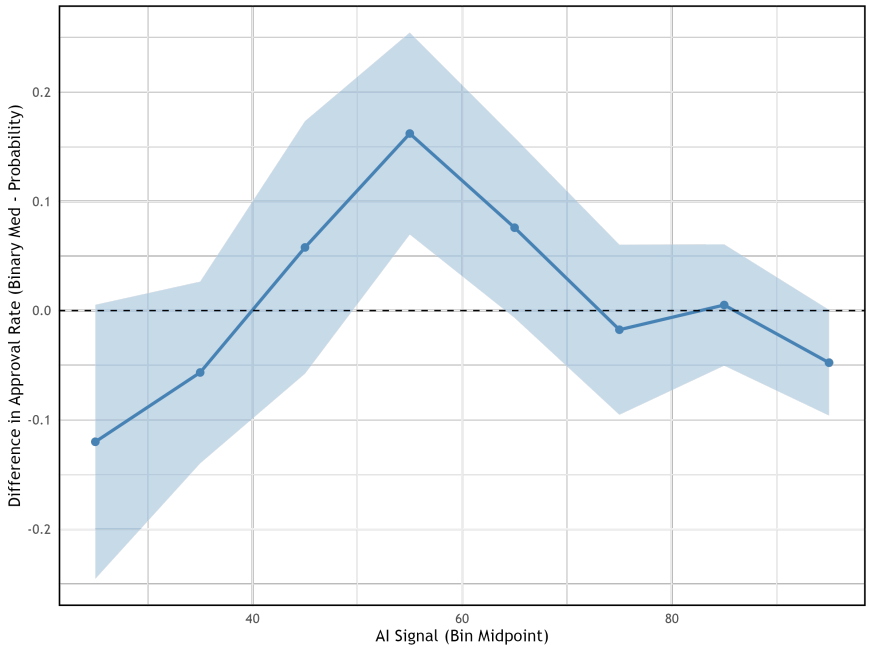
Table C.23-C.25 present the shares of true/false positives and negatives across treatments. These tables condition on rounds with an elicited posterior r: *Human-Only* elicits r every round, while the treatment arms elicit it in roughly 81% of rounds, which is why their counts and overall accuracy differ slightly from the full-sample figures. These results are broken down by the main treatment arms *Binary Medium*, *Human-only*, and *Probability*, and further categorized by buckets of AI signals s and human-elicited signals r (where available), offering insight into how treatments influence decision-making accuracy across different levels of signal strength.

Table C.23: Share of True/False Positives and Negatives by Treatment

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)
Binary Med (50)	56.12	18.73	13.39	11.75	74.86
Human Only	48.60	19.20	15.33	16.87	67.80
Probability	50.83	18.97	14.98	15.22	69.80

As seen in Table C.23, regardless of treatment, loan specialists are much better at approving

Figure C.14: Difference in Approval Probability: Binary Medium vs. Probability



Note: This figure plots the difference in approval probability between the *Binary Medium* and *Probability* treatments as a function of the underlying AI score. Positive values indicate higher approval rates under binary coarsening; negative values indicate higher approval rates under full probability. Shaded regions indicate 95% confidence intervals. The pattern shows that participants under binary coarsening are more likely to approve when the AI score is above 50 (following the "Yes" recommendation) and more likely to deny when below 50 (following the "No" recommendation), particularly for marginal AI scores (50–70).

correctly[73] than denying correctly,[74]. This aligns with anecdotal evidence of real-life loan specialists' behavior: they are often trained to focus on "cherries" (strong applicants), approving applicants based on clear positive indicators, such as meeting thresholds of creditworthiness (e.g., credit scores above a certain range or a debt-to-income ratio within acceptable limits), rather than identifying "lemons" (poor applicants). The coarsened treatment appears to improve decision-making primarily by reducing the share of false negatives and increasing true positives.

Breaking this down further by the human signal inferred from observing the loan application (where elicited), Table C.24 shows that loan specialists perform worst in the implied "Deny" region (own signal below 50), and perform best when their own signal is high (71-100). The AI treatments do not appear to have consistently positive effects on accuracy when the human signal is low (0-30). The *Binary Medium* treatment offers the greatest improvement when the human signal falls in

[73] True positive rate = TP/(TP+FN), or the sensitivity
[74] True negative rate = TN/(TN+FP), or the specificity

Table C.24: Share of True/False Positives and Negatives by r Bucket and Treatment

r	Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)
0-30	Binary Med (50)	8.87	46.77	1.61	42.74	55.65
0-30	Human Only	0.91	58.18	0.91	40.00	59.09
0-30	Probability	9.03	52.78	4.17	34.03	61.81
31-50	Binary Med (50)	29.84	35.66	10.08	24.42	65.50
31-50	Human Only	19.64	36.56	7.25	36.56	56.19
31-50	Probability	27.04	30.37	9.63	32.96	57.41
51-70	Binary Med (50)	65.75	12.56	16.67	5.02	78.31
51-70	Human Only	63.29	7.71	21.30	7.71	70.99
51-70	Probability	59.17	12.92	20.41	7.49	72.09
71-100	Binary Med (50)	77.33	5.79	15.62	1.26	83.12
71-100	Human Only	76.75	0.22	21.71	1.32	76.97
71-100	Probability	73.82	4.99	17.21	3.99	78.80

the uncertain middle range (31-70).

We can alternatively break this out by the underlying AI risk score of each loan case, as seen in Table C.25. Interestingly, when the AI signal is low, across all treatments there are no true positives or false negatives—in other words, the AI is very good at picking out lemons. This suggests that loan specialists would unambiguously benefit from never approving loans with sufficiently low AI risk scores.[75] The *Binary Medium* treatment shows the greatest improvement over *Probability* when the AI score falls in the ambiguous middle range (31-70), largely by converting false negatives into true positives: in the 51-70 bucket, the pooled "Yes" message cuts false negatives from 27.1% to 18.2% of decisions while raising true positives from 49.5% to 56.6%. Providing a coarsened signal implying a clear action in these cases is particularly helpful, as the full probability risk score may deliver a 'weaker' signal, making loan specialists less likely to align their decisions with the implied action.

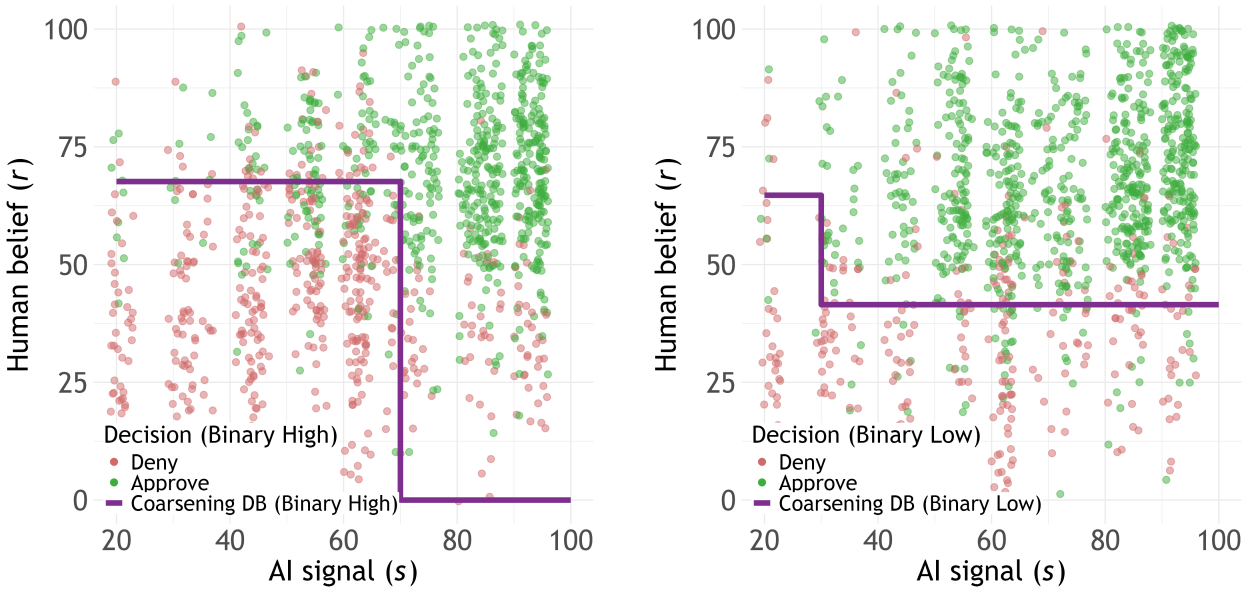
[75] Note, of course, that they don't actually see this signal in the *Human-only* condition, and only a coarsened version of the signal in the *Binary Medium*.

