Calibrated Coarsening: Designing Information for AI-Assisted Decisions

Bnaya Dreyfuss
Harvard University

Ruru Hoong*
Harvard Business School

Job Market Paper

This version: March 27, 2026
Click here for latest version

Abstract

Artificial intelligence (AI) is increasingly used to aid human decision-making across critical applications, but errors in human probabilistic reasoning (e.g., cognitive biases) often undermine its effectiveness. This raises the central design 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 chosen thresholds—as a way to do so that (i) ensures humans retain final decision authority, (ii) modifies signals without deception, and (iii) adapts flexibly to diverse biases and contexts. 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 a randomised experiment with professional loan specialists, we empirically demonstrate that coarsening AI signals at the theory-derived threshold significantly improves decision-making outcomes, 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 decision heterogeneity amongst loan officers and estimate a Bayesian hierarchical linear model (HLM), which we use to simulate and test the benefits of personalised coarsening policies.

*Contact: ruruhoong@g.harvard.edu. Ruru Hoong is 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, Josh Schwartzstein, Bennett Smith-Worthington, Clare Suter, Chris Walker and the behavioural and labour 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, the NUS Development Grant, and the 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.

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—whether due to institutional constraints such as legal liability,[2] or technical constraints such as the ability to access information beyond the AI's scope.[3] 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.

Behavioural science has long established that humans struggle to process and aggregate information, documenting how 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 flexible enough to accommodate 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 chosen thresholds—as a broad approach to improving decision-making outcomes that: (i) retains humans' final decision rights, (ii) transparently modifies signals without deception, and (iii) is adaptable to a wide range of cognitive biases or decision-making contexts. Coarsening is already commonplace in information-provision settings, from discretised scores (e.g., 1, 2, 3) to colour-coded risk (e.g.,

[1] e.g., Algorithmic risk assessments have not improved sentencing outcomes (Stevenson and Doleac, 2019; 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, 2011; 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] E-commerce firm Zalando uses human oversight in algorithmic pricing for this reason (Huelden et al., 2024). GDPR's Article 9 bans processing certain data types (incl. for AI training) unless explicit consent is given or data is public.

green vs. red),[4] underscoring its intuitive appeal over raw AI probability 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 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, irrespective of the human biases at play. When more is known about individual behaviour, further calibration is possible.

In this paper, we develop a theoretical framework to derive an approximately optimal binary coarsening policy, and empirically validate it in a randomised experiment with 150 loan officers, underwriters, and processors (henceforth, "loan specialists") recruited through professional channels—including job postings on Indeed and LinkedIn, and targeted outreach to members of the National Association of Mortgage Underwriters and Processors—with an average of 12 years of industry experience. Participants in all treatments review loan applications and make approval decisions; conditions vary in whether they receive an AI signal, and if so, how it is coarsened. We find that coarsening AI signals at the theory-derived threshold significantly improves approval accuracy, outperforming both human-only (based solely on the application) and continuous AI (assisted with uncoarsened AI risk-score) benchmarks, and two arbitrary coarsening benchmarks. 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 behaviour, moderated by both demographics and elicited bias measures. This heterogeneity motivates calibrated coarsening in a double sense: choosing a non-arbitrary threshold and, when feasible, calibrating to individuals. We explore the scope for such personalisation by estimating a Bayesian HLM on past data, predicting modest gains of ~ 1.5 percentage points that we test in a two-stage experiment.

How can coarsening signals—making them strictly *less* informative—lead to better decisions? Intuitively, this can occur when specialists' decisions deviate from those of a fully rational Bayesian, even when given the same information. We refer to these divergences as "disagreement regions." Calibrated coarsening shrinks these regions by pooling them with nearby areas of agreement—a core insight from information design (Kamenica and Gentzkow, 2011)—aligning non-Bayesian actions more closely with those of Bayesians. For example, if a Bayesian would approve a loan at a given AI signal but a non-Bayesian would not, we can pool that signal together with higher

[4]E.g., Pretrial risk is scored on a discrete 1–5 scale (see Figure 2 in Angelova et al. (2023)), and dental platform VideaAI colour-codes disease risk into 3 bands instead of providing the raw probabilities (see Appendix Figure 14).

ones where both approve, pushing the non-Bayesian towards approval. The same logic applies in reverse, leading non-Bayesians to make the same decision as Bayesians more often.

To generalise this pooling intuition, we draw on canonical (non-)Bayesian persuasion models (Kamenica, 2019); however, our theoretical framework also allows for other mechanisms, such as differences in belief updating under coarse signals, to affect behaviour. A Receiver (e.g., loan specialist) takes a binary action to match a binary state (e.g., approve/deny the loan depending on whether or not they think it will be repaid). Before taking action, the Receiver forms posterior beliefs about the state based on (i) privately observed information (e.g., by reviewing a loan application), which we refer to as the *human signal*, and (ii) a message from the Sender (e.g., an AI designer or bank manager). The Sender publicly commits to a deterministic monotone *disclosure mechanism* mapping signals to messages, privately observes their own *AI signal* (e.g., a risk score from AI model output), and generates a message accordingly. We assume a continuously differentiable joint distribution of human and AI signals satisfying MLRP, allowing for any arbitrary correlation conditional on state. While the Sender and Receiver have aligned incentives, their beliefs may differ. The Sender is consistent with the normative benchmark, aggregating all available information using Bayes' rule, whereas the Receiver's belief may deviate arbitrarily—subject only to mild continuity and monotonicity. This belief misalignment distorts Receiver choices, as in non-Bayesian persuasion (de Clippel and Zhang, 2022). When choosing a disclosure mechanism, the Sender's goal is to align the Receiver's behaviour as closely as possible with the Bayesian benchmark.

Focusing on binary coarsenings, we show that when the Sender cannot condition the mechanism on the Receiver's signal, an approximately optimal binary threshold is the *AI-only threshold*, above which a Bayesian relying solely on the AI signal would approve. The approximation is good when (i) the *adhering types*—those who follow the AI's implied recommendation—are representative of the overall population, and (ii) shifts in the set of adhering types are approximately welfare-neutral. If, however, the human signal provides no residual information—that is, it is not predictive of repayment conditional on the AI signal—then the first condition holds automatically. We assess the approximation with simulations calibrated to parameters estimated from our experimental data: $\hat{t}$ and the optimal binary threshold yield essentially the same welfare in all of our simulations. Coarsening yields substantial gains over full revelation for some but not all parameter configurations — whether it helps in a given setting is ultimately an empirical question.

To that end, our main experiment empirically tests whether coarsening signals at the AI-only

threshold improves decision-making outcomes, demonstrating the real-world applicability of this approach. We ask 150 loan specialists to approve loan applications they believe will be repaid on time, and deny otherwise. All specialists review simplified versions of real loan applications, allowing them to form beliefs about the probability of repayment (human signal). We generate a risk-score for each application (AI signal) using an AI model we trained on repayment data from Home Credit.

We randomly assign participants to five treatment conditions that vary how the AI signal is shown: *Human-only* (no AI), *Probability* (the raw AI risk-score), and three *Binary* treatments which display only whether the AI signal is above or below a given threshold (e.g., "Yes, this applicant's AI score is above [threshold]"). Participants are incentivised with a bonus for every correct decision.[5] Given this incentive structure, 50% is the approximately optimal binary coarsening policy (AI-only threshold). We refer to this as *Binary Medium*, and also include two placebo binary conditions, *Binary Low* (30%) and *Binary High* (70%). This enables us to evaluate the value of theoretically-informed coarsening over arbitrary coarsening. We also estimate in the experiment data that 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.[6]

We find that providing loan specialists with AI signals coarsened at the *Binary Medium* threshold significantly improves decision accuracy compared to all other conditions. Participants are 4 percentage points (pp) more accurate than under *Probability*—closing 40% of the gap to the Bayesian benchmark—and 6pp more accurate than under *Human-only*, underscoring the power of effective information design in closing the gap between actual and optimal decision-making.

Coarsening not only improves accuracy but also significantly reduces decision time—a critical factor in high-throughput settings like loan underwriting. Participants are 7% faster in *Binary Medium* than in the *Probability* condition, and many preferred coarsened signals: in an incentivised choice task, over half opted for either binary messages or no AI assistance over a full probability score, and those who chose 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

[5]Participants receive a base payment of $16-60 depending on the recruitment channel, plus a performance-linked bonus averaging 10–25% of the base.

[6]e.g., 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.

and cognitive processing costs (Guan et al., 2025, 2023; Exley and Kessler, 2023), and also echo interviews we conducted with senior lending professionals, who acknowledged the value of AI tools but found them "challenging to use" and easily "overridden". Taken together, these findings suggest that coarsened signals can reduce cognitive burden and encourage adoption in practice.

Performance gains, however, are not uniform across specialists, offering insight into the mechanisms driving the outcome gap and corresponding gains from coarsening. Gender emerges as a striking moderator: men benefit nearly twice as much from AI assistance than women, reflecting both a lower baseline and greater willingness to rely on AI advice. Moreover, consistent with theories of algorithmic underweighting (AU)—the tendency to underweight algorithmic information—specialists exhibiting AU on an elicitation task perform better with coarsened than full AI risk scores, while those more Bayesian on this scale show the converse, implicating AU as a potential driver of the deviation. Taken together with other bias moderators, these findings suggest that suboptimal AI use stems not from a single factor but from a combination of behavioral deviations. This underscores the value of calibrated coarsening as a robust intervention that can address multiple biases simultaneously,[7] and also sets the stage for our personalised approach: rather than imposing more structural assumptions on how individuals update, we use past actions to estimate and assign the most effective coarsening policy.

To explore this, we estimate a Bayesian hierarchical model on our main experiment data to predict the gains from personalising coarsened policies. Our model suggests modest additional gains from personalisation relative to the calibrated uniform policy—approximately 1.5pp—representing an upper bound that remains below the Bayesian benchmark and smaller than the 4pp improvement from universal calibrated coarsening relative to full AI assistance. We then test this in a two-stage experiment with new participants. The model uses decision data from the first survey to estimate individual behaviour under different coarsening policies. In the second survey, participants are randomised to receive either their model-predicted best policy or universal policy benchmarks. We find no average gains in practice—likely due to both the modest predicted gains and behavioural nonstationarity across stages: marginally significant personalisation gains occur only when individual behaviour remains stable between estimation and deployment. These results highlight the challenges of operationalising personalisation in dynamic environments where behavioural shifts

[7] A multi-parameter structural bias model could fit almost any choice patterns under strong auxiliary assumptions, so we instead take a less structural hierarchical Bayesian approach geared toward predicting which coarsening works best for whom. For settings where a single dominant behavioural 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 characterising optimal coarsening and its expected gains.

are common—and suggest that in such settings, a universal calibrated coarsening rule may already offer a simple, robust, 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 (Tversky 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 one of the first empirical applications of information design to address human errors in probabilistic reasoning. Our theoretical framework draws on (non-)Bayesian persuasion (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 Sender and Receiver preferences are aligned and the Receiver updates beliefs correctly, full disclosure is optimal (Kamenica and Gentzkow, 2011). When instead the 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), some other information structure—of which coarse partitions is a subset—can be optimal. Other work treats coarsening as a *constraint* limiting persuasion (Aybas and Turkel, 2019) 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, 2024; Avoyan et al., 2024), 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 coarsening improves outcomes by addressing cognitive biases.

Second, 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., 2019; Kiani et al., 2020; Maron et al., 2020; Bastani et al., 2021; 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., 2024; Kim et al., 2024; Snyder et al., 2024; Otis et al., 2024; Grimon and Mills, 2025; Caplin et al., 2024), and a managerial prediction literature forecasts AI's positive impact on organisational 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). Realising 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), rational inattention (Boyacı et al., 2024), prospect theory (Ye et al., 2022), and correlation neglect (Agarwal et al., 2023)—that shape how humans interpret and act upon AI predictions. A key advantage of coarsening is its potential to address many such biases simultaneously.[8] Other proposed solutions include explanations and uncertainty information (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, 2022; Ibrahim et al., 2021; McLaughlin and Spiess, 2024; Dai and Singh, 2024).

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 true and false positives and negatives, rather

[8]Coarsening also connects to behavioural work on categorical 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. That said, if coarsened signals more closely resemble the categories that a categorical thinker would naturally gravitate towards, such signals may be easier to process and more readily adopted, potentially a reason for why we observe higher demand for coarsened AI signals in our experiment.

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. Theoretically, our papers make substantively different and non-nested assumptions: we assume monotonicity of actions in signals, while AMW assume human judgments depend on AI only via the resulting posterior on the state—an assumption that 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)).[9] These distinct modeling choices lead to different conclusions about when and how coarsening improves outcomes. Empirically, the papers yield opposing results in different contexts: AMW find no benefit to coarsening among Prolific participants performing fact-checking tasks; we find coarsening improves approval accuracy among professional loan specialists, relative to full disclosure.

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. Moreover, even beyond AI predictions, our framework applies more broadly to decision-making involving multiple signals.[10] We focus on algorithmic signals because they enable precise, transparent adjustments through threshold choice. 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.

The paper proceeds as follows. Section 2 presents a model of decision-making and illustrates how coarsening can address various behavioural 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 introduces a Bayesian HLM of decision-making and uses it to simulate gains from policy personalisation. Section 7 outlines the two-stage design of the personalisation experiment and estimates gains from personalisation. Section 8 concludes.

[9]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).

[10]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 organisations (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).

2 Model

2.1 Basic framework

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 $\omega \in \Omega = \{0, 1\}$ (e.g., whether the applicant repays on time). The Receiver's payoff from action $a = 1$ conditional on the state ω is

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

where c_{FP} is the cost of a false positive (wrongly approving an applicant who does not repay) and c_{TP} is the benefit of a true positive (correctly approving one who repays). The payoff from inaction ($a = 0$) is normalized to zero without loss of generality.[11]

The Receiver holds prior belief $p = \Pr(\omega = 1)$ and privately observes a *human signal* $r \in R \subseteq \mathbb{R}$ (e.g., from reviewing the application), which we also call their *type*. A *Sender* (e.g., a bank manager controlling AI system output, or the system designer) shares the Receiver's prior and payoffs,[12] and privately observes an *AI signal* $s \in S \subseteq \mathbb{R}$ (e.g., the applicant's risk score). Conditional on state ω, the signals (r, s) are jointly distributed according to $H_\omega(r, s)$ with marginals $G_\omega(r)$ and $F_\omega(s)$, and conditionals $G_\omega(r \mid s)$ and $F_\omega(s \mid r)$. Throughout, lowercase letters denote the corresponding densities (e.g., h_ω), which are all continuously differentiable and strictly positive. State-unconditional distributions are denoted by omitting ω from the subscript.

We allow r and s to be correlated even conditional on ω, but impose that H_ω satisfies the multivariate monotone likelihood ratio property (MLRP): higher signals imply a higher likelihood of $\omega = 1$. This ensures the Bayesian posterior $\pi(r, s) = \frac{p \cdot h_1(r,s)}{h(r,s)}$—the probability of $\omega = 1$ given the pair of signals r, s—is monotonically increasing in both arguments.

Disclosure mechanisms. The Sender does not act directly, but can influence the Receiver by choosing whether and how to reveal information about s via a weakly monotonic, deterministic *disclosure mechanism* $\Phi : S \to \mathbb{R}$ that maps signals to messages. We focus mainly on three types of mechanisms. *Full revelation*, denoted Φ^*, reveals s directly, i.e., $\Phi^*(s) = s$ for all $s \in S$. *Full censorship*, denoted $\Phi_\emptyset$, is uninformative about s, i.e., $\Phi_\emptyset(s) = k$ (where k is a constant)

[11] Equation 1 is derived by normalizing inaction without loss of generality from the utility function $\hat{u}(a|\omega) = -c_{FP}\mathbb{I}(a = 1, \omega = 0) + c_{TP}\mathbb{I}(a = 1, \omega = 1) - c_{FN}\mathbb{I}(a = 0, \omega = 1) + c_{TN}\mathbb{I}(a = 0, \omega = 0)$.

[12] To keep the focus on deviation from Bayesian updating, we assume shared priors and incentive alignment, though the framework can accommodate misaligned utilities and heterogeneous priors as in Kamenica and Gentzkow (2011).

for all $s \in S$. A *binary coarsening* at threshold t, denoted Φ_t, pools signals into two messages: $\Phi_t(s) = \mathbf{1}\{s \geq t\}$, with signals below t mapped to "No" ($k = 0$) and signals above to "Yes" ($k = 1$).[13]

Timing. The Sender publicly selects Φ; the state ω is realized and signals (r, s) are drawn according to H_ω. The Receiver observes $(r, \Phi(s))$ and chooses action a; utilities are realized.[14]

Posteriors. Given a disclosure mechanism Φ, the Bayesian posterior from signal r and message $\Phi(s)$ is denoted $\pi_\Phi(r, \Phi(s))$.[15] Since Φ is monotonic and H satisfies MLRP, $\pi_\Phi(r, \Phi(s))$ is monotonically increasing in both r and $\Phi(s)$, and is continuously differentiable in r, as well as in $\Phi(s)$ whenever Φ is continuous. While payoffs are aligned, the Sender's and Receiver's *beliefs* may diverge, generating a wedge between their subjective utility-maximizing actions. The Sender is Bayesian:[16] while they typically do not observe r, $\pi(r, s)$ is the posterior they would hold under full information. The Receiver's posterior may differ from the Bayesian benchmark, and is denoted $\tilde{\pi}_\Phi(r, \Phi(s))$. Under full revelation, we omit the subscript altogether and write $\tilde{\pi}(r, s)$. Note that in general the Receiver's posterior $\tilde{\pi}_\Phi$ can depend on the disclosure mechanism Φ itself (in addition to taking $\Phi(s)$ as input), allowing the updating heuristic to vary with Φ.

We remain agnostic about the source of any deviation from Bayesian updating—whether from belief misspecification or cognitive bias[17]—imposing only that, like the Bayesian posterior, the Receiver's posterior is also increasing in both signals:

Assumption 1. *For any monotonic disclosure mechanism Φ, the Receiver's posterior $\tilde{\pi}_\Phi(r, \Phi(s))$ is monotonic in both elements, continuously differentiable in r, and continuously differentiable in $\Phi(s)$ whenever Φ is continuous.*

Assumption 1 disciplines the Receiver's bias by preventing the Receiver from interpreting a higher signal as indicative of a *lower* likelihood of $\omega = 1$. Models such as Grether's (1992) under-

[13] More generally, a *K-coarsening* at thresholds $T = (t_1, \ldots, t_{K-1})$ generates K potential messages where $\Phi_T(s) = k$ for $s \in [t_k, t_{k+1})$.

[14] We later consider the case where the Sender can condition Φ on r.

[15] Formally, $\pi_\Phi(r, k) = \frac{\int_{s:\Phi(s)=k} \pi(r,s)h(r,s)ds}{\int_{s:\Phi(s)=k} h(r,s)ds}$ for message $k = \Phi(s)$.

[16] This is not a behavioral assumption: a utility-maximizing Bayesian serves as a normative benchmark against which we measure the Receiver's deviations.

[17] Standard deviations from Bayesian updating (e.g., confirmation bias or base-rate neglect) generate posteriors that are well-defined and internally consistent for any Φ.

or over-reaction, and correlation neglect Enke and Zimmermann (2019) are consistent with this assumption for a wide range of parameters.[18] However, certain heuristics violate it: for instance, the posterior of a Receiver who distrusts highly confident AI signals ("if the AI is too sure, something must be wrong") may be non-monotonic in the signal. Notice also that Assumption 1 is distinct from the distortion function approach in de Clippel and Zhang (2022), where the biased posterior is a fixed function of the Bayesian posterior. In our setting, two signal pairs that generate the same Bayesian posterior may yield different biased posteriors, since the Receiver's bias may apply to one of the signals and not the other.

To illustrate our framework and main results, we consider a specific, well-documented deviation from Bayesian updating: AI misweighting, the tendency to overweight (or underweight) signals from AI relative to one's own information.

Example (AI misweighting)**.** *Conditional on the state, the signals* (r, s) *are jointly normally distributed with means* $\mu_r^\omega = \mu_s^\omega = 2\omega - 1 \in \{-1, 1\}$*, variances* $\sigma_r^2 \geq \sigma_s^2 > 0$*, and correlation* ρ *which are constant across states and satisfy joint MLRP (*$\frac{\sigma_r}{\sigma_s}\rho < 1$*).*

The Receiver treats the AI signal as if it has variance $\tilde{\sigma}_s^2 = \frac{\sigma_s^2}{\alpha}$ *while correctly perceiving the other aforementioned parameters. This can reflect misperception of the information structure, as well as broader psychological forces that cause systematic under- or overreaction to algorithmic information. When* $\alpha < 1$*, the receiver underestimates that precision—capturing mistrust or algorithmic aversion (Dietvorst et al., 2015). When* $\alpha > 1$*, the receiver overestimates it—capturing automation bias or algorithmic appreciation (Logg et al., 2019). In Appendix 9.2.2 we discuss the relationship of this formulation to the Grether (1992) framework, which has been used to model the under or overweighting of algorithmic information in human-AI contexts (Agarwal et al., 2023).*

The Receiver's log posterior ratio given signals (r, s) *is:*

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

For this example, we assume the Receiver applies their misspecified model consistently when inverting the disclosure mechanism, weighting each signal in the pre-image of k *according to their*

[18] In principle, the conditional correlation between r and s can allow for these biases to generate a reversal in the perceived correlation between ω and r. While technically possible, we view this case as less relevant for capturing the underlying psychology in our context, namely under- and over-reacting to the signal (in the case of Grether (1992)), or double-counting the common information between human and AI (in the case of correlation neglect).

perceived density $\tilde{f}(s \mid r)$,[19] *though we later more generally allow the Receiver's updating rule to vary with the disclosure mechanism* Φ.

Choice. Given mechanism Φ and observed signals $r, \Phi(s)$, the Receiver chooses $a = 1$ iff

$$E_{\tilde{\pi}_\Phi}(U(\omega) \mid r, \Phi(s)) \geq 0.$$

By Assumption 1, we can separate the s-r plane into action and no-action regions:

Result 1 (*Decision Boundaries*)**.** *Fix a monotonic information disclosure mechanism* Φ *associated with a posterior* $\tilde{\pi}_\Phi$*. For each message* k *there exists a* ***decision boundary*** $r(k; \tilde{\pi}_\Phi)$ *such that the Receiver chooses* $a = 1$ *if and only if* $r \geq r(k; \tilde{\pi}_\Phi)$*. The boundary* $r(k; \tilde{\pi}_\Phi)$ *is decreasing in* k*.*

For a given disclosure mechanism, each message induces Receivers whose type is above a certain threshold to act.[20] Notice that decision boundaries suffice to fully characterize the Receiver's actions while imposing minimal structure on the precise updating bias. This will become useful in Section 6, where we estimate boundaries directly from observed Receiver decisions and signal pairs.

In general, the Receiver's decisions deviate from the Bayesian benchmark: $r(k; \tilde{\pi}_\Phi) \neq r(k; \pi_\Phi)$. This discrepancy is central to our framework: since a Bayesian with full information would act according to $r(s; \pi)$—yielding the first-best outcome—the Sender seeks to choose Φ such that $r(\Phi(s); \tilde{\pi}_\Phi)$ is "as close as possible" to $r(s; \pi)$. We say that the Sender can *fully implement* $r(s; \pi)$ if there exists a mechanism Φ such that $r(\Phi(s); \tilde{\pi}_\Phi) = r(s; \pi)$ for all $s \in S$. That is, for any pair (r, s), a Receiver observing $(r, \Phi(s))$ takes the same action as a Bayesian observing (r, s).

To streamline the exposition, we denote the Bayesian decision boundary under full disclosure by $r^*(s) = r(s; \pi)$ and the boundary of the biased Receiver $\tilde{r}(s) = r(s; \tilde{\pi})$, and define their inverse—the threshold AI signal required to induce a known type r to act—as $s^*(r)$ and $\tilde{s}(r)$, respectively.