Table C.25: Share of True/False Positives and Negatives by s Bucket and Treatment

s	Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)
0-30	Binary Med (50)	0.00	90.00	10.00	0.00	90.00
0-30	Human Only	0.00	79.61	20.39	0.00	79.61
0-30	Probability	0.00	85.19	14.81	0.00	85.19
31-50	Binary Med (50)	1.27	70.70	18.47	9.55	71.97
31-50	Human Only	0.00	57.65	33.16	9.18	57.65
31-50	Probability	0.62	64.81	19.14	15.43	65.43
51-70	Binary Med (50)	56.64	10.03	15.18	18.16	66.67
51-70	Human Only	52.46	12.50	11.27	23.77	64.96
51-70	Probability	49.48	10.82	12.63	27.06	60.31
71-100	Binary Med (50)	76.01	2.74	11.43	9.82	78.74
71-100	Human Only	66.34	4.49	12.48	16.69	70.83
71-100	Probability	73.20	2.10	15.41	9.28	75.31

C.2.6 Decision Boundaries, Estimated from Experiment Data

Figure C.15: Decision Boundaries, Estimated from Experiment Data (Pooled)



(a) Coarsening DB (*Binary High*, Pooled)

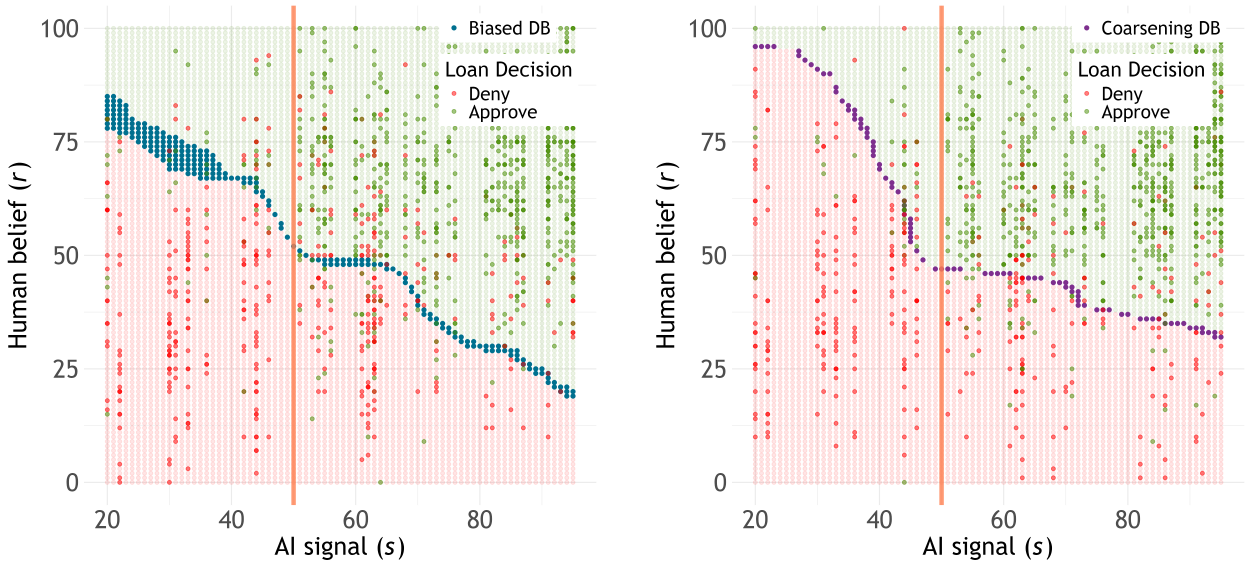
(b) Coarsening DB (*Binary Low*, Pooled)

Note: All panels plot the AI signal s (Sender's type) against the loan specialist's private signal r (Receiver's type) expressed as 0–100 repayment probability. Purple lines plot the posterior median of the HLM-estimated distribution of loan-specialist decision boundaries, with points showing all specialist decisions under the relevant treatment. Panel (a) plots *Binary High*, and Panel (b) plots *Binary Low*.

As a non-parametric check on the hierarchical estimates, we re-estimate the full-revelation decision boundary directly from realized choices. Pooling all *Probability* decisions for which the human signal r was elicited, we estimate the approval probability $\Pr(a = 1 \mid r, s)$ by a Nadaraya–Watson kernel regression (bandwidths chosen by cross-validation), monotonize it in r and s via the rearrangement of Chernozhukov et al. (2009), and trace the $\Pr(a = 1 \mid r, s) = 0.5$ contour (Figure C.16a). The resulting boundary closely matches the HLM estimate in the main text (Figure VIb). Repeating the exercise under *Binary Medium* (Figure C.16b) shows that thresholding the AI signal at $s = 50$ pulls decisions toward the Bayesian benchmark: within $r \in [35, 75]$, specialists on average adhere to the binary recommendation.[76]

[76]The s-axis is retained only for comparability with panel (a): under *Binary Medium* specialists observe only the binary message $\mathbb{1}[s \geq 50]$, not the continuous score, so any within-bin variation of the contour in s reflects the composition of cases in each bin rather than a behavioral response to the hidden score.

Figure C.16: Non-Parametric Decision Boundaries



(a) Loan Specialist Decision Boundary (Probability Treatment, Pooled)

(b) Loan Specialist Decision Boundary (Binary Medium Treatment, Pooled)

Note: Non-parametric analogues of the specialist decision boundaries estimated by the hierarchical model in the main text (the teal boundary in Figure VIb and the step boundary in Figure VIc). Both panels plot the AI signal s (Sender's type, x-axis) against the specialist's human signal r (Receiver's type, y-axis). Each panel is estimated *only* from its own treatment's decisions for which r was elicited—panel (a) from *Probability*, panel (b) from *Binary Medium* — pooled across specialists; the points are those individual decisions, colored by action (approve / deny). The curve is the $\Pr(a = 1 \mid r, s) = 0.5$ contour of a Nadaraya–Watson kernel estimate of the approval probability (bandwidths by cross-validation, monotonized in (r, s) via the rearrangement of Chernozhukov et al., 2009).

C.3 Robustness checks - Main Experiment

C.3.1 Restricting to participants recruited through industry channels

We conduct a robustness check on the inclusion of Prolific participants who were subject to stricter screening but recruited through a general-purpose survey platform. While Prolific offers flexibility and cost advantages, it differs from targeted industry outreach. Our pilots found that Prolific participants with Accounting and Business backgrounds (common among loan specialists) was ineffective due to differences in industry-specific knowledge and decision-making. Only after implementing rigorous screening did we identify Prolific participants resembling loan specialists, highlighting the need for targeted recruitment to ensure domain expertise.

Table C.26: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Human-Only	-0.059^{***}	-0.055^{***}
	(0.018)	(0.015)
Probability	-0.040^{*}	-0.016
	(0.021)	(0.018)
Binary High (70)	-0.069^{***}	-0.039^{**}
	(0.020)	(0.018)
Binary Low (30)	-0.042^{**}	-0.019
	(0.020)	(0.017)
Observations	4,800	4,800
R^2	0.003	0.323
Adjusted R^2	0.002	0.300

Note: Column (1) reports estimates without fixed effects; Column (2) includes individual, case, and block order fixed effects. Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table C.26 presents the results shown in Table II, restricting data to those recruited over LinkedIn (NAMU/NAMP inclusive) and Indeed, representing around 65% of our full sample. Our main results still hold: *Binary Medium* performs better than both *Probability* and *Human-only*. These differences are significant for Binary Medium vs. *Human-only*, but due to the smaller sample size are not significant when compared to *Probability*.

C.3.2 Whether AI or human signal came first, elicitation of first signal

Recall that we implemented case-level randomization of (i) **order of information presentation:** whether the AI signal or loan application shown first {AI (10%), Loan (90%)}, and (ii) **elicitation of posterior before the second signal:** whether posterior is additionally elicited before the second signal shown {Elicit after only (10%), Elicit both before and after (90%)}.

We conduct robustness checks regressing (i) final loan decision or (ii) final posterior beliefs on these; as shown in Table C.27, neither the order of the AI signal nor the elicitation of posterior beliefs significantly influences the outcomes.

Table C.27: Robustness Checks

	Dependent variable:			
	Loan Decision	Posterior Belief	Loan Decision	Posterior Belief
	(1)	(2)	(3)	(4)
First Signal	−0.021	−1.035		
	(0.019)	(0.978)		
Elicitation Order			0.002	−0.357
			(0.018)	(0.838)
Constant	0.678***	60.465***	0.657***	59.847***
	(0.020)	(1.104)	(0.020)	(1.060)
Observations	7,500	7,500	7,500	7,500
R^2	0.0002	0.0002	0.00000	0.00002
Adjusted R^2	0.00003	0.00004	-0.0001	-0.0001

Note: Regression results for the final loan decision (Columns 1, 3) and posterior belief (Columns 2, 4), on whether or not the AI signal was shown first (Columns 1, 2) and whether posterior beliefs are elicited after the second signal (Columns 3, 4). Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

C.3.3 Restricting to each decision block

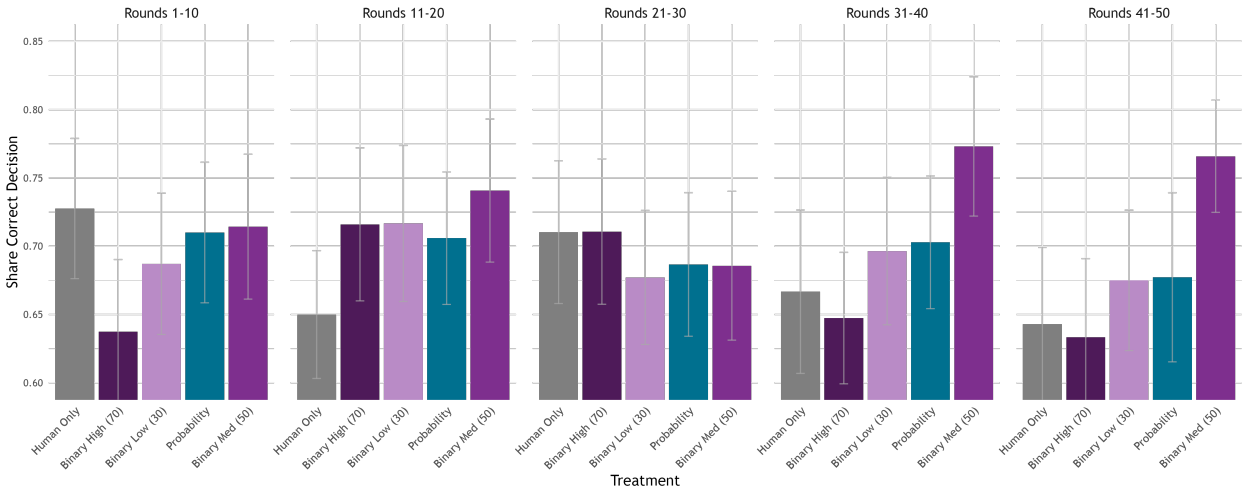
Figure C.17 presents the results shown in Figure IV, breaking down the results by each decision block. *Binary Medium* outperforms *Probability* in nearly every decision block—most importantly the first block, where the comparison is cleanest (before any within-participant carryover across blocks).[77], indicating that its advantages are not driven by block order. Two patterns are particularly noteworthy, though it is important to note these are not statistically significant (we did not plan to be powered for these differences). First, the highest *Human-only* accuracy appears in the first decision block, suggesting that practice on the loan applications—without AI assistance—has an immediate positive effect that drops after seeing a round of AI signals. Second, accuracy under the *Binary Medium* condition improves in later rounds, which may indicate that participants are learning to interpret and apply the signal more effectively over time.

C.4 Decision Model: Specification and Estimation

This appendix provides the full specification and estimation details for the decision model summarized in Section 5.2.2, together with the out-of-sample evaluation used to personalize policies.

[77] Note the two are roughly tied in Rounds 21–30.

Figure C.17: Decision-making outcomes under different treatment conditions, by decision block



Note: Bars represent the share of correct decisions under each treatment condition. Error bars represent 95% confidence intervals.

C.4.1 Decision model specification

We model loan approval using a latent probit model. Each loan application n is evaluated by a loan specialist i under treatment condition T, where $T = \{t_1, ..., t_{K-1}\}$ is the set of thresholds that coarsen the space under a *K*-coarsening. Let $s_n \in S \subseteq \mathbb{R}$ be the underlying continuous *AI signal* (from the AI output) for application n. We define the *treatment-specific* AI message $\Phi_T(s_n)$ as:

$$\Phi_T(s_n) = \begin{cases} 1, & \text{if } (|T| = 1 \text{ and } s_n \geq t_1) \text{ or } (|T| = 2 \text{ and } s_n \geq t_2), \\ \mu_M(t), & \text{if } |T| = 2 \text{ and } t_1 \leq s_n < t_2, \\ 0, & \text{if } (|T| = 1 \text{ and } s_n < t_1) \text{ or } (|T| = 2 \text{ and } s_n < t_1) \text{ or } |T| = 0, \\ s_n, & \text{if } |T| = \infty \text{ (full revelation).} \end{cases}$$

In other words, when the treatment is *binary* (i.e., $|T| = 1$), the AI score is coarsened to $\{0, 1\}$ depending on whether s_n falls below or above the threshold t_1. When the treatment is *trinary* (i.e., $|T| = 2$), the two thresholds (t_1, t_2) demarcate the score distribution into three bands. The low and high bands are anchored at 0 and 1, while the middle band is mapped to a single constant

$\mu_M(t) = \frac{\mathrm{E}[\,s_n | t_1 \le s_n < t_2\,] - \mathrm{E}[\,s_n | s_n < t_1\,]}{\mathrm{E}[\,s_n | s_n \ge t_2\,] - \mathrm{E}[\,s_n | s_n < t_1\,]}$, so that all three support points lie on $[0,1]$.[78] If the treatment is *human-only* (i.e., $|T| = 0$), we set $\Phi_T(s_n) = 0$ for all n. Finally, under *full revelation* (the fine limit $|T| \to \infty$, i.e., no coarsening), the continuous score is directly revealed by setting $\Phi_T(s_n) = s_n$.

For each application decision, the loan specialist i observes their own human signal $r_{i,T,n} \in R \subseteq \mathbb{R}$ and the coarsened AI signal $\Phi_T(s_n)$ as defined above, for all loan applications n assigned to them under treatment T. Based on this information, the specialist makes an approval decision $a_{i,T,n} \in \{0,1\}$. We assume that the binary decision $a_{i,T,n}$ arises from a continuous variable $y^*_{i,T,n}$ which is not directly observed, i.e., that $a_{i,T,n} = \mathbb{1}[y^*_{i,T,n} \ge 0]$. The variable $y^*_{i,T,n}$ is given by:

$$y^*_{i,T,n} = \alpha_{i,T} + r_{i,T,n} + \beta_{i,T} \cdot \Phi_T(s_n) + \epsilon_{i,T,n} \qquad \epsilon_{i,T,n} \sim \mathcal{N}(0, \sigma^2_{\epsilon,T})$$

$$\alpha_{i,T} \sim \mathcal{N}(Z_i \cdot \gamma_{\alpha,T}, \sigma^2_{\alpha,T}) \quad \text{and} \quad \beta_{i,T} \sim \mathcal{N}_{(0,\infty)}(Z_i \cdot \gamma_{s,T}, \sigma^2_{s,T})$$

where $\alpha_{i,T}$ is a treatment-specific intercept and $\beta_{i,T}$ is the weight assigned to the AI signal. To accommodate specialist-level heterogeneity without overfitting, we employ a hierarchical approach, allowing specialist-level parameters $(\alpha_{i,T}, \beta_{i,T})$ to vary while being informed by population-level distributions. Specifically, each set of specialist-level effects is drawn from a normal distribution whose mean depends on specialist-specific covariates Z_i, where $\gamma_{\alpha,T}$ and $\gamma_{s,T}$ are the population-level coefficient vectors, and $\sigma^2_{\alpha,T}$ and $\sigma^2_{s,T}$ represent residual heterogeneity in intercepts and slopes, respectively. This structure allows specialists with similar characteristics to have similar coefficients while permitting idiosyncratic variation. Note that the specification restricts the weight on the own signal $r_{i,T,n}$ and the decision-noise variance $\sigma^2_{\epsilon,T}$ to be common across specialists within a treatment: individual heterogeneity enters only through the intercepts $\alpha_{i,T}$ and the AI-signal weights $\beta_{i,T}$, so the personalization exercise is driven by differences in responsiveness to the AI message. We assume that both $\alpha_{i,T}, \beta_{i,T}$ are independent of each other and of r,[79] and normalize the coefficient on the human signal to fix scale.[80] The term $\epsilon_{i,T,n}$ captures residual idiosyncratic

[78] Since the latent index is linear in $\Phi_T(s)$, the low-to-middle jump ($0 \to \mu_M(t)$) shifts latent approval by $\beta_{i,T}\,\mu_M(t)$ and the middle-to-high jump ($\mu_M(t) \to 1$) by $\beta_{i,T}\,(1 - \mu_M(t))$. This normalization places the middle band at the relative position of its conditional-mean AI score between the low and high bands, so all three support points lie on $[0,1]$.

[79] The independence assumption is reasonable given that there is no *a priori* reason to expect why someone with higher baseline approval (α) should be more/less sensitive to the AI score (β). In practice, omitting correlation between the coefficients leads to more stable estimates, as there is not enough data to reliably estimate the correlation parameters. In settings with more data, one could consider relaxing this assumption.

[80] To ensure stable hierarchical estimation, in practice we fix the coefficient on the human signal to $1/\sigma_\epsilon$ and work on

variation in decision-making not explained by observed signals, where $\sigma^2_{\epsilon,T}$ represents residual heterogeneity at the decision-level within treatment T.

To prevent illogical decision boundaries (e.g., where some loan specialists may Deny under a "Yes" message but Approve under a "No" message), we impose the constraint $\beta_{i,T} > 0$ for all specialists and treatment conditions, ensuring that higher AI signals are weakly associated with a higher likelihood of loan approval.

C.4.2 Decision model estimation

As we wish to retain our flexible conceptual framework and permit different updating procedures under varying information policies, we estimate the model separately for each treatment condition T, allowing the coefficients $\alpha_{i,T}$ and $\beta_{i,T}$ to vary across both specialists and treatments. As such, we recover one set of decision boundary parameters per specialist-treatment.

Each loan specialist makes 10 loan application decisions in each treatment condition: for our main experiment, this includes $T \in \{30\}, \{50\}, \{70\}, \emptyset$, and $S \setminus \{\min(S)\}$. The econometrician observes $D = \{a_{i,T,n}, \mathbb{1}^{\text{obs}}_{i,T,n}, \hat{r}_{i,T,n}, s_n, \Phi_T(s_n)\}$, where $\hat{r}_{i,T,n}$ is the loan specialists' human signal *when elicited* and is only observed when $\mathbb{1}^{\text{obs}}_{i,T,n}$ is equal to 1. Since this missingness is completely at random,[81] we treat each unobserved signal $\tilde{r}_{i,T,n}$ as a latent variable:

$$r_{i,T,n} = \mathbb{1}^{\text{obs}}_{i,T,n}\, \hat{r}_{i,T,n} + \big(1 - \mathbb{1}^{\text{obs}}_{i,T,n}\big)\, \tilde{r}_{i,T,n}, \qquad \tilde{r}_{i,T,n} \sim \mathcal{N}\big(Z_i^\top \delta + X_n^\top \lambda,\, \sigma^2_{r,T}\big),$$

where $\tilde{r}_{i,T,n}$ is modeled as a function of observable application characteristics X_n (e.g., loan amount, annuity) and a vector of specialist-specific covariates Z_i (e.g., demographics, baseline biases). The residual idiosyncratic variation is governed by the population-level variance parameter $\sigma^2_{r,T}$, which is allowed to vary across treatment conditions.