[19] Given a disclosure mechanism Φ, the Receiver's posterior is $\tilde{\pi}_\Phi(r, k) = \int \tilde{\pi}(r, s), \tilde{\gamma}_\Phi(s \mid r, k), ds$, where

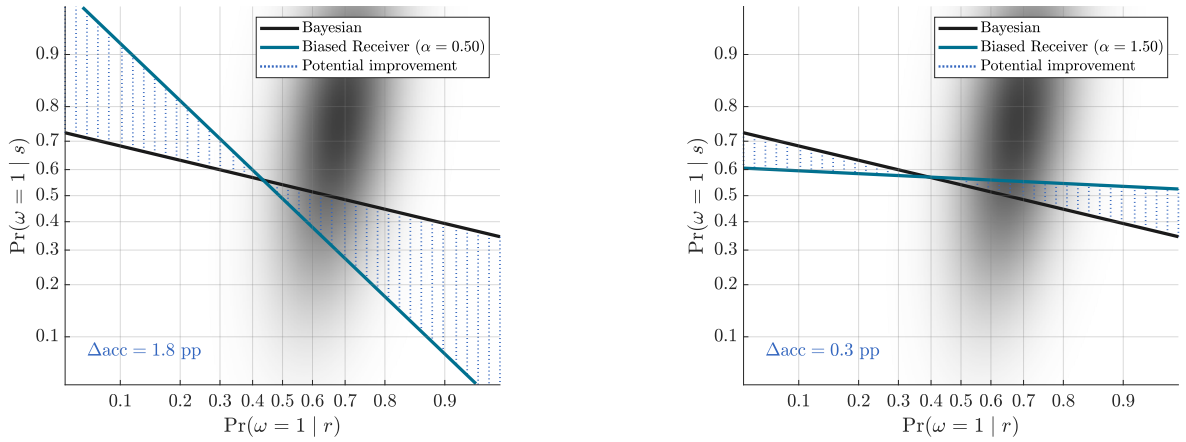
$$\tilde{\gamma}_\Phi(s \mid r, k) = \frac{\tilde{f}(s \mid r)\, \mathbf{1}[\Phi(s) = k]}{\int_s \tilde{f}(s' \mid r)\, \mathbf{1}[\Phi(s') = k]\, ds'}$$

is the Receiver's perceived conditional density of s, given a signal r and a message k from Φ.

[20] This is sometimes called a "cutoff rule" in the literature.

Example (AI misweighting (cont.))**.** *Continuing our example, the misweighting parameter α governs the slope of the decision boundary. Figure 1 plots the Bayesian boundary $s^*(r)$ (black solid line) and the biased boundary $\tilde{s}(r)$ (teal) for two different values of α. Relative to the Bayesian benchmark, $\alpha > 1$ implies over-reaction to s (steeper slope), and $\alpha < 1$ implies under-reaction (flatter slope). The region between the two boundaries—where actions under full revelation are suboptimal—is the scope for improvement, illustrated in dotted blue.*

Figure 1: Decision boundaries: Bayesian vs. AI-misweighting



(a) AI underweighting

(b) AI overweighting

Note: Decision boundaries for different values of α. Each boundary separates the plane into $a = 1$ (above the line) and $a = 0$ (below the line) areas. The Bayesian decision boundary is in black, and the biased is in teal. Decisions disagree in the area between the curves, where there is scope for potential improvement (dotted blue). Potential accuracy gains are denoted Δacc. The joint density is illustrated in shading.

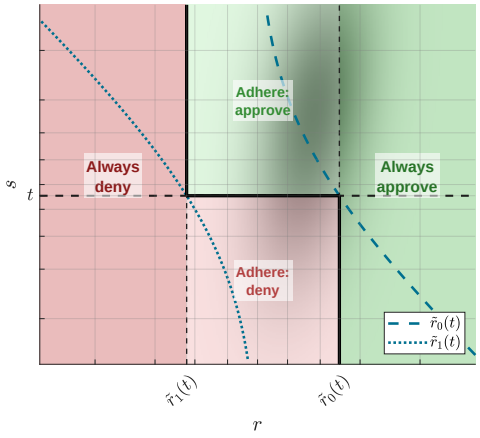
When the Receiver is Bayesian ($\tilde{\pi} = \pi$), fully revealing s is optimal (Proposition 1 in the Appendix). When the Receiver exhibits biased updating, however, the Sender may improve outcomes by choosing a disclosure mechanism that does not fully reveal s.

2.2 Binary coarsening

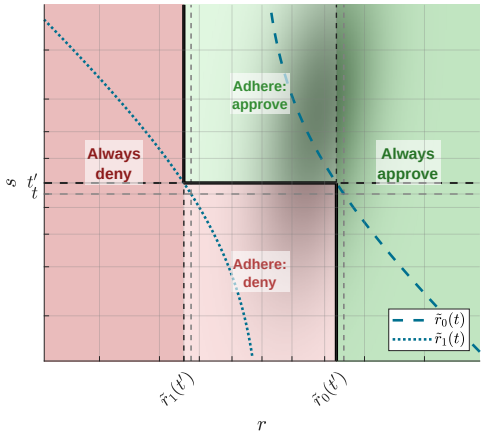
We focus on binary coarsenings. In a binary-action binary-state setting, this is a natural starting point with an intuitive interpretation: the continuous output is transformed into a recommendation that the receiver may overrule. Our main motivation for taking it as a starting point is its potential practical advantage: binary coarsenings are simple, minimizing the confusion that may limit the gains from more sophisticated mechanisms (even if in theory they should perform better).

We refer to the two Receiver's boundaries induced by a binary coarsening as the *adherence curves*, denoted $\tilde{r}_1(t) = r(1, \tilde{\pi}_{\Phi_t})$ and $\tilde{r}_0(t) = r(0, \tilde{\pi}_{\Phi_t})$. Each adherence curve $\tilde{r}_k(t)$ traces out the lowest Receiver type that would choose to act upon observing message k as a function of the threshold t. Given a binary coarsening at threshold t, the *adhering types* are Receiver types whose actions align with the message's implied recommendation, i.e., those with $r \in [\tilde{r}_1(t), \tilde{r}_0(t))$. Figure 2a illustrates the adherence curves and the choices induced by a particular binary coarsening.

Figure 2: Behavior under binary coarsening



(a) Binary coarsening at t.



(b) When t increases to t'

Can binary coarsening improve outcomes over full revelation? If the Sender knows r when choosing Φ, the optimal policy is always a binary coarsening; moreover, it strictly improves outcomes over full revelation when the Receiver's beliefs deviate from Bayes (see Proposition 2 in Appendix).[21] This result is consistent with the Bayesian persuasion literature, such as Kamenica and Gentzkow (2011) and de Clippel and Zhang (2022).

2.2.1 Binary coarsening with unknown r

In practice the Sender often does not observe the Receiver's human signal r when choosing Φ, as is likely in the loans setting. This setup corresponds to the extension discussed in Section VI.A of KG, and relates to broader work in which the Receiver may possess private information over the state (Guo and Shmaya, 2019) or heterogeneous priors over the state (Alonso and Câmara,

[21] The same results hold if r is unanimously irrelevant, i.e., *both* the Sender and the Receiver agree it contains no information about the state.

2016).[22] We first ask *how* to choose the coarsening threshold, before turning to the question of *whether* binary coarsening improves on full revelation.

The Sender would like the Receiver to act whenever the Sender's subjective expected utility, denoted $v(r,s) \equiv E_\pi[U(\omega) \mid r,s]$ is positive. The optimal threshold is given by:

$$\max_{t\in\mathbb{R}} V(t) = \int_{-\infty}^{t}\int_{\tilde{r}_0(t)}^{\infty} v(r,s)h(r,s)drds + \int_{t}^{\infty}\int_{\tilde{r}_1(t)}^{\infty} v(r,s)h(r,s)drds.$$

The first term represents the Sender's utility from Receiver types who choose to approve the loan despite receiving a "No" ($\Phi_t(s) = 0$) message, illustrated in Figure 2a by the region to right of $\tilde{r}_0(t)$ and below threshold t. The second term is utility from Receivers who approve after receiving a "Yes" ($\Phi_t(s) = 1$) message, illustrated by the area right of $\tilde{r}_1(t)$, and above threshold t.

Suppose $V(t)$ is concave. The first-order condition with respect to t yields:

$$\begin{aligned} 0 = \; & \left.\int_{\tilde{r}_1(t)}^{\tilde{r}_0(t)} v(r,t)\, h(r,t)\, dr \;\right\} \textit{Recommendation} \\ & \left.\begin{aligned} &+\tilde{r}_0'(t) \textstyle\int_{-\infty}^{t} v(\tilde{r}_0(t),s)\, h(\tilde{r}_0(t),s)\, ds \\ &+\tilde{r}_1'(t) \textstyle\int_{t}^{\infty} v(\tilde{r}_1(t),s)\, h(\tilde{r}_1(t),s)\, ds \end{aligned} \right\} \textit{Adherence selection} \end{aligned}$$

Intuitively, this condition captures the two effects of changing the threshold, as illustrated in Figure 2b. The first is an *adherence selection* effect: as the threshold changes from t to t', the informational content of the binary message changes and therefore the set of adhering types $[\tilde{r}_1(t), \tilde{r}_0(t))$ shifts to $[\tilde{r}_1(t'), \tilde{r}_0(t'))$, captured by the shift of the vertical lines. The second is a *recommendation* effect: as the threshold changes, the implied recommendations—and hence, actions—change, captured by the horizontal line shift. In other words, changing the threshold affects both *who* follows the recommendation, and *when* action is recommended for these people.

The exact solution to the problem requires knowledge of the adherence curves, which requires either imposing strong assumptions or collecting data on behavior under different thresholds, or both. A simple approximation sidesteps this. Let $\hat{t}$ denote the *AI-only threshold*, defined by the condition $E_r[v(r,\hat{t})] = 0$.

The threshold $\hat{t}$ provides a good approximation for the optimal binary coarsening t^* when two conditions hold. The first is that adhering types are representative of the population wrt to $v(r,\hat{t})$, so that the recommended action for the adhering types is similar to the recommended action

[22] For models where the Sender elicits information from the Receiver, see Kolotilin et al. (2017); Li and Shi (2017). In practice, even if Receiver and Sender incentives are aligned, measuring r can be noisy, distorted, or biased. Moreover, eliciting r and customising the threshold for each individual may be impractical in empirical settings.

of the overall population. Formally, this condition holds when $\mid E_r\left[v(r,t) \mid r \in [r_1(t), r_0(t)]\right] - E_r[v(r,t)] \mid$ is small. This condition is always satisfied when the human signal provides no residual information beyond the AI signal—that is, $\pi(r,s) = \pi(r',s)$ for all r, r'—since $v(r,s) = v(s)$ for all r (whether adhering or non-adhering) and the condition equals 0 by construction.[23] This scenario is increasingly common in a world of big-data and multi-modal models, where AI systems now have access to the same—or even more—underlying data than humans, yet it is still humans who must make the final call.

The second condition is that adherence selection is approximately welfare-neutral, i.e., the total welfare effect of the composition change of marginal Receiver types (near the adherence bound) is small. When these two conditions hold approximately, $\hat{t}$ is approximately-optimal; in the limit as they hold exactly, $\hat{t}$ is the optimal binary threshold.