From the latent probit model, we can write the conditional probability that loan n is approved—

a unit-variance probit scale. Dividing the structural equation by $\sigma_{\epsilon,T}$ gives

$$\tilde{y}^*_{i,T,n} \;=\; \frac{y^*_{i,T,n}}{\sigma_{\epsilon,T}} \;=\; \tilde{\alpha}_i + \tilde{\beta}_{r,i,T}\, r_{i,T,n} + \tilde{\beta}_{s,i,T}\, \Phi_T(s_n) + \varepsilon^{\text{std}}_{i,T,n}, \qquad \varepsilon^{\text{std}}_{i,T,n} \sim \mathcal{N}(0,1),$$

where $\tilde{\alpha}_i = \alpha_i/\sigma_{\epsilon,T}$, $\tilde{\beta}_{r,i,T} = \frac{1}{\sigma_{\epsilon,T}}$, $\tilde{\beta}_{s,i,T} = \frac{\beta_{i,T}}{\sigma_{\epsilon,T}}$. This allows us to sample the rescaled coefficients $\tilde{\alpha}_i$ and $\tilde{\beta}_{s,i,T}$ hierarchically and interpret them directly, while $\sigma_{\epsilon,T}$ captures the residual decision noise.

[81] For the *Human-only* condition ($T = \emptyset$), the econometrician observes $\hat{r}_{i,T,n}$ in every round; in the AI-assisted rounds, it is observed completely at random for 81% of observations—that is, the probability that $\hat{r}_{i,T,n}$ is unobserved does not depend on any observed or unobserved variables.

the decision likelihood—given the model parameters $\Theta = \{\alpha_{i,T},\ \beta_{i,T},\ \lambda,\ \delta,\ \sigma^2_{\epsilon,T},\ \sigma^2_{r,T}\}$:

$$p(a_{i,T,n} \mid \mathbb{1}^{\text{obs}}_{i,T,n},\ \hat{r}_{i,T,n},\ X_n,\ Z_i,\ \Phi_T(s_n),\ \Theta)$$
$$= \text{Bernoulli}\left(a_{i,T,n} \ \middle|\ \Phi\Big(\frac{\alpha_{i,T}+\mathbb{1}^{\text{obs}}_{i,T,n}\,\hat{r}_{i,T,n}+(1-\mathbb{1}^{\text{obs}}_{i,T,n})\,(Z_i^\top\delta+X_n^\top\lambda)+\beta_{i,T}\,\Phi_T(s_n)}{\sqrt{\sigma^2_{\epsilon,T}+(1-\mathbb{1}^{\text{obs}}_{i,T,n})\,\sigma^2_{r,T}}}\Big)\right)$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function. This reflects the full predictive structure: when the human signal $\hat{r}_{i,T,n}$ is observed, it enters the model directly; when unobserved, the model marginalizes over the latent signal $\tilde{r}_{i,T,n}$. This marginal form and the joint-posterior formulation below are equivalent representations—integrating the probit likelihood over the Gaussian latent signal yields exactly the inflated-variance expression above; in estimation we implement the latent formulation, sampling the missing $\tilde{r}_{i,T,n}$ jointly with Θ via HMC. Note that whenever the human signal is observed ($\mathbb{1}^{\text{obs}}_{i,T,n} = 1$) we also explicitly model $\hat{r}_{i,T,n} \sim \mathcal{N}\big(Z_i^\top\delta + X_n^\top\lambda,\ \sigma^2_{r,T}\big)$ as part of the joint likelihood, via the *latent signal model*:

$$p\big(r_{i,T,n} \mid \mathbb{1}^{\text{obs}}_{i,T,n}, \hat{r}_{i,T,n}, X_n, Z_i, \Theta\big) = \begin{cases} \mathcal{N}\Big(\hat{r}_{i,T,n} \ \Big|\ Z_i^\top\delta + X_n^\top\lambda,\ \sigma^2_{r,T}\Big), & \text{if } \mathbb{1}^{\text{obs}}_{i,T,n} = 1, \\ \mathcal{N}\Big(\tilde{r}_{i,T,n} \ \Big|\ Z_i^\top\delta + X_n^\top\lambda,\ \sigma^2_{r,T}\Big), & \text{if } \mathbb{1}^{\text{obs}}_{i,T,n} = 0, \end{cases}$$

ensuring that the variance $\sigma^2_{r,T}$ is directly identified from the dispersion of the observed $\hat{r}_{i,T,n}$ values around its regression mean. Once $\sigma^2_{r,T}$ is pinned down, the approval likelihood attributes the remaining unexplained variation to the decision-noise variance $\sigma^2_{\epsilon,T}$, allowing the two variance components to be separately identified.

We estimate the model in Stan using Hamiltonian Monte Carlo (HMC) to jointly sample from the posterior distribution of all parameters. Given that r is missing completely at random (MCAR), we sample all model parameters Θ and latent signals $\tilde{r}_{i,T,n}$ from the joint posterior:

$$p(\Theta, \tilde{r}_{i,T,n} \mid \mathbb{1}^{\text{obs}}_{i,T,n}, \hat{r}_{i,T,n},\ a_{i,T,n},\ X_n,\ Z_i,\ \Phi_T(s_n)) \propto$$
$$\underbrace{p(a_{i,T,n} \mid r_{i,T,n}, \Phi_T(s_n), \Theta)}_{\text{decision likelihood}}\ \underbrace{p(r_{i,T,n} \mid \mathbb{1}^{\text{obs}}_{i,T,n}, \hat{r}_{i,T,n}, X_n, Z_i, \Theta)}_{\text{latent signal model}}\ \underbrace{p(\Theta)}_{\text{prior}},$$

where the first term represents the likelihood of the observed decision given the (partially latent) human and AI signals, as laid out in the conditional likelihood above; the second is the signal model for $r_{i,T,n}$; and the third encodes priors over all parameters. No adjustment to the likelihood

is needed as the missingness mechanism is ignorable under MCAR (Gelman et al., 1995).

Lastly, we use weakly informative priors to regularize estimation while allowing the data to drive inference. Specifically, we use $\mathcal{N}(0, 5)$ priors for the group-level hyperparameters $\gamma_{\alpha,T}$ and $\gamma_{s,T}$, and assign half-Cauchy (0, 2.5) priors for the standard deviations $\sigma_{\alpha,T}$, $\sigma_{s,T}$ and σ_ϵ following Gelman's recommendation for weak regularization in probit models. We also place data-driven priors on the human-signal coefficients. Using only observations where $\hat{r}$ is reported, we regress $\hat{r}$ on the specialist and loan covariates Z_i, X_n; the resulting point estimates $(\hat{\delta}, \hat{\lambda})$ and squared standard errors $(\hat{\sigma}_\delta^2, \hat{\sigma}_\lambda^2)$ become the means and variances of our priors: $\delta \sim \mathcal{N}\big(\hat{\delta}, \hat{\sigma}_\delta^2\big)$ and $\lambda \sim \mathcal{N}\big(\hat{\lambda}, \hat{\sigma}_\lambda^2\big)$. The residual signal variance is given a truncated normal prior $\sigma_r \sim \mathcal{N}\big(\hat{\sigma}_r, \hat{\sigma}_{r,se}^2\big)\, \mathbb{1}_{(0,\infty)}$.

C.4.3 Estimating individual specialist decision boundaries

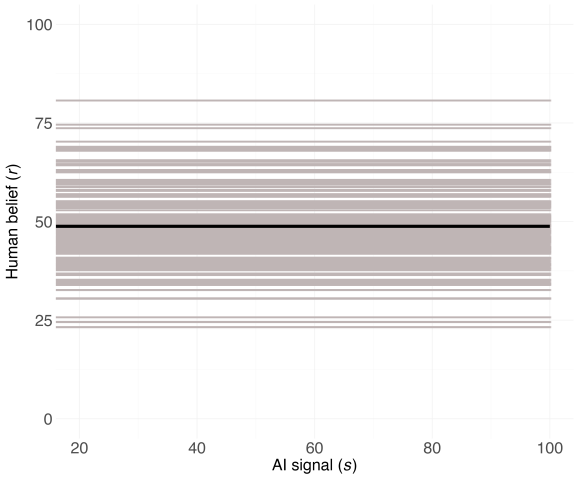
Posterior draws from our hierarchical model allow us to infer each specialist's decision boundary under different treatment conditions. Each specialist's decision boundary is defined as the set of points (r, s) at which their latent utility $y_i^* = 0$. We can express the decision boundary for each treatment as follows:

- **Probability:** $\alpha_{i,T} + r + \beta_{i,T} \cdot s = 0 \implies s = -\frac{\alpha_{i,T}+r}{\beta_{i,T}}$ for $t = S \setminus \{\min(S)\}$ and $\beta_{i,T} \neq 0$ (this is necessarily the case since we impose $\beta_{i,T} > 0$ in our estimation).
- **Human-only Treatment:** $r_0 = -\alpha_{i,T}$ for $t = \emptyset$.
- **Binary Treatments:** $r_1 = -\alpha_{i,T} - \beta_{i,T}, \quad r_0 = -\alpha_{i,T}$ for $t \in \{30\}, \{50\}, \{70\}$, where the actions of the Receivers for each binary treatment are characterized by $\{t_1, r_0, r_1\}$, i.e., the binary threshold and region where the specialist would adhere to the AI recommendation.
- **Trinary Treatments:** $r_2 = -\alpha_{i,T} - \beta_{i,T}, \quad r_1 = -\alpha_{i,T} - \beta_{i,T} \cdot \mu_M(t), \quad r_0 = -\alpha_{i,T}$ for $t \in \{30, 50\}, \{50, 70\}$, where the actions of the Receivers for each trinary treatment are characterized by $\{t_1, t_2, r_0, r_1, r_2\}$, i.e., the two thresholds and the minimum human signals needed to "Approve" under each induced message.

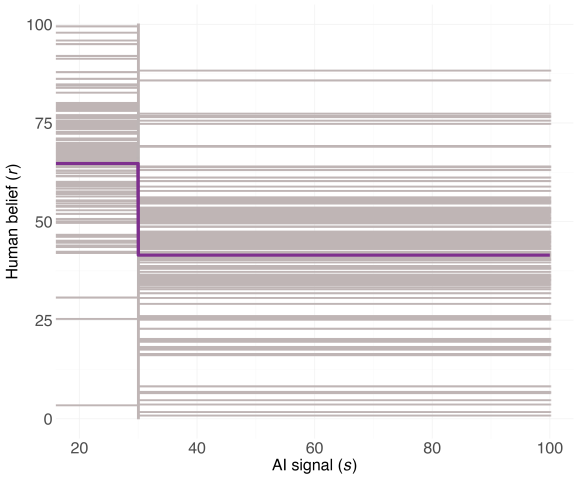
C.4.4 Out-of-sample policy evaluation

Using the specialist-level parameters estimated above, we evaluate how each loan specialist would perform under each treatment condition on a shared set of $N = 1{,}000$ out-of-sample loan applications, maintaining the same population prior.

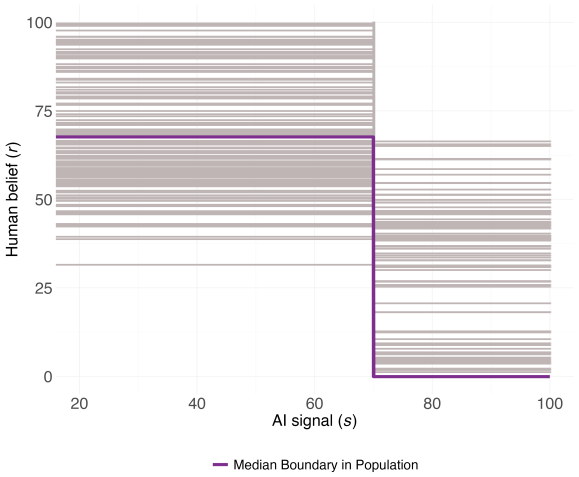
Figure C.18: Heterogeneity in decision-boundaries under remaining treatments



(a) Human-Only



(b) Binary Low



(c) Binary High

Note: Estimated median decision boundary for individual specialists under the *Human-only*, *Binary Low*, and *Binary High* treatments (see Figure IX for *Probability* and *Binary Medium*), based on posterior samples from our hierarchical model. Grey lines represent individual decision boundaries (median across posterior samples). The solid black line shows the median boundary of each relevant point of the adherence curve (e.g., $r_0(t), r_1(t)$) across specialists.

First, for each specialist-treatment pair, we draw S posterior samples of the decision parameters $\alpha_{i,T}^{(s)}$, $\beta_{i,T}^{(s)}$, $\lambda_t^{(s)}$, and $\sigma_{i,T}^{2\,(s)}$, and use these to compute predicted decisions $\hat{a}_{i,T,n}^{(s)}$ for each out-of-sample loan application n and posterior sample s.

Specifically, for each posterior sample and loan application, we draw the human signal: $r_{i,T,n}^{(s)} \sim \mathcal{N}(X_n\lambda_t^{(s)} + Z_i\delta_t^{(s)}, \sigma_{r,T}^{2\,(s)})$ and then compute the predicted decision $\hat{a}_{i,T,n}^{(s)} = \mathbb{1}[\alpha_{i,T}^{(s)} + \frac{1}{\sigma_{\epsilon,T}} \cdot r_{i,T,n}^{(s)} + \frac{\beta_{i,T}^{(s)}}{\sigma_{\epsilon,T}} \cdot \Phi_T(s_n) \geq 0]$. For each draw, we compute the predicted accuracy for each treatment by averaging over the evaluation set and posterior draws:

$$\text{Accuracy}_{i,T} = \frac{1}{S}\frac{1}{N}\sum_{s=1}^{S}\sum_{n=1}^{N}\mathbb{1}\left[\hat{a}_{i,T,n}^{(s)} = y_n\right]$$

where N is the number of applications and $y_n \in \{0, 1\}$ is the ground-truth label for application n.

For each specialist, their optimal treatment is the one that maximizes this expected accuracy, $t_i^{\text{opt}} = \arg\max_t \text{Accuracy}_{i,T}$.

C.5 Personalization Experiment Design

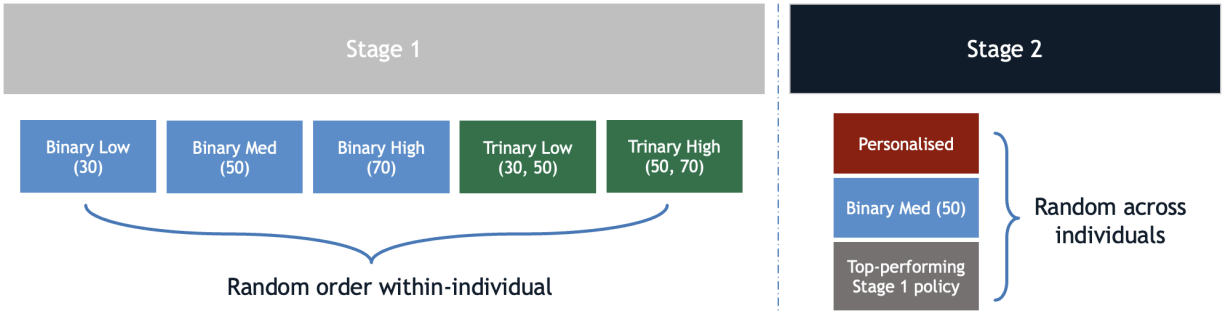
C.5.1 Survey flow and sample

Figure C.19 summarizes the two-stage survey flow. Our experiment took place between May 17 and June 24, 2025, recruiting 218 US, UK, Canada, Australia, and New Zealand-based participants aged 18 or older over Prolific. Qualified participants completed the first online survey (Stage 1) on the oTree platform via Heroku (approximately 60 minutes; 50 loan application decisions plus 10 practice questions, alongside additional questions on biases and demographics detailed in Appendix C.1.1). We oversampled some loan types for power, and offered optional loan specialist training (a short < 2 min video) before the practice questions. Table C.28 summarizes the demographic characteristics of the 218 recruited participants, 173 of whom completed both stages. Loan application decisions are made as in the main experiment (Section 4.1); the primary difference is that AI signals are shown on a dial with accompanying text (e.g., "This applicant's score is between 0–50"), minimizing visual-design differences across binary and trinary conditions (Figures C.20–C.21).

Table C.28: Participant Demographics in Personalization Experiment

	Sample Mean or Percentage
N (participants)	218
Age	38
Female	48.62%
Bachelor's Degree or Above	83.49%
Income	101,388
Loan Specialist Experience	100.00%
Loan Specialist Years	6
Bank / Credit Union Experience	95.41%
Applied to Loans Before	95.41%
Loan Amount (USD$)	216,912
Used Supervised AI Before	51.38%
Never Used any AI	7.80%

Figure C.19: Personalization experiment survey flow



C.5.2 Personalization Experiment Additional Instructions

Figure C.20: Instructions for *Trinary High (50, 70)* condition in the personalization experiment

Instructions for next 2 tasks More details

Instructions - Assistance from Emery the AI

For the next 2 questions, you will be provided assistance from an artificial intelligence (AI) named Emery that will show you one of three options:

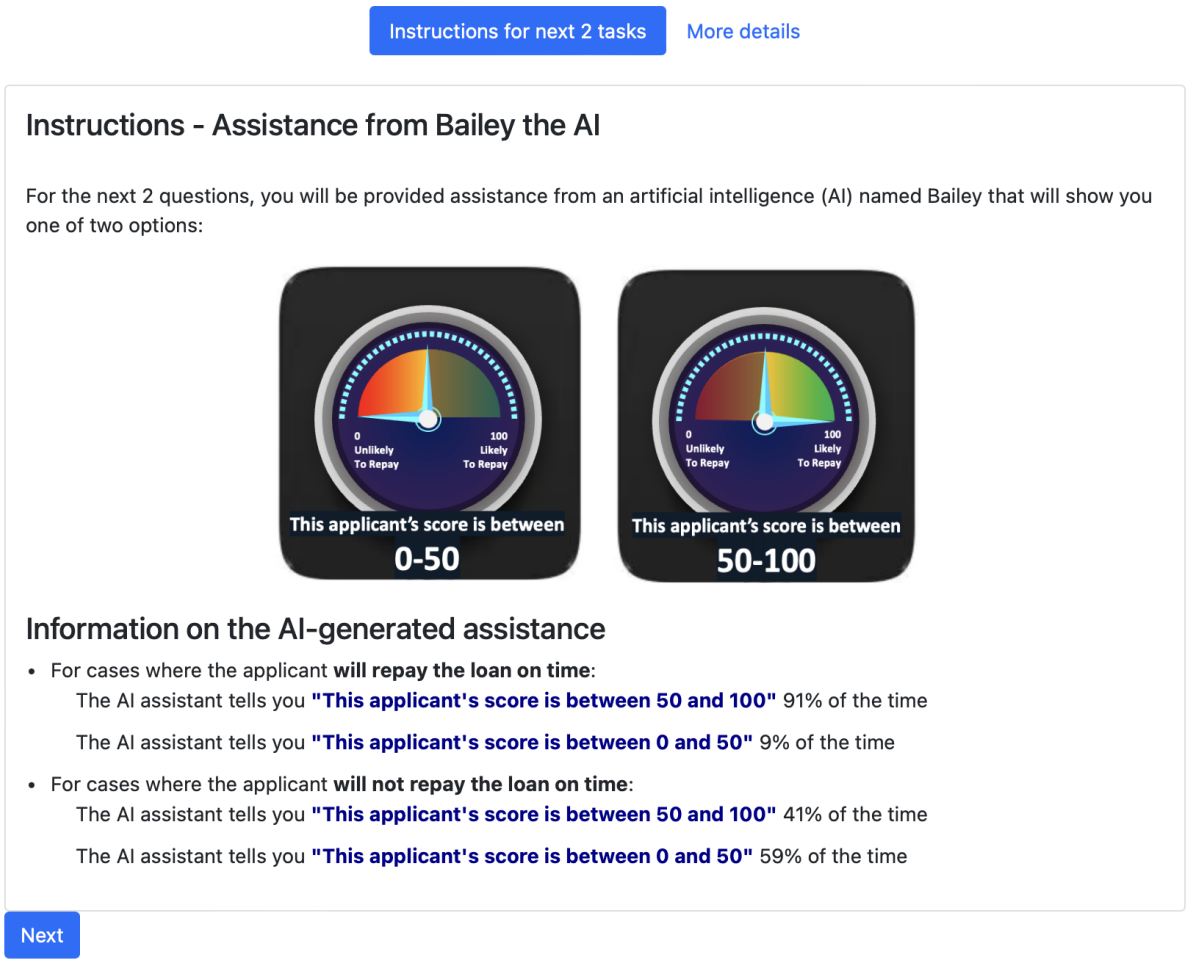


Information on the AI-generated assistance

- For cases where the applicant **will repay the loan on time**:
  - The AI assistant tells you **"The applicant's score is between 70 and 100"** 73% of the time
  - The AI assistant tells you **"The applicant's score is between 50 and 70"** 18% of the time
  - The AI assistant tells you **"The applicant's score is between 0 and 50"** 9% of the time
- For cases where the applicant **will not repay the loan on time**:
  - The AI assistant tells you **"The applicant's score is between 70 and 100"** 18% of the time
  - The AI assistant tells you **"The applicant's score is between 50 and 70"** 24% of the time
  - The AI assistant tells you **"The applicant's score is between 0 and 50"** 59% of the time
- More details on the loan applicants, human, and AI-generated assistance can be found by clicking on "More details"

Next

Figure C.21: Instructions for *Binary Medium* condition in the personalization experiment



C.6 Empirical Evidence—Personalization Experiment

C.6.1 Main results

Table C.29 presents regression results from Equation 5, using the *Personalized* treatment as the reference level; there are no significant differences in approval accuracy between the different treatment conditions.

Table C.29: Decision-making outcomes with and without fixed effects

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Binary Medium (Randomized)	−0.002	−0.002
	(0.015)	(0.015)
Trinary High (Randomized)	0.0005	0.0005
	(0.015)	(0.015)
Constant	0.670***	0.630***
	(0.009)	(0.038)
Observations	5,189	5,189
R^2	0.00000	0.364
Adjusted R^2	-0.0004	0.360

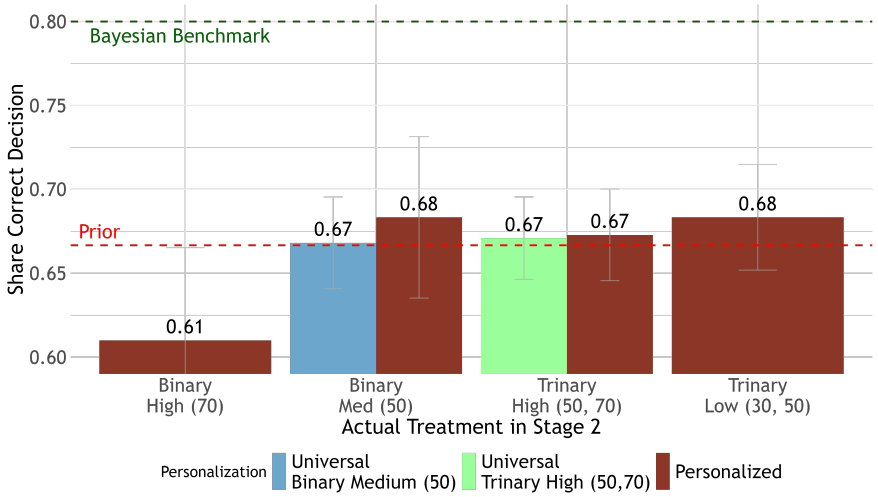
Note: Regression estimates per Equation 5 of loan approval accuracy by treatment condition, with *Personalized* as the omitted reference group. Column (1) reports estimates without fixed effects; Column (2) includes application (case-level) fixed effects. We do not include individual or block order fixed effects because the personalization treatment is assigned at the participant level and does not vary across blocks within a participant. Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01.

As a sense check, we decompose Figure Xb by the treatments participants actually received in Stage 2. For the *Binary Medium* and *Trinary High* treatments, we compare performance between (i) participants assigned via the Universal condition (random assignment, independent of model recommendations) and (ii) those assigned via the Personalized condition (model recommending that treatment). If personalization works as intended, accuracy should be higher in the latter group. As noted in the pre-analysis plan, the study was not powered to detect effects for this comparison.

Figure C.22 shows slightly higher accuracy in the Personalized conditions (blue bars) relative to Universal (purple and grey), though differences are not statistically significant. This suggests the model may select individuals who benefit somewhat more from these treatments, and crucially,

we find no evidence that personalization worsens outcomes.

Figure C.22: Experiment 2 Results by Actual Treatment



C.6.2 Main results, reweighted

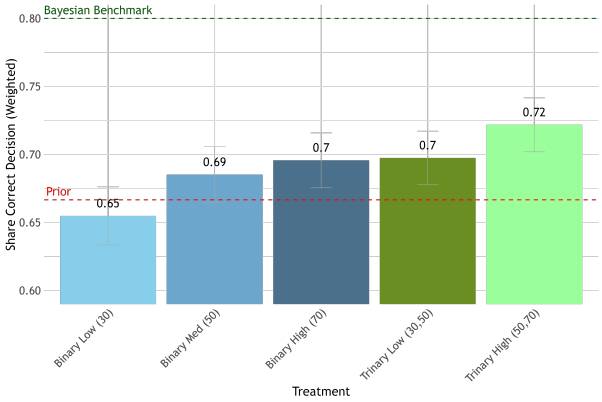
We apply reweighting procedure to correct for our strategic oversampling design. During data collection, we oversampled cases in the 10-30th and 60-80th percentiles of the *predicted* human signal—regions near the margins of adherence, where non-*Binary Medium* treatments are expected to have greater impact—to increase statistical power (sampling weight = 2). In our reweighted analysis, we assign a weight of 0.5 to these oversampled groups to restore the original population distribution. Standard errors are adjusted accordingly to account for the reweighting.

C.6.3 Attrition

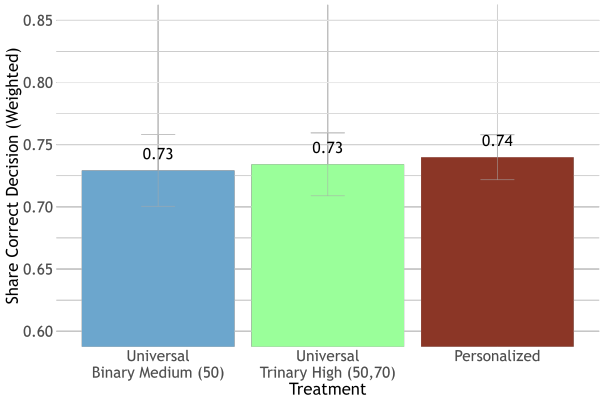
Table C.30 reports detailed attrition rates by final treatment condition and whether the treatment was assigned via personalization. Attrition is notably higher in the non-personalized *Trinary High* (20.3%) and *Binary Medium* (25.5%) groups. By contrast, the Personalized Binary High condition has a relatively low attrition rate (9.1%), despite showing low performance in Stage 2. While this raises potential concerns about selection bias—e.g., if those who remained in the Personalized condition may have been systematically weaker—the completers who form our analysis sample are balanced on baseline ability: Stage 1 accuracy among completers is statistically indistinguishable across treatment groups.

To probe this further, Table C.31 compares Stage 1 performance across treatment groups. Participants in the Personalized condition have slightly lower Stage 1 accuracy (0.665) than those in the Universal *Trinary High* group (0.669), though the difference is small and not statistically significant (SE = 0.005–0.007).

Figure C.23: Stage 1 and Stage 2 results (reweighted), personalization experiment



(a) **Stage 1.** Bars represent the share of correct decisions under each treatment condition. Dashed red line indicates prior; dashed green line marks Bayesian benchmark. Error bars represent 95% CIs.



(b) **Stage 2.** Bars represent the share of correct decisions in each treatment condition. Error bars represent 95% CIs.