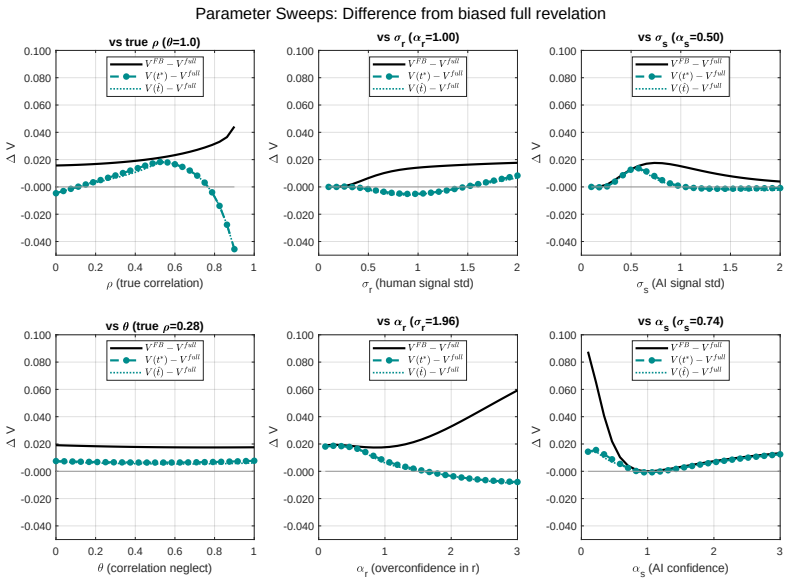
Simulations. How good is the $\hat{t}$ approximation in practice? We assess this with simulations calibrated to our experimental setting. For analytical tractability, we assume $(r,s) \sim N(\mu_\omega, \Sigma)$ and estimate the parameters from our experimental data by Maximum Likelihood, yielding three parameters: $\hat{\sigma}_r = 1.89$, $\hat{\sigma}_s = 0.74$, and $\hat{\rho} = 0.29$. The prior is set by design at $p = 2/3$.[24] We set the baseline bias to AI underweighting ($\alpha = 0.5$), motivated by the AI underweighting we document in our experimental data (Section 9.4.2). Each panel in Figure 3 plots the performance of three benchmarks, relative to full revelation: (i) the gains from binary coarsening at the (approximately optimal) AI-only threshold ($V(\hat{t}) - V^{full}$, dotted teal), (ii) the gains from coarsening at the optimal binary threshold ($V(t^*) - V^{full}$, dashed teal) and, as benchmark, (iii) the gap between biased full revelation and the first-best benchmark, i.e., a Bayesian with full information ($V^{firstbest} - V^{full}$, black solid). The top row varies the parameters governing the joint distribution (ρ, σ_r, σ_s); the bottom row varies the bias parameters (θ, α_r, α_s), allowing the Receiver to neglect the correlation between the signals (parameterized by $\theta \in [0,1]$), misperceive the precision of their own signal (parameterized by $\alpha_r > 0$), or the precision of the AI signal (parameterized by $\alpha_s > 0$).

Two patterns emerge. First, our approximation $\hat{t}$ closely tracks t^* across the parameter space: conditional on coarsening, the AI-only threshold achieves virtually the same utility as the fully optimal threshold. Therefore, the recommendation and adherence selection effects are second order

[23] No residual information is a weaker condition than irrelevance: the former only imposes that r does not affect the Bayesian posterior whereas the latter additionally assumes that the Receiver also ignores it.

[24] We impose the same covariance matrix across states and normalize means to $\mu_0 = (0,0)$, $\mu_1 = (1,1)$ which pins down Σ. See Section 3 for details on the experimental data.

Figure 3: Gains from binary coarsening across parameters



Note: Each panel varies one parameter from the baseline DGP estimated from experimental data ($\hat{\sigma}_r = 1.89$, $\hat{\sigma}_s = 0.74$, $\hat{\rho} = 0.29$, $p = 2/3$) and $\alpha = 0.5$. Black solid: distance from biased full revelation to the Bayesian benchmark. Dashed teal: gain from optimal binary coarsening over full revelation. Dotted teal: gain from the AI-only threshold $\hat{t}$ over full revelation. The zero line corresponds to biased full revelation.

at $\hat{t}$. In a symmetric distribution, this is intuitive. First, the always-acting and never-acting types occupy the opposing tails of the distribution and approximately cancel each other out, making the adhering types representative. Second, the welfare effect of selection is typically small: for a Bayesian receiver, it is zero by definition: boundary types ($r_1(\hat{t})$ and $r_0(\hat{t})$) are indifferent between actions, so (by the envelope theorem) including or excluding them at the margin has zero welfare impact. Under a biased model, boundary types are indifferent according to the wrong model, creating a wedge between Sender and Receiver indifference. However, the bias distorts both boundaries in the same direction: at one, the receiver acts when they should not have; at the other, they do not act when they should have. These errors are opposite in sign and partially cancel, making the overall adherence selection effect small.

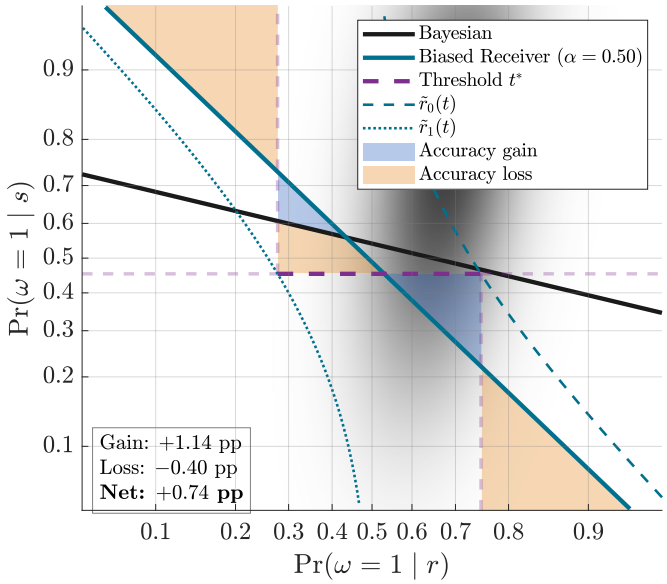
Second, coarsening can yield substantial gains over full revelation. In the top row, which varies the DGP parameters, the gains are largest when the AI signal is moderately precise and the correlation between the two signals is moderately high—in these cases, coarsening essentially closes the gap to the first-best benchmark. In the bottom row, which varies the bias parameters, coarsening helps under both AI underweighting and overweighting. Correlation neglect and own-

signal misweighting can also be improved by coarsening, though in the latter case only for a subset of the range.

In summary, coarsening can, but does not always, improve outcomes—its value depends on the magnitude and direction of the bias relative to the information environment, and whether it helps in a given setting is ultimately an empirical question.

Figure 4 visualises the expected gains and losses from coarsening at $\hat{t}$ under the estimated parameters from our experiment. The figure overlays the Bayesian and Receiver's decision boundaries along with the adherence curves $\tilde{r}_0(t)$ and $\tilde{r}_1(t)$, as pictured previously in Figures 1a and 2a. The figure shades regions where coarsening at $\hat{t}$ yields accuracy gains (blue) or losses (orange) relative to biased full revelation. Gains arise at signal pairs (r, s) where the biased Receiver's action under full revelation disagrees with the Bayesian benchmark, but coarsening corrects it; losses arise where the reverse holds. Gains are concentrated among adhering types, between $\tilde{r}_1$ and $\tilde{r}_0$, where AI underweighting would otherwise cause the largest deviations from the Bayesian decision boundary. Losses arise mainly for non-adhering types who would have correctly followed a strong AI signal that is now pooled with weaker ones.

Figure 4: Simulated gains and losses from binary coarsening under estimated parameters



Note: Gains (blue shading) and losses (yellow shading) from binary coarsening at the AI-only threshold $\hat{t}$, using estimated parameters ($\hat{\sigma}_r = 1.89$, $\hat{\sigma}_s = 0.74$, $\hat{\rho} = 0.29$, $p = 2/3$, $\alpha = 0.5$).

Under the estimated parameters, coarsening yields net gains, suggesting that our empirical setting is one where the theoretical conditions for improvement are met. Of course, this con-

clusion rests on the maintained assumptions—the joint normal signal structure, the misweighting model, and the estimated parameters. Moreover, although our heterogeneity analysis later provides suggestive evidence that AI underweighting is a key bias in our empirical context, deviations from Bayesian decision-making may arise from any combination of misspecifications or biases. Whether $\hat{t}$ is approximately optimal is therefore ultimately an empirical question. We test this in the next sections with an experiment on loan specialists (see Section 3 to skip ahead), comparing decision accuracy under coarsening at $t = \hat{t}$ (*Binary Medium*) to several benchmarks: full revelation (access to the full AI score), human-only (no AI input), and two alternative coarsenings—*Binary High* ($t \gg \hat{t}$) and *Binary Low* ($t \ll \hat{t}$).

3 Experimental Context: Consumer Loan Approvals

Our randomised 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. Financial institutions—and the decision-makers within them—face binary prediction policy problems: approve loans for applicants likely to repay, and deny otherwise. Since repaid loans not only generate revenue through interest and fees but also promote financial inclusion, institutions are incentivised to approve when timely repayment is expected.

In our experiments, we utilise 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, credits, cash), demographic information (e.g., gender, age, car/house ownership, residential information, family, education, and employment), credit card information, normalised credit scores from external sources, amongst others. A key advantage of this particular context is the availability of ground truth: the dataset labels whether each applicant repaid their loans on time. Ground truth for the direct outcome of interest is often unavailable in many other domains, e.g., radiology where "ground truth" is usually defined by expert consensus (Agarwal et al., 2023). Here, our diagnostic standard exactly reflects the state(s) that decision-makers care about, such as loan repayment. Given that our central goal is improving decision-making outcomes in this high-stakes context, having ground truth is of immense value.

A challenge this introduces, however, is the selective labels problem, where the outcome of interest is observed only for a subset of the population. Specifically, repayment outcomes are

available only for applicants who were approved for loans in the original Home Credit setting, and our AI model is trained on this subset. However, this is not a significant concern in our experiment, as 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. Importantly, 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 substantial checks to verify participants understood this prior. Instead, the primary consideration is external validity: whether differences in this application pool (as a result of the initial selected labels) interact with the treatments to skew results. For this to be an issue, it's not enough for selected applications to differ from the general population; these differences would need to differentially affect how loan specialists respond to a full probability AI score relative to a coarsened signal. Moreover, firms deploying calibrated coarsening will typically have access to richer datasets, including data on applicants for whom the repayment outcome is unobserved. With these datasets, selective label issues could be mitigated more directly through reweighting or other causal inference methods.

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 by Fair, Isaac and Company (FICO, 2024). Algorithmic approaches were met with resistance and skepticism, 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 advanced 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) that generate algorithmic signals regarding loan eligibility. These systems summarise recommendations into categories like "Approve/Eligible," "Refer," or "Ineligible" (Freddie Mac, 2025)—effectively a trinary coarsening—which are then provided to lenders to use in decision-making. While DU is not explicitly AI-driven, LPA has been integrated with third-party AI-powered platforms to enhance

underwriting assessments (BusinessWire, 2021; PRNewswire, 2024). Additionally, third-party AI-driven firms have grown increasingly prominent. In an interview we conducted, Brent Chandler, CEO of FormFree, emphasized the growing adoption of AI-generated scores like their product "RIKI," particularly highlighting the value of combining diverse algorithmic assessments with traditional credit scoring methods in order to offer a better understanding of consumers' ability to pay back a loan, especially when they may have limited credit history or FICO scores.

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". Even when a GSE's system issues unfavorable recommendations, lenders may still offer portfolio loans at higher interest rates or with larger down payments. However, it is important to note that the recommendations from these two underwriting systems also affect the ease with which lenders can subsequently bundle and resell these loans to GSEs, affecting algorithmic compliance.[25]

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.[26] 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

[25] This differs from our experimental context, as algorithmic recommendations directly influence lender preferences/incentives through secondary market considerations. In contrast, we assume that the preferences over true/false positives/negatives are not a function of the AI signal.

[26] 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).

lending,[27] 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, aiming to test whether providing signals that are universally coarsened at the model-implied approximately optimal threshold improves decision-making outcomes.

4.1 Survey overview

Figure 5 summarises the design of our experiment, which took place between October 3 and November 4, 2024. We recruited a total of 149 US-based participants aged 18 or older, with a minimum of one year of experience as a loan specialist. Recruitment primarily took place through temporary job postings on LinkedIn and Indeed, supplemented by targeted outreach to members of the National Association of Mortgage Underwriters (NAMU) and the National Association of Mortgage Processors (NAMP), as well as contract hiring on Upwork. Recruiting practicing professionals through industry channels was a deliberate choice: given our focus on improving real-world decision-making outcomes, it was important that participants reflect the population of specialists who would use such tools in practice. A quarter of the loan specialists were recruited through custom screening on Prolific, with stringent screening processes implemented to filter out scammers.[28] Participants recruited over Prolific additionally met the following requirements: they needed to (1) report over one 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. Appendix 9.5.1 shows that our main results remain robust even when Prolific participants are excluded.

Table 1 presents the demographic summary of the 149 participants who completed our experiment. The majority were recruited through Indeed and LinkedIn, including members of NAMU

[27] 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).

[28] Prolific loan specialists in this experiment were recruited after October 21, 2024, when we updated our pre-registration to include stringent both pre-screening and post-screening criteria detailed below.

and NAMP contacted via LinkedIn groups. All participants have loan specialist experience, with an average of 12 years in the field. Additionally, most reported prior exposure to AI, though only a fifth had experience using supervised AI models, such as those commonly employed in loans.

Table 1: Participant Demographics

	Sample Mean or Percentage
Age	44
Female	58.39%
Bachelor's Degree or Above	61.07%
Income	107,836
Loan Specialist Experience	100.00%
Loan Specialist Years	12
Bank / Credit Union Experience	34.23%
Applied to Loans Before	34.90%
Loan Amount (USD$)	193,615
Used Supervised AI Before	20.81%
Never Used any AI	24.16%
Recruited through LinkedIn (NAMU/NAMP inclusive) and Indeed	64.43%
Recruited through Upwork	9.40%
Recruited through Prolific	26.17%

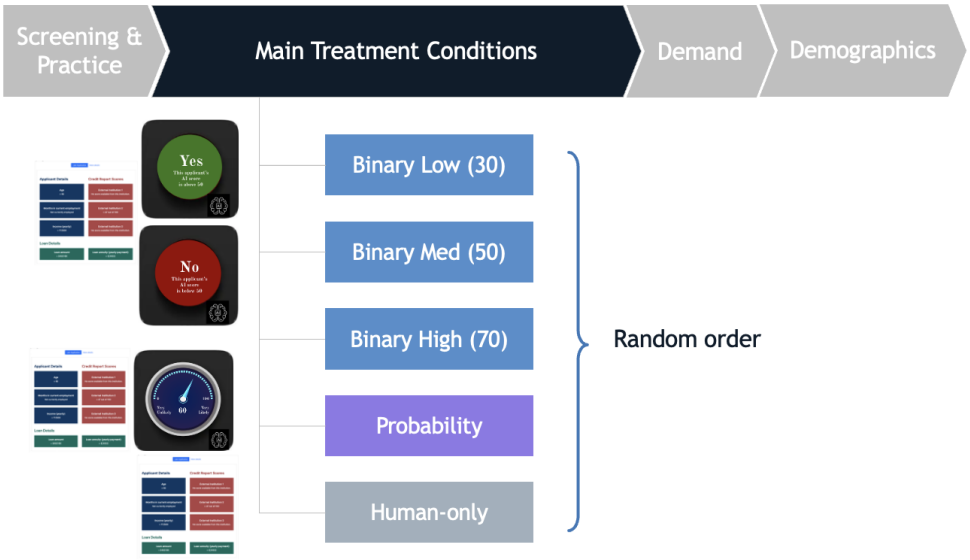
Once through with screening, qualified participants were directed to complete a 60 minute survey hosted on the Otree platform via Heroku, comprising 60 loan decisions and additional questions on demand for AI, biases, demographics, amongst others, detailed in Appendix 9.3.2.

4.2 Loan application decisions

The main thrust of our survey presents participants with binary prediction problems: approve the loan to applicants expected to make timely payments, and deny the loan otherwise. Participants are shown part of a real loan application sourced from Home Credit. The experiment consists of five treatment conditions described in the section below, each containing ten decisions presented in randomised order. Depending on their assigned treatment group, participants may also receive a signal from an AI model trained on a loan repayment dataset from Home Credit. After reviewing this information, they are directed to a landing page where 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.]"

Figure 5: Main experiment survey flow



- "Do you want to approve this applicant for a loan? [Approve / Deny]"

On this page, participants can view both the application details and the AI signal (if applicable) and take as much time as they need to submit their answers.

After a practice round of 10 decisions (without AI assistance), participants evaluate 50 loan applications in the same manner as part of each of the five treatment arms. Before making decisions, they are briefed on several key aspects of the decision-making environment: (1) **the prior**, or the share of applicants in the population expected to repay; (2) **their performance in the practice rounds**; (3) **details of the AI model**, including its training data, score interpretation, performance metrics, and signal density conditional on repayment outcomes; and (4) **details on the assisting AI's decision rule and behaviour** (e.g., threshold and conditional performance for *Binary* formats, or average scores by repayment status for *Probability* scores). Participants can access this information throughout the experiment via a "More details" tab on the right-hand side. The full experiment instructions can be found in Appendix 9.3.1.

Incentives are structured to balance the cost between false positives and negatives: a bonus of \$0.10 is awarded for each loan correctly denied (for default applicants) and each loan correctly approved (for on-time payers). Participants are explicitly told this incentive structure and undergo checks for understanding. Since equalising the costs of false positives and negatives is equivalent to maximising accuracy, we focus mainly on diagnostic accuracy as our main outcome. Elicited posteriors are additionally incentivized via a binarized scoring rule (Hossain and Okui, 2013).

Participants are expected to earn $4 in bonuses on average, in addition to a base rate which varied from $16 (Prolific) to $40-60 per survey (Indeed, LinkedIn).

4.2.1 Loan application details

Participants are shown several key variables as part of a loan application (see Figure 20 in the Appendix for an example), 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 debt-to-income ratio, months in employment (relative to age), as well as external credit scores. Variables include: loan amount, loan annuity, age, months in current employment, income (yearly), and normalised credit scores from three sources, each ranging from 0–100, with higher values indicating greater creditworthiness.

4.2.2 AI model details

We train an AI model on the Home Credit dataset of 300,000 loan applicants and their repayment outcomes, as described in Section 3.[29] The AI model has access to all details in the loan application available to human participants, as is explicitly made clear in the experimental instructions (see Appendix Figure 16). For each loan application, the AI generates a score from 0-100, which we then convert to the Bayesian posterior probability of repayment (based on the AI score and prior).

4.3 Treatment conditions

In our main experimental variation, we manipulate the type of AI assistance participants receive (if any). The treatment conditions are structured as follows: {*Human-only, Probability, Binary Low (30), Binary Medium (50)*, and *Binary High (70)*}.

Participants see the loan application in all conditions. In the *Human-only* condition, participants make decisions based solely on the loan application information, without any AI assistance. In the *Probability* condition, they additionally see the numerical output of the AI model from 0-100, and are told it is the AI's estimate of the likelihood the person would repay the loan on time. In each of the three *Binary* conditions, participants are given one of two messages: "Yes, this applicant's AI score is above [threshold]," or "No, this applicant's AI score is below [threshold],"

[29] We use a top-performing AI model from a public competition on the same dataset. The model employs Bayesian optimisation to automate the tuning of hyper-parameters, resulting in a cross-validation AUC-ROC score of 0.79.

depending on the specified threshold for that condition (30, 50, or 70). Note that in this setting, the model-implied approximately optimal binary treatment is the *Binary Medium* condition with the threshold of 50.

To increase power, all participants complete all five decision blocks, which appear in randomised order. The order of the underlying loan applications are also randomised, and thus cases typically differ across treatment conditions for different individuals. As a robustness check, we report block-level results in Appendix 9.5.3 and show that *Binary Medium* (50) consistently outperforms *Probability* across all five blocks, including in the first round, which is a 'pure' comparison due to randomised block order across individuals, uncontaminated by prior exposure.

4.3.1 Other sources of randomisation

We also cross-randomise our treatments with case-level randomisation of:

- **Order of information presentation:** Whether the AI signal or loan application shown first {AI (10%), Loan (90%)}
- **Elicitation of posterior beliefs before observing the second signal:** Whether posterior is additionally elicited before the second signal shown {Elicit after only (10%), Elicit both before and after (90%)}

This design enables us to capture the posterior the loan specialist forms after just seeing the loan application (i.e., human signal) in cases where: (i) the posterior is elicited after both the first and second signal, and (ii), the loan application is presented first. Randomising these conditions also allows us to test for potential anchoring effects. While our model does not assume or rule out order effects, it is important we check for it empirically. If participants systematically respond differentially to treatments depending on whether they see the loan application or AI signal first, or whether they report their posterior before or after seeing a second signal, this could affect our estimates of how coarsening affects belief updating and decision accuracy. In Appendix 9.5.2, we show that neither the order of AI signal presentation nor the timing of posterior elicitation systematically affects participants' decisions or final posteriors.

4.4 Empirical strategy

We focus on loan approval accuracy as our main outcome, given that we incentivise for accuracy. Our main specification is as follows:

$$Y_{ilt} = \beta_0 + \sum_t \beta_t \text{Treatment}_{il}^t + \epsilon_{ilt} \tag{2}$$

where Y_{ilt} denotes the outcome variable for case i decided by loan officer l in treatment arm t, $\text{Treatment}_{il}^t \in \{0, 1\}$ indicates that the case i decided by loan officer l is in the treatment group t, setting the *Binary Medium* treatment as the reference base level. We estimate the regression with and without fixed effects. Identification is established by random treatment assignment. Our main objects of interest are the β_t coefficients. In particular, we test the (pre-registered) hypotheses:

- Whether the *Binary Medium* arm performs differently from the *Probability* treatment, i.e., testing the null hypothesis that $\beta_{\text{probability}} = 0$
- Whether the *Binary Medium* arm performs differently from the *Human-only* treatment, i.e., testing the null hypothesis that $\beta_{\text{Human-only}} = 0$

The above tests were pre-registered in our pre-analysis plan (AEARCTR-0013716). We also documented in the plan that we expected that *Binary Medium* would perform better than both the benchmarks above, as well as the *Binary High/Low* treatment conditions, although we remained agnostic to whether *Binary High/Low* would outperform the *Human-only* or *Probability* treatment.

5 Main Experiment Results: Universal Calibrated Coarsening Improves Outcomes

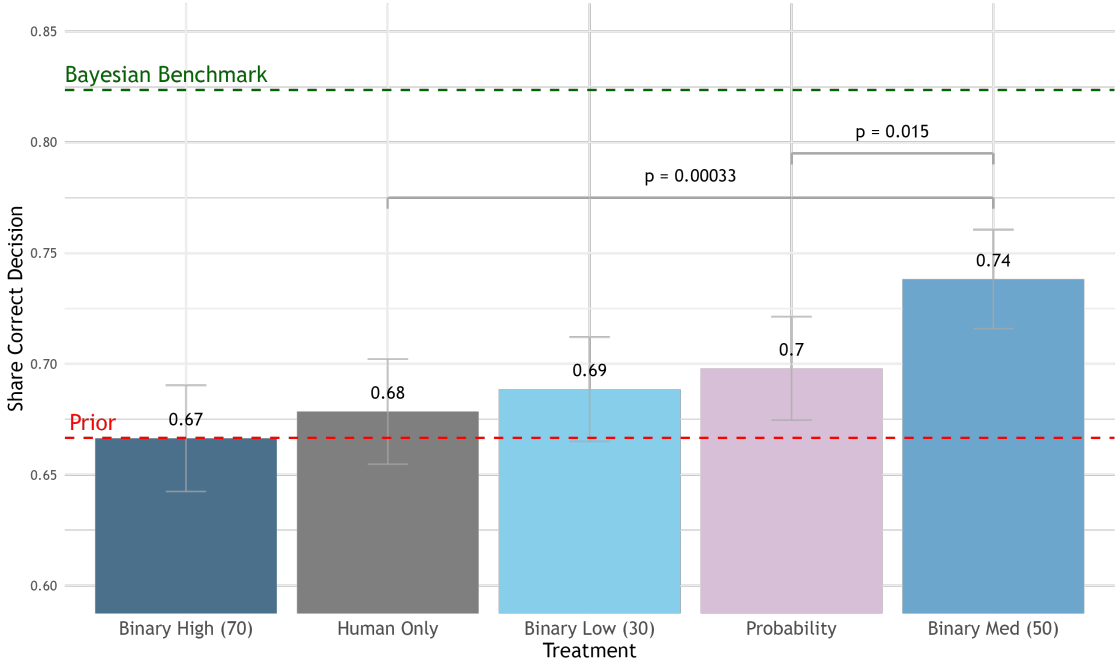
We present evidence from the main experiment that providing loan specialists with signals coarsened at the model-implied approximately optimal binary threshold leads to better decision-making outcomes than human judgment alone, providing the full AI output, or other binary AI coarsenings.