Table C.30: Detailed Attrition Analysis by Actual Treatment (Stage 2) and Randomized Condition

Actual Treatment S2	Condition	Assigned	Completed	Attrition Rate
Binary High	Personalized	11	10	9.1%
Binary Medium	Universal	51	38	25.5%
Binary Medium	Personalized	13	12	7.7%
Trinary High	Universal	59	47	20.3%
Trinary High	Personalized	45	38	15.6%
Trinary Low	Personalized	39	28	28.2%

Table C.31: First Stage Performance by Second Stage Treatment

Second Stage Treatment	Share Correct (First Stage)	Std Error	N
Universal Binary Medium (50)	0.664	0.007	38
Universal Trinary High (50,70)	0.669	0.006	47
Personalized	0.665	0.005	88

Taken together, these patterns suggest that while attrition varies across groups, differential attrition did not leave the compared arms with different baseline (Stage 1) accuracy, and the small average differences in Stage 1 accuracy are insufficient to account for the lack of observed gains from personalization. Because this balance is assessed on observables among completers, we cannot rule out selection on unobservables.

C.6.4 Overprediction of personalization gains

To test whether the model overpredicts the benefits of personalization, we compare predicted and observed gains across treatment conditions. We define:

$$\Delta_{\text{pred}} = \mathbb{E}[\hat{y} \mid \text{Personalized}] - \mathbb{E}[\hat{y} \mid \text{Universal}], \quad \Delta_{\text{actual}} = \mathbb{E}[y \mid \text{Personalized}] - \mathbb{E}[y \mid \text{Universal}]$$

where $\hat{y}$ is the model-predicted accuracy and y is the observed Stage 2 accuracy. The *gain gap*, $\Delta_{\text{pred}} - \Delta_{\text{actual}}$, captures any systematic over- or under-prediction. Using a non-parametric bootstrap (5,000 resamples) and treating $\hat{y}$ as fixed, we find a 95% confidence interval for the gap of $[0.009, 0.057]$, indicating statistically significant overprediction. To incorporate posterior uncertainty, we perform a Bayesian analogue using posterior draws from the hierarchical model: for each posterior draw s,

$$\text{Gap}^{(s)} = \left(\mathbb{E}^{(s)}[\hat{y} \mid \text{Personalized}] - \mathbb{E}^{(s)}[\hat{y} \mid \text{Universal}]\right) - \left(\mathbb{E}[y \mid \text{Personalized}] - \mathbb{E}[y \mid \text{Universal}]\right),$$

yielding a 95% Bayesian credible interval of $[0.021, 0.047]$, again supporting overprediction.

C.6.5 Model calibration

Table C.32 presents a direct test of the model's calibration, regressing actual performance on model-predicted performance. The coefficients are positive and significant across specifications, ranging from 0.414 to 0.544. This implies that when the model predicts a participant will perform 1 percentage point better under their assigned treatment, they actually perform roughly 0.4-0.5pp better on average. These results remain stable when controlling for treatment fixed effects, suggesting that the model captures perfomance variation consistently across treatments.

C.6.6 Placebo test: Does the model perform better than random?

To verify that the model's predictions are not simply an artifact of noise or overfitting to random structure in the data, we conduct a placebo test.

We rerun the model on a version of the data in which Stage 2 treatment assignments are randomly reshuffled across participants. This preserves observed decisions, signals, and the model structure, but breaks any systematic relationship between participants and their assigned treat-

Table C.32: Quality of Fit: Stage 2 Accuracy on Model-Predicted Accuracy

	Dependent variable:					
	Stage 2 Accuracy					
	Basic	Pers.	Baseline	Demog.	AI Exp.	Treatment FE
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted Accuracy	0.455***	0.544***	0.414***	0.427***	0.432***	0.420**
	(0.141)	(0.154)	(0.158)	(0.161)	(0.162)	(0.165)
Personalized		−0.018	−0.013	−0.013	−0.013	−0.008
		(0.013)	(0.013)	(0.013)	(0.013)	(0.016)
Baseline Performance			0.074***	0.075***	0.074***	0.067**
			(0.026)	(0.026)	(0.027)	(0.027)
Age				0.001	0.001	0.001
				(0.001)	(0.001)	(0.001)
Underwriting Years				0.00001	−0.00004	0.00000
				(0.001)	(0.001)	(0.001)
Female				−0.005	−0.005	−0.004
				(0.012)	(0.012)	(0.012)
AI Never					0.008	0.007
					(0.022)	(0.023)
Actual Treatment (Binary Med)						0.046
						(0.029)
Actual Treatment (Trinary High)						0.035
						(0.028)
Actual Treatment (Trinary Low)						0.044
						(0.029)
Constant	0.362***	0.311***	0.150	0.115	0.113	0.100
	(0.095)	(0.102)	(0.115)	(0.120)	(0.121)	(0.124)
Observations	173	173	173	173	173	173
R^2	0.058	0.068	0.110	0.121	0.121	0.138
Adjusted R^2	0.052	0.057	0.094	0.089	0.084	0.085
Note:						*p<0.1; **p<0.05; ***p<0.01

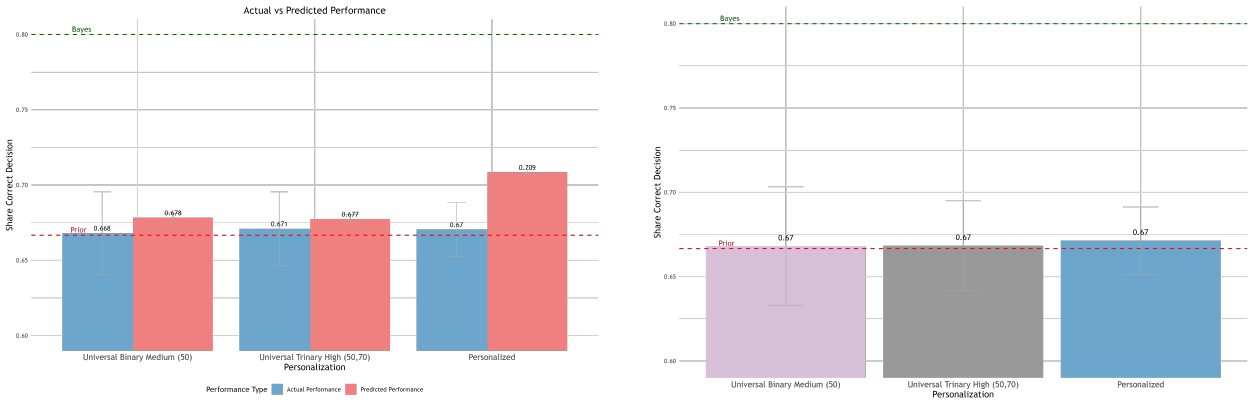
ments. If the model is working well, it should not predict gains to personalization with this reshuffled data. Reassuringly, both the bootstrap 95% confidence interval ($[-0.016, 0.013]$) and Bayesian credible interval ($[-0.026, 0.023]$) are centered around zero. While this test cannot assess whether the model is well-calibrated, it provides reassuring evidence that the model does not generate spurious treatment effects when none exist.

C.6.7 Is the model mispredicting r?

When predicting loan approval decisions, the model treats r (the human signal) as latent:

$$r_{i,T,n}^{(s)} \sim \mathcal{N}(X_n \lambda_t^{(s)} + Z_i \delta_t^{(s)}, \sigma_{r,T}^{2\,(s)}).$$

However, since we observe $\hat{r}$ for 90% of the observed cases in Stage 2, we can plug in $\hat{r}$ instead of treat r as latent. Figure C.24a shows that even when we plug in the observed $\hat{r}$ directly, the model continues to predict sizable gains from personalization.



(a) Stage 2 actual accuracy vs. predicted accuracy using plug-in $\hat{r}$

(b) Stage 2 results, restricting to participants with matching treatments under plug-in $\hat{r}$

Figure C.24: Stage 2 predictions and outcomes using observed $\hat{r}$.

We then also compare treatment assignments under the latent-r and plug-in $\hat{r}$ versions. Assignments match in 77% of cases, differing for 39 participants. We re-run Stage 2 analyses excluding participants whose treatments differ across the two methods. As shown in Figure C.24b, this does not improve performance—suggesting that misprediction of r is unlikely to explain the lack of observed gains from personalization.

C.6.8 Stage 1 fit: out-of-sample predictions on holdout set

To test stage 1 fit, we assess out-of-sample predictive accuracy using a holdout procedure. One random decision per participant-treatment is held out, and the model is fit on the remaining data. We then compute posterior-predictive accuracy on the held-out cases using the full posterior distribution of model parameters.

Results show strong out-of-sample predictive accuracy: the model correctly replicates participants' approve/deny decisions in 77% to 84% of held-out cases on average, across the treatments. This indicates that the model captures individual-level decision behavior and generalizes well to unseen cases.

However, because only one decision per treatment per participant is held out, we cannot estimate individual decision accuracy or assess the quality of personalized treatment assignment. Rather, this analysis confirms that the model captures individual-level decision rules well enough to make reasonably accurate predictions on new data, but leaves open the question of whether it is useful for personalization.

C.6.9 Stage 2 fit, out-of-sample predictions on holdout set

Another test we can run is fitting the decision model on Stage 2 data, and running the aforementioned test. If the model is well-calibrated, then a Stage 2-trained model should perform well in-sample. Results again show strong out-of-sample predictive accuracy: the model correctly replicates participants' approve/deny decisions in 85% to 90% of held-out cases across treatments.[82]

C.6.10 Stage 1 fit: in-sample predictions controlling for fixed effects

We next assess in-sample fit by asking whether specialists perform better in their model-predicted best treatment, controlling for case difficulty. Since treatments may have received harder or easier cases by chance, we include case fixed effects to isolate the impact of signal presentation on decision accuracy.