5.1 Main results

5.1.1 Impact of coarsening on decision-making outcomes

Our main finding provides empirical evidence that calibrated coarsening improves decision-making. Figure 6 clearly demonstrates that the *Binary Medium (50)* treatment outperformed all other treatments, achieving an accuracy of 74%. Table 2 complements this by presenting regression results from Equation 2, using the *Binary Medium (50)* treatment as the reference level.

Figure 6: 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; dashed green line marks the Bayesian benchmark. p-values indicate significance of pairwise differences relative to *Binary Medium (50)*. Error bars represent 95% confidence intervals.

Loan specialists made, on average, 4 percentage points more correct decisions in this condition than in *Probability*. This effect remains significant even after controlling for individual participant, loan application, and block order fixed effects (Column 2).

To contextualise the magnitude of these effects, we see that performance under this condition far exceeds that of the prior of 66% (dotted red line)—the best achievable outcome without information, obtained by approving every case—as well as the *Human-only* condition. Moreover, it substantially narrows the gap to the Bayesian benchmark of 82%, which represents the accuracy a fully Bayesian agent would achieve with complete access to AI probability signals (dotted green

line).[30] Moreover, while providing participants with the full AI score in the *Probability* condition leads to some improvement over *Human-only*, the difference is not statistically significant. These results underscore the effectiveness of calibrated coarsening—closing more than one-third of the gap between the *Probability* condition and the Bayesian benchmark, and almost half the gap between the *Human-only* condition and the Bayesian benchmark.

Table 2: Decision-making outcomes under different treatment conditions

	Dependent variable: Share of Decisions Correct			
	No Fixed Effects	*Fixed Effects (Individual + Case + Block Order)*		
	(1) Coefficient	(2) Coefficient	RI p-val (2-sided)	RI p-value (left)
Human-Only	-0.060	-0.058	0.001	0.001
	(0.016)	(0.015)		
Probability	-0.040	-0.030	0.055	0.032
	(0.018)	(0.017)		
Binary High (70)	-0.072	-0.058	0.001	0.001
	(0.017)	(0.016)		
Binary Low (30)	-0.050	-0.040	0.015	0.005
	(0.016)	(0.014)		
Constant	0.738			
	(0.012)			
Observations	7450	7450		
R-squared	0.003	0.307		
Adj. R-squared	0.002	0.287		
Omnibus F-test		5.600	0.001	

Note: Regression estimates per Equation 2 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-level fixed effects. Asymptotic standard errors clustered at the individual level in parentheses. Randomisation-inference (RI) p-val columns report Fisher p-values based on 1,000 design-consistent permutations of treatment–block assignments within participants.

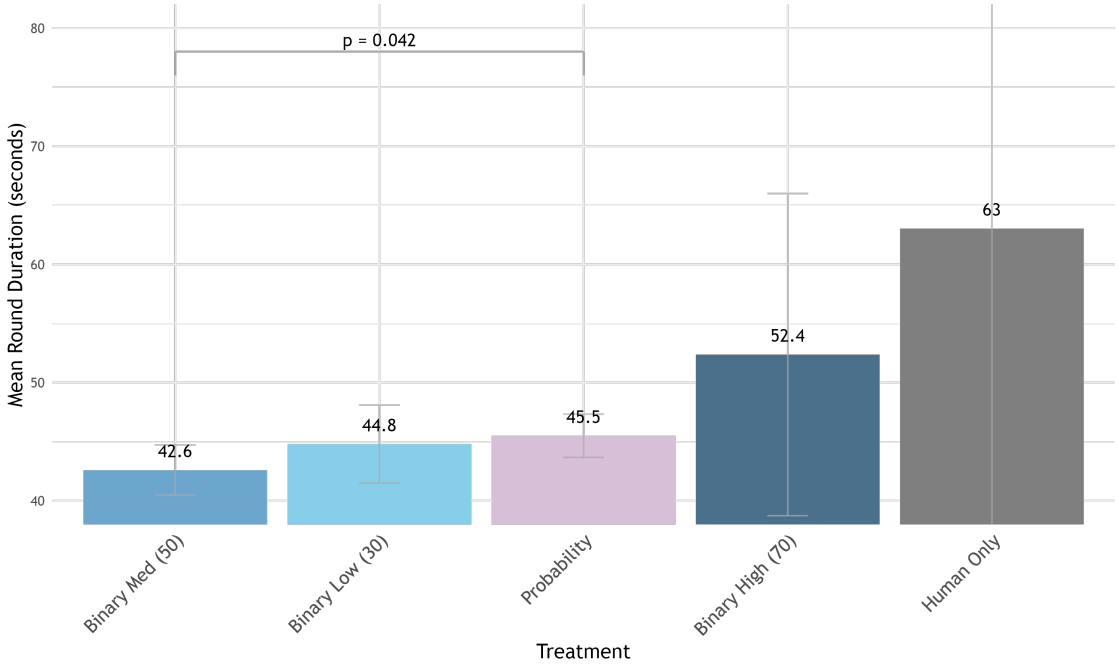
The other comparisons underscore the importance of *calibrated* coarsening as a solution: coarsening at the arbitrary alternative thresholds *Binary High (70)* and *Binary Low (30)* yields accuracy rates comparable to the *Human-only* condition, suggesting that not all forms of coarsening are beneficial. By selecting an optimised threshold, as in the *Binary Medium (50)* condition, decision-making performance improves.

[30] Note that this is also the AI-only benchmark, as the human does not have residual information over the AI here. The difference between the prior and the Bayesian benchmark is the value of AI information in our context.

5.1.2 Impact of coarsening on time spent

In addition to improving accuracy, coarsening AI signals can significantly reduce decision time—a critical consideration in real-world applications where experts' time is valuable. Figure 7 shows the average decision time across conditions. We find that participants in the *Binary Medium (50)* condition spent 42.6 seconds on each application, making decisions approximately 7% (around 3 seconds per application) faster than those in the *Probability* condition. This suggests a meaningful reduction in cognitive load, as individuals no longer need to interpret and calibrate a continuous score but instead act on a clearly actionable binary recommendation.

Figure 7: Average decision time across treatment conditions



Note: Bars represent mean decision time (in seconds) across treatment conditions. Participants in the *Binary Medium (50)* condition made decisions significantly faster than those in the *Probability* condition (p = 0.042). Error bars indicate 95% confidence intervals.

This time-savings is particularly important in high-throughput decision environments such as financial underwriting or medical triage, where expert bottlenecks are costly and reducing time per case—even marginally—can unlock substantial operational gains. Businesses may sometimes even care more about throughput than marginal improvements in accuracy, and this is where AI often delivers the most tangible value: by accelerating decisions without sacrificing (and in this case, improving) performance.

We also observe substantial time savings compared to the *Human-only* condition, although with

considerable variance in the latter group. Some participants took much longer to reach decisions when unaided by AI, while others were relatively quick. While our study was not powered to detect differences in time spent, results suggest that AI tools not only boost performance but can also streamline workflows, representing large operational gains.

5.1.3 Demand for coarsened AI

It is not enough for a solution to be effective. An important practical question remains: do people adopt it? Moreover, do people choose the signals that are most effective for them?

At the end of this experiment, participants were given the choice between three options: receiving no AI assistance, the binary AI signal (at their preferred threshold), or the full probability score. As shown in Table 3, less than half of the participants chose to receive AI assistance in the form of the probability score, while others preferred less than full information, either opting for no AI assistance[31] or preferred a binary signal instead.

These preferences are mirrored in participants' willingness to pay (WTP), which we elicited in an incentivised manner. Those who declined AI assistance were, on average, willing to pay less than 10% of the 10-cent per-case bonus (i.e., $< \$0.01$) for binary signals, and would have to be paid to use a probability score. In contrast, those who opted for binary and probability AI assistance were willing to forgo non-trivial amounts: 38% (3.8 cents) and 25% (2.5 cents) of their per-case bonus to access them, respectively—which is *higher* than the gains in accuracy that these types of AI assistance brings.

Table 3: Willingness to Pay and Highest Payoff Treatment, by Choice of AI Assistant

		Highest Payoff Treatment				
Choice of AI	Total (#)	Human Only (% total)	Binary (% total)	Prob (% total)	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	34	3.8 (4.10)	2.4 (4.04)
Probability	69	22	50	53	3.1 (4.06)	2.5 (3.72)

Note: Summary of participants' choice of AI assistant, their highest payoff treatment, and mean willingness to pay (WTP), expressed in cents per case.

There is some suggestion that people are selecting into the types of signals that are most effec-

[31] "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 something we capture in our theoretical framework.

tive for them—even though they did not receive any feedback on how they did in each of the different treatments. The majority of participants for whom the highest payoff treatment was *Human-only* chose not to use any AI,[32] 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 of them preferring AI assistance in the form of a probability risk score instead of a binary.

In Appendix Table 5, we also explore how demographic characteristics relate to AI assistant choice. Notably, 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.

Before asking participants to choose between the three AI assistance types, we also elicited their ideal binary threshold—i.e., the score above which they would want the model to output a "Yes, this applicant's score is above [threshold]" recommendation. Figure 24 in the Appendix shows the distribution of these thresholds. Choices are skewed to the right of 50: many participants selected thresholds above the best-performing decision threshold. This pattern may reflect a tendency to emulate a "cherry-picking" strategy that loan specialists are often trained under, prioritising identification of applicants highly likely to repay over eliminating lemons.

5.2 Understanding how participants make decisions

Given that what we care about is decision-making, we want to further understand how our participants are making these decisions—for instance, to what extent participants are following the AI recommendations—and what factors are driving these outcomes.

One important aspect we verify is whether loan specialists behave consistently with the decision rule implied by the incentive structure. Since correct approvals and denials are rewarded equally, participants should approve when their posterior exceeds 50% and deny otherwise. Of course, participants may also be influenced by professional norms or their own sense of what constitutes a "good" lending decision, which could lead them to apply different implicit thresholds. We find little evidence of this: Appendix Figure 26 shows that pooled approval rates transition sharply around the 50% posterior threshold, and Appendix Figure 25 shows that for most individ-

[32] 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.

ual participants, approve/deny decisions are cleanly separated around this threshold. This suggests that most deviations from the Bayesian benchmark stem from belief formation rather than decision noise or miscalibrated action thresholds.

Having established that participants generally apply consistent action thresholds, we examine how they respond to variation in the binary AI thresholds. As shown in Appendix Figure 27, approval rates increase systematically with the binary threshold: higher thresholds increase approvals under a "Yes" recommendation whilst weakening the force of a "No" and reducing denials. 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 defer to, AI advice (albeit erroneously in this case, where they don't have any residual information over the AI).

5.2.1 Estimating pooled decision boundaries

If participants are not simply following the AI signal, how are they actually making decisions? To investigate this, we estimate participants' pooled decision boundaries using our experimental data. Note that figures 8c-8d present the empirical analogues of the decision boundary graphs in Figures 1a and 2a in the model section, estimated from observed choices.

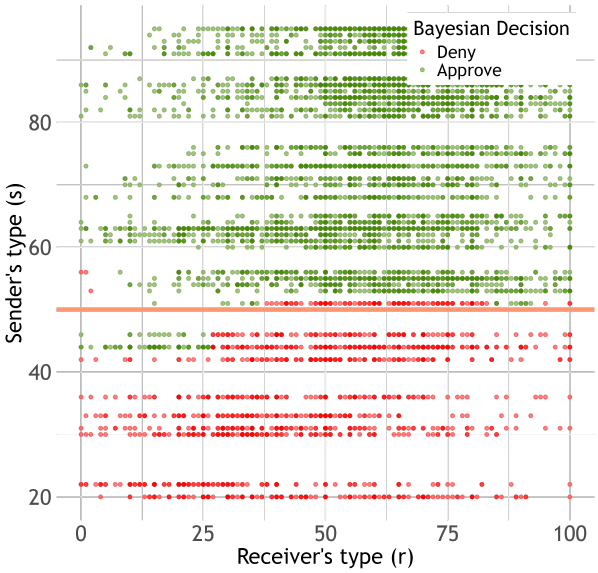
First, we estimate the Bayesian decision boundary—the best a decision-maker can do given their information—represented by the orange line in Figure 8a.[33] Green and red dots represent cases where the Bayesian would approve and deny, respectively, and the orange line demarcates the decision boundary between them. Ex ante, we did not know whether humans would have residual information over the AI—even with access to the same or fewer variables, humans might be better at mapping variables to state. Empirically, however, the boundary is approximately flat at $r = 50$, suggesting that in this setting, humans have approximately no residual information over the AI. A Bayesian decision-maker would perfectly follow the AI signal, approving any case where the AI score exceeds 50 and denying otherwise.[34]

Next, we estimate the decision boundary of the loan specialists in our experiment. To do this,

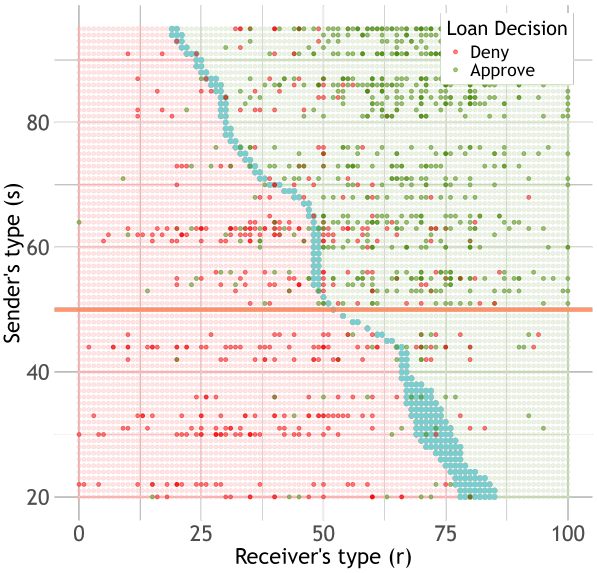
[33] For all cases where we have elicited the human signal over the loan application (r)—$\sim 81\%$ of cases, since 90% of cases present the loan application first (before the AI signal), and we independently randomise whether the posterior is elicited after the first signal in 90% of cases—we compute the Bayesian posterior using the joint distribution of r and s conditional on the state ω. We then apply the thresholding rule: a Bayesian would approve if their posterior probability is greater than 0.5.

[34] Slight deviations arise for two reasons: (i) although the AI observes more information than the human for any given application, humans may retain residual information about the *mapping* from variables to state; and (ii) more likely, estimation noise from sparse data in some parts of the distribution.

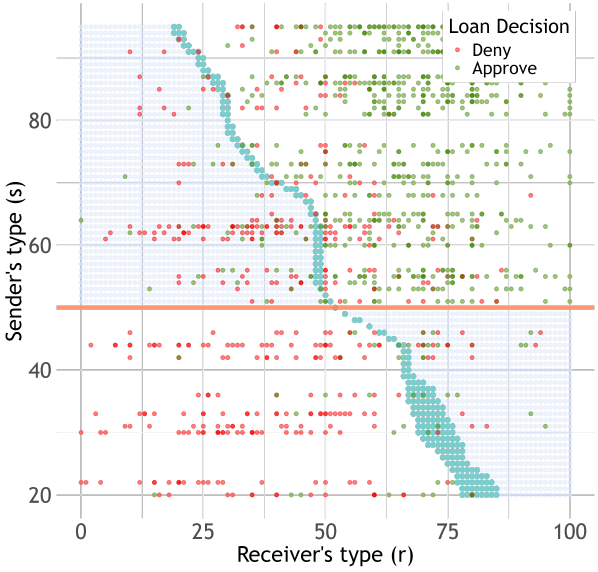
Figure 8: Pooled Decision boundaries, Visualised Using Experiment Data



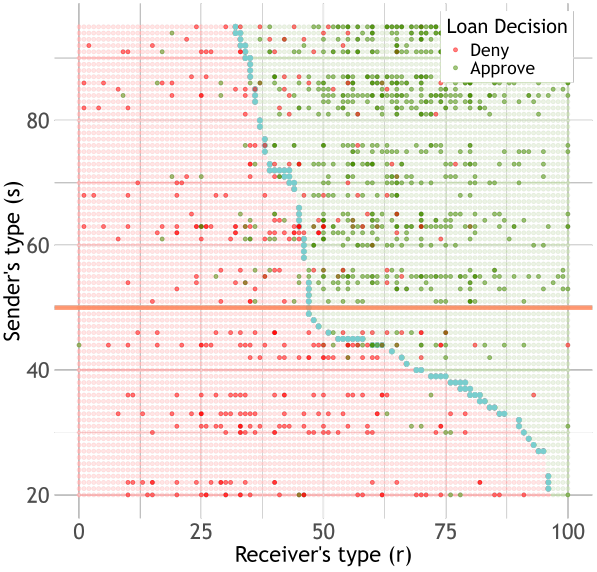
(a) Bayesian Decision Boundary
(All cases with elicited r, pooled)



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



(c) Scope for Outcome Improvement
(Probability Treatment, Pooled)



(d) Loan Specialist Decision Boundary
(Binary Medium (50) Treatment, Pooled)

Note: Panels show estimated decision boundaries for a Bayesian benchmark (top left), pooled loan specialist decisions under the *Probability* (top right and bottom left) and *Binary Medium (50)* (bottom right) treatments. The blue area in Panel (c) illustrates the scope for improvement by showing where participants' decisions diverge from the Bayesian-optimal boundary.

we pool together all decisions under the *Probability* condition where we also elicited the human signal. This allows us to examine how loan specialists integrate both the human and probability AI signals in their decision-making process. Using a Nadaraya-Watson kernel with bandwidth selection via cross-validation, we estimate the probability of approval, $\Pr(a^r = 1 \mid r, s)$. To ensure the estimated probabilities are monotonic with respect to both r and s, we apply the rearrangement procedure described in Chernozhukov et al. (2009). Finally, we solve for pairs of r and s satisfying $\Pr(a^r = 1 \mid r, s) = 0.5$, yielding the blue estimated decision boundary in Figure 8b.

There is a clear deviation from the orange Bayesian decision boundary: the blue decision boundary of pooled loan specialists is steeper, indicating that the loan specialists place greater weight on their own (human) signal compared to a Bayesian decision-maker. Ideally, to attain better decision-making outcomes, loan officers should approve cases above the Bayesian line and deny cases below. Figure 8c highlights the scope for improvement in loan specialists' decision-making, with the blue area representing the discrepancy between the Bayesian decision boundary (orange line) and the estimated decision boundary of loan specialists (blue curve). This misalignment suggests opportunities to refine decision-making by structuring signals in a way that brings loan specialists closer to Bayesian-optimal behaviour.

Examining decisions under the *Binary Medium (50)* treatment reveals that coarsening the AI signal can indeed improve decision-making outcomes. Figure 8d illustrates how thresholding the AI signal at the model-implied optimised threshold ($s = 50$) helps participants adhere more closely to the AI recommendation within the range of $r = 35$ to $r = 75$.[35] Intuitively, because participants systematically underweight the AI signal relative to their own, pooling a strong AI score (e.g., 100%) with a marginal one (e.g., 51%) can increase the perceived strength of the latter without altering the recommendation from the former. As a result, individuals are more likely to follow the AI when their own signal is sufficiently low (e.g., below $r < 90$), with similar logic for $r > 40$. Outside this range, however, participants continue to rely on their own human signal, which—when strong enough—can lead them to override the AI recommendation.

[35] See Figure 11e: we estimate a Bayesian HLM which allows us to back out the median receiver type (r) corresponding to each binary message. Appendix Figure 28 provides supplemental evidence: it plots the difference in approval probability between *Binary Medium* and *Probability* treatments by underlying AI score. Participants under *Binary Medium* are more likely to approve when the AI score exceeds 50 (following the "Yes" recommendation), with the largest effects for marginal AI scores.

5.2.2 Where loan specialists are performing well

To better understand how coarsening affects different types of decisions, we examine where specialists are good at or particularly struggle in approving and denying loans. Across treatments, loan specialists are generally better at correctly approving than correctly denying applicants. The *Binary Medium* treatment primarily improves performance by reducing false negatives—helping participants identify and approve creditworthy applicants who might otherwise be denied. This effect is especially pronounced when participants' own signal falls in the uncertain range (34–66), where decision-making is more ambiguous. Full results broken down by true/false positives and negatives, stratified by AI and human signals, are reported in Appendix Tables 26–28.

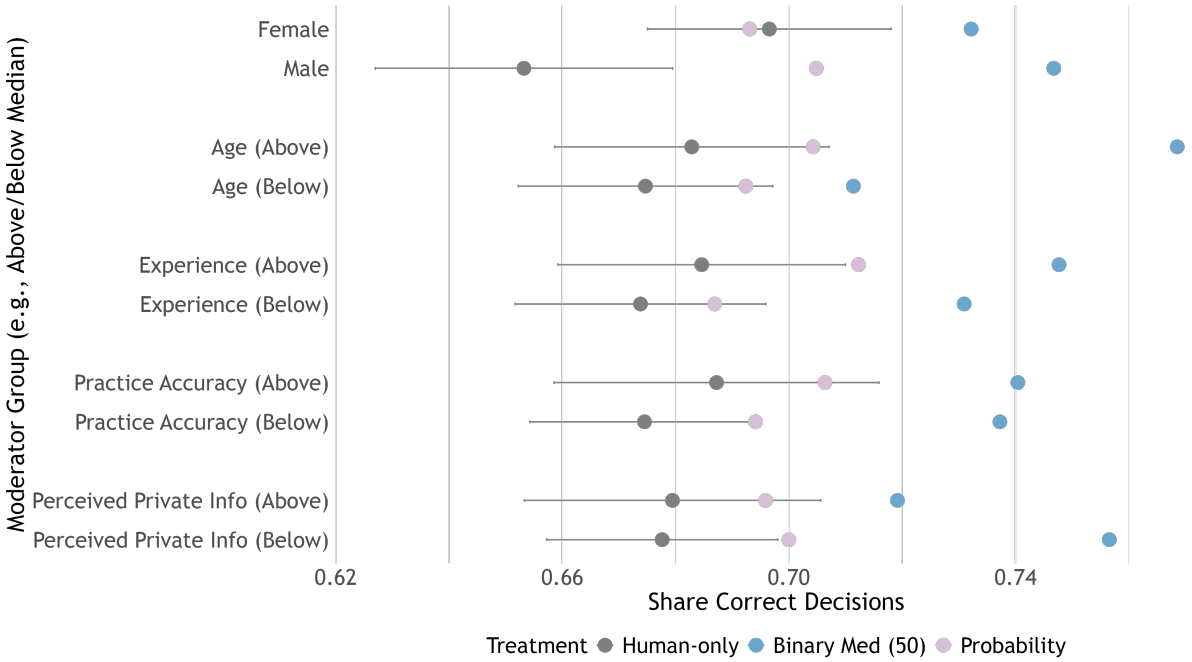
5.3 Heterogeneous Treatment Effects (HTEs)

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). Figures 9 and 10 show the average share of correct decisions across subgroups defined by demographic and bias-related moderators. Corresponding regressions, based on Equation 2, and including individual, case-level, and block order fixed effects, are reported in Appendix 9.4.2.

5.3.1 HTEs by demographics

First, notice that in nearly all of these cuts, *Binary Medium* outperforms *Probability*, which in turn outperforms *Human-only*. Gender emerges as a key moderator, with men benefiting significantly more from AI assistance. The *Binary Medium (50)* treatment increases men's accuracy by 9.8 percentage points over *Human-only*, compared to 5.2 percentage points for women. Men also gain substantially from the *Probability* treatment (6.1 percentage points over *Human-only*), whereas women see essentially no improvement from *Probability*. These differential gains are partly explained by baseline disparities: women exhibit higher accuracy (69.7%) than men (65.3%) in the *Human-only* condition, suggesting that men, starting from a lower baseline, have more room to benefit from AI assistance. The pronounced gender gap in AI-driven performance gains also aligns with existing literature suggesting that men may be more willing to adopt, adapt to and leverage AI-generated recommendations effectively (Bick et al., 2024; Carvajal et al., 2024; Humlum and Vestergaard, 2024; Liu and Wang, 2024), though some evidence suggests that this gender gap in

Figure 9: HTEs by Demographics



Note: Each point represents the average share of correct decisions for a given subgroup under three treatment conditions: *Human-only, Binary Medium*, and *Probability*. Subgroups are defined by gender (male/female) as well as median splits on the other variables. 'Above' and 'Below' refer to participants above or below the median for each variable. Error bars represent 95% confidence intervals, displayed only for *Human-only* for simplicity (note that most comparisons are not powered to be significant). Formal tests of differences in treatment effects across subgroups are reported in Appendix 9.4.2.

adoption has narrowed substantially over time (Chatterji et al., 2025).