Table C.33 reports results. The regressions are at the decision level: the dependent variable is an indicator for a correct decision (218 participants $\times$ 50 Stage 1 rounds $=$ $10{,}900$ decisions), with case fixed effects. Column (1) shows that participants performed significantly better in their

[82] It is unsurprising that this performs better than the Stage 1 out-of-sample test, since we observe 30 decisions per treatment-specialist in Stage 2, as opposed to 10.

predicted top treatment: the coefficient of 0.070 implies a 7 percentage point gain. Column (2) adds treatment-by-top-treatment interactions and finds the effect remains robust.

These results provide in-sample evidence that the model successfully identifies high-performing treatments within Stage 1. However, because the same data is used to estimate the model and test outcomes, this analysis reflects in-sample fit, not generalization. Note that we cannot run the in-sample test described here using a model fitted on Stage 2 data, as we only observe one treatment per participant in Stage 2.

Table C.33: Decision-making outcomes under different treatment conditions, S1 data

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Binary High	0.012	0.007
	(0.012)	(0.012)
Binary Low	-0.033^{***}	-0.036^{***}
	(0.012)	(0.012)
Trinary High	0.021^{*}	0.007
	(0.012)	(0.014)
Trinary Low	0.008	0.009
	(0.012)	(0.014)
Top Treatment (Model Predicted)	0.070^{***}	0.041^{*}
	(0.010)	(0.025)
Binary High × Top		0.046
		(0.039)
Binary Low × Top		−0.039
		(0.091)
Trinary High × Top		0.052^{*}
		(0.030)
Trinary Low × Top		0.016
		(0.030)
Constant	0.772^{***}	0.774^{***}
	(0.032)	(0.032)
Observations	10,900	10,900
R^2	0.333	0.333
Adjusted R^2	0.327	0.327
Note:	*p<0.1; **p<0.05; ***p<0.01	

Note: Reference category is the *Binary Medium (50)* treatment condition with non-top performers. Column (1) includes additive treatment and top-treatment; Column (2) adds treatment × top-treatment interactions. Both columns include case fixed effects, omitted from the table for brevity.

We can run the same regression framework using Stage 2 outcomes (with the top treatment defined by that recommended by the model fitted on Stage 1 data). Consistent with the lack of observed gains from personalization, we find no significant evidence of model fit in this setting.

Table C.34: Decision-making outcomes under different treatment conditions, S2 data

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Binary High	-0.063^{**}	-0.070^{**}
	(0.025)	(0.027)
Trinary High	-0.001	0.008
	(0.013)	(0.018)
Trinary Low	0.010	0.003
	(0.018)	(0.021)
Top Treatment (Model Predicted)	0.002	0.013
	(0.013)	(0.020)
Binary High × Top		
Trinary High × Top		-0.018
		(0.026)
Trinary Low × Top		
Constant	0.631^{***}	0.628^{***}
	(0.030)	(0.031)
Observations	5,189	5,189
R^2	0.365	0.365
Adjusted R^2	0.361	0.361
Note:	*p<0.1; **p<0.05; ***p<0.01	

Note: Reference category is the *Binary Medium (50)* treatment condition with non-top performers. Case fixed effects included but omitted from the table for brevity. Column (1) includes additive treatment and top-treatment; Column (2) adds treatment × top-treatment interactions. Trinary Low × Top and Binary High × Top do not have coefficients because *all* observations in Binary High and Trinary Low are (by definition) in the top treatment, since those conditions can only be assigned when personalized.

C.6.11 Stage 1 and Stage 2 adherence as moderators

Figure C.25 plots the joint distributions of adherence curve differences across the two stages.[83] While there is some consistency, we observe systematic shifts—particularly in the Trinary conditions. These shifts are non-trivial in scale, undermining the assumption of behavioral stationarity

[83] Estimates are shown for three of the five treatment conditions; no participants were assigned *Binary Low*, and too few were assigned *Binary High* in Stage 2 for decision model convergence.

between stages.

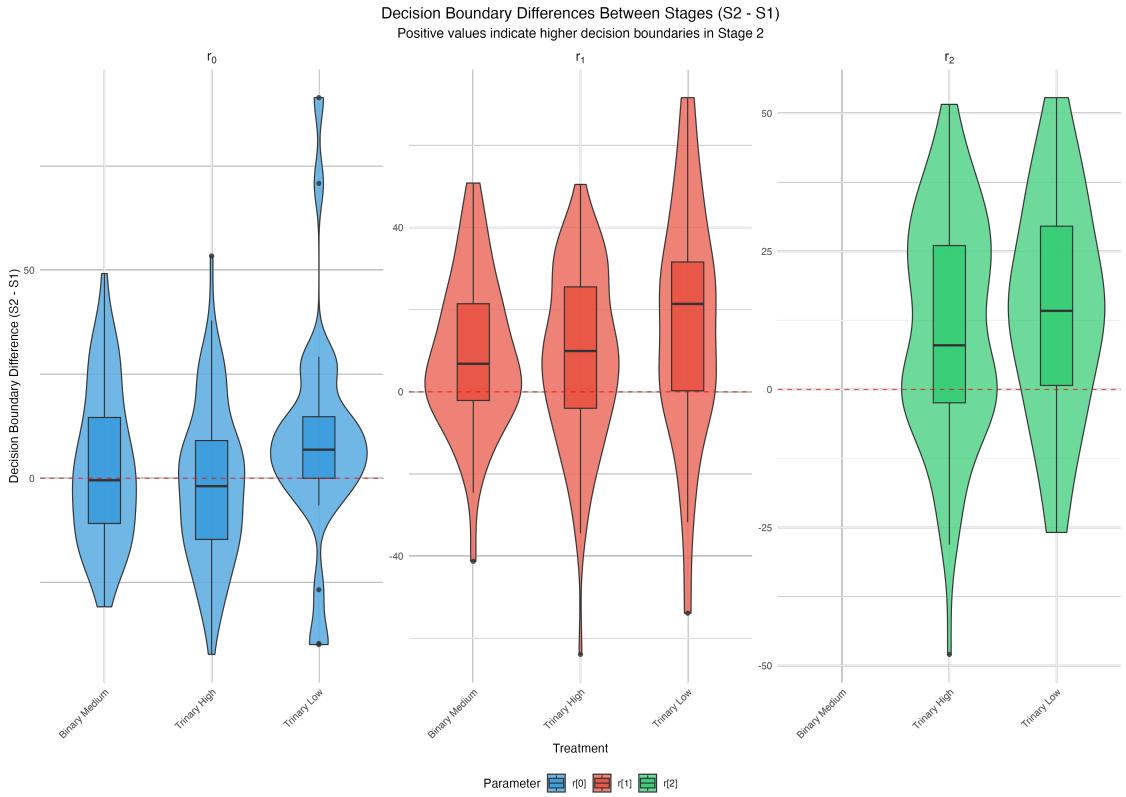


Figure C.25: Comparison of r_0, r_1 and r_2 adherence estimated from Stage 1 vs. Stage 2

Note: Y-axis denotes the difference between Stage 2 and Stage 1 adherence curve values of r_0, r_1, r_2 (S2 - S1), split out by treatment. Positive values indicate higher values in Stage 2. The violin plots show the full distribution of all posterior sample differences, while the boxplots show the median (center line), interquartile range (box), and whiskers extending to 1.5 × IQR.

We test whether these Stage 1–Stage 2 adherence shifts moderate the effect of personalization. These estimates exclude the ten Personalized *Binary High* completers (300 rounds), for whom the adherence-bound moderator is not defined. Table C.35 splits participants by whether their average absolute adherence difference across the boundaries defined under their policy (r_0 and r_1 for binary, plus r_2 for trinary coarsenings) is above or below the median. Those below median exhibit positive gains from personalization (though not statistically significant), whereas those above median do not.

Table C.35: Decision-making outcomes, without fixed effects

	Dependent variable:	
	Share of Decisions Correct	
	Below Median Adherence Curve Diff	Above Median
	(1)	(2)
Binary Medium (Randomized)	−0.024	−0.000
	(0.020)	(0.026)
Trinary High (Randomized)	−0.018	−0.000
	(0.021)	(0.021)
Constant	0.693***	0.667***
	(0.015)	(0.011)
Observations	2,459	2,430
R^2	0.001	0.000
Adjusted R^2	-0.0003	-0.001

Note: Regression estimates per Equation 5 of loan approval accuracy by treatment condition, with *Personalized* as the omitted reference group. Column (1) reports results for those below median average absolute adherence difference across the boundaries defined under the participant's policy (r_0, r_1 for binary; r_0, r_1, r_2 for trinary); Column (2) for above median. Asymptotic standard errors clustered at the individual level in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table C.36 examines this interaction: among participants whose adherence curves are stable (adherence difference near zero), the effect of personalization is positive and significant at the 10% level, though the interaction term itself is not statistically significant. This is suggestive that personalization may perform better when decision strategies remain stable, rather than establishing stability as a necessary condition.

Table C.36: Decision-making outcomes with fixed effects

	Dependent variable:
	Share of Decisions Correct
Personalized	0.039^{*}
	(0.023)
Adherence Diff (S2-S1)	0.0001
	(0.001)
Personalized x Adherence Diff (S2-S1)	-0.002
	(0.001)
Constant	0.638^{***}
	(0.039)
Observations	4,889
R^2	0.370
Adjusted R^2	0.366

Note: Dependent variable is an indicator for whether the subject's decision was correct. Coefficients on *Personalized* report mean differences in correctness relative to when unpersonalized, evaluated when the adherence boundary difference is zero. *Adherence Bound Diff (S2–S1)* measures how accuracy changes with a one-unit increase in adherence boundaries for the baseline group, and the interaction terms show how this slope differs across personalization conditions. Case fixed effects included. Asymptotic standard errors (in parentheses) are clustered at the individual level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.