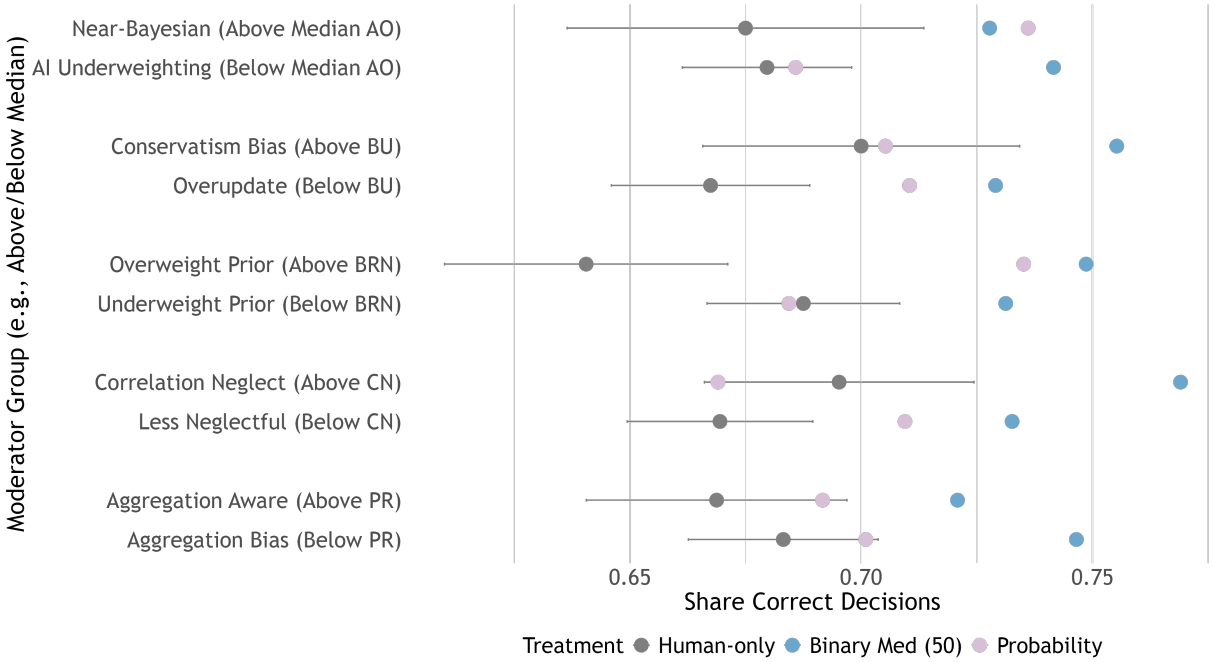
Other demographic factors also matter. While age does not meaningfully moderate treatment effects, older loan specialists—those above the median age—perform somewhat better across all treatments, with a 0.8 percentage point advantage in the *Human-only* condition. Experience shows a similar pattern: more experienced specialists have slightly higher baseline accuracy (a 1.1 percentage point gap in *Human-only*) and directionally larger gains from AI treatments, though the difference in treatment effects is not statistically significant. The literature on whether experience complements or substitutes for AI assistance is mixed, with some evidence that less experienced workers benefit disproportionately (Dell'Acqua et al., 2023; Brynjolfsson et al., 2025) and other evidence that domain expertise complements AI (Otis et al., 2024). Similarly, those with below-median practice accuracy (proxy for baseline performance) benefit slightly more from coarsened signals.

Perceived private information also influences responsiveness to AI. Those with above- and below-median perceived private information exhibit nearly identical baseline accuracy in the *Human-*

only condition (68.0% vs. 67.8%). However, those with below-median perceived private information benefit more from AI assistance, particularly *Binary Medium (50)* (8.5 vs. 5.6 percentage point gains). This pattern is intuitive—individuals who (correctly) recognise their limited private information are more receptive to AI signals, making them more likely to incorporate AI assistance effectively.

5.3.2 HTEs by cognitive biases

Figure 10: HTEs by Bias Elicitations



Note: Each point represents the average share of correct decisions for a given subgroup under three treatment conditions: *Human-only, Binary Medium*, and *Probability*. Subgroups are defined by median splits: 'Above' and 'Below' indicate participants above or below the median. Error bars represent 95% confidence intervals, displayed only for *Human-only* for simplicity. Formal tests of differences in treatment effects across subgroups are reported in Appendix 9.4.2.

Baseline cognitive biases also shape treatment effects, offering suggestive evidence on which types of bias are most likely to be at play in this decision context. 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 9.3.3.

Heterogeneity by algorithmic underweighting is striking. While baseline accuracy in the *Human-only* condition is similar across groups, outcomes diverge sharply once AI input is introduced. In this setting, where there is no residual human information over the AI, there is no scope for AI overweighting (AO). Instead, the relevant concern is algorithmic underweighting, as specialists may rely too heavily on their own information rather than following the AI's signal (the best approach in this context). Among agents who underweight AI (below median AO), the *Binary Medium* treatment outperforms *Probability*. Strikingly, this pattern reverses among those who are more Bayesian, who perform better with the *Probability* treatment. This sharp reversal aligns with the conceptual insights from the simulated example in Section 2: when specialists underweight the AI signal, coarsening pools signals together and helps them weight the AI more heavily.

Base-rate neglect (BRN) also plays a role: those who face above-median BRN start with lower accuracy in the *Human-only* condition and experience correspondingly larger gains to AI assistance. They also exhibit a statistically significantly larger treatment effect from *Binary Medium* (relative to *Probability*) than those with below-median BRN (see Appendix Table 22), consistent with the idea that individuals who discount priors more highly (i.e., higher BRN) update more in response to signals. The *balls-urns* (BU) measure also shows a stark contrast: those with above-median BU—who exhibit close-to-Bayesian updating or conservatism bias—have much higher baseline accuracy in the *Human-only* condition (12.8pp higher) than those below median, who tend to over-update. However, treatment effects are stronger for the latter group, which aligns with expectations, as individuals who over-update are more sensitive to AI signals. *Correlation neglect* (CN) and *aggregation bias* (PR) show a distinct pattern: individuals with higher bias perform worse under the *Probability* condition, but this gap is more than closed by coarsening, suggesting that these biases may also be at play in this context.

Taken together, these findings suggest that suboptimal use of AI is not driven by any single bias, but by a mix of behavioral deviations. Calibrated coarsening offers a robust solution that mitigates the effects of multiple biases at once. This motivates our next step: rather than making structural assumptions about any particular bias, we adopt a data-driven personalisation strategy that uses past actions to identify the most effective coarsening policy.

6 Simulating the Gains From Personalised Policies

The previous section documents substantial heterogeneity in how specialists respond to different types of AI assistance, suggesting that a one-size-fits-all approach may leave additional gains on the table. While a universal coarsening rule (such as the theory-implied *Binary Medium*) improves average performance, the presence of systematic heterogeneity raises the possibility of further gains through personalisation. In this section, we develop a Bayesian model of decision-making to estimate individual decision boundaries from past data, with the goal of simulating the value of AI-assistance policies which we then tailor to each specialist. Our aim is to design an approach that can adjust flexibly to individual heterogeneity while performing well under data constraints. In our setting, we have less than 10 observations per specialist-treatment for which the human signal (r) is elicited, making traditional estimation methods unreliable. To address the small sample size, we adopt a hierarchical approach (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.

When estimated on data from the main experiment, the model predicts an average additional gain of ~ 1.5 percentage points, offering a suggestive upper bound on what personalised assistance can achieve. This potential improvement is modest: performance would still fall short of the Bayesian benchmark of 82% accuracy, and the gain is smaller than that generated by the universal calibrated coarsening rule (over assistance with the full AI risk-score).

6.1 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:

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

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 $[0, 1]$.[36] If the treatment is *human-only* (i.e., $|T| = 0$), we set $\Phi_T(s_n) = 0$ for all n. Finally, under *full revelation* (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. We assume that both $\alpha_{i,T}, \beta_{i,T}$ are independent of each other and of r,[37] and normalise the coefficient on the human signal to fix scale.[38] The term $\epsilon_{i,T,n}$ captures residual idiosyncratic variation in decision-making not explained by observed

[36] This ensures that a jump from $0 \to \mu$ has the same marginal effect on latent approval as from $\mu \to 1$.

[37] 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.

[38] To ensure stable hierarchical estimation, in practice we fix the coefficient on the human signal to $1/\sigma_\epsilon$ and work on a unit-variance probit scale. Dividing the structural equation by $\sigma_{\epsilon,T}$ gives

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

signals, where $\sigma^2_{\epsilon,T}$ represents residual heterogeneity at the decision-level within treatment T.

To reflect the single-crossing property in Result 1 and 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.

6.2 Decision model estimation

As we wish to retain our flexible conceptual framework and permit different updating procedures under varying information policies,[39] 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. This occurs completely at random for 90% of the cases—that is, the probability that $\hat{r}_{i,T,n}$ is unobserved does not depend on any observed or unobserved variables—so 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 modelled 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—*the decision likelihood*—given the model parameters $\Theta = \{\alpha_{i,T},\ \beta_{i,T},\ \lambda,\ \delta,\ \sigma^2_{\epsilon,T},\ \sigma^2_{r,T}\}$:

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

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.

[39] i.e., π_Φ is indexed by Φ as per our conceptual framework.

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 marginalises over the latent signal $\tilde{r}_{i,T,n}$. 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 regularise 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 regularisation 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 $(\hat{\sigma}^2_\delta, \hat{\sigma}^2_\lambda)$ become the means and variances of our priors: $\delta \sim \mathcal{N}\big(\hat{\delta}, \hat{\sigma}^2_\delta\big)$ and $\lambda \sim \mathcal{N}\big(\hat{\lambda}, \hat{\sigma}^2_\lambda\big)$. The residual signal variance is given a truncated normal prior $\sigma_r \sim \mathcal{N}\big(\hat{\sigma}_r, \hat{\sigma}^2_{r,se}\big)\ \mathbb{1}_{(0,\infty)}$.

6.3 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 $t \in \{30\}, \{50\}, \{70\}$, recalling that the actions of the Receivers for each binary treatment are characterised by $\{t_1, r_0, r_1\}$, where t_1 is the binary threshold and $r \in [r_1, r_0)$ is the region where the specialist would adhere to the AI recommendation.

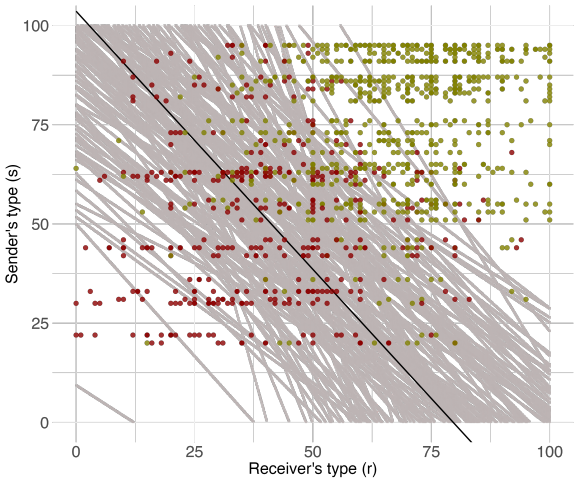
Trinary Treatments: $r_2 = -\alpha_{i,T} - \beta_{i,T}, \quad r_1 = -\alpha_{i,T} - \beta_{i,T} \cdot \mu_M(t), \quad r_0 = -\alpha_{i,T}$ for $t \in \{30, 50\}, \{50, 70\}$, where the actions of the Receivers for each trinary treatment are characterised by $\{t_1, t_2, r_0, r_1, r_2\}$, where t_1, t_2 are the thresholds and r_0, r_1, r_2 are the minimum human signals needed to "Approve" under each induced message.

6.4 Heterogeneity in decision-making: example from main experiment data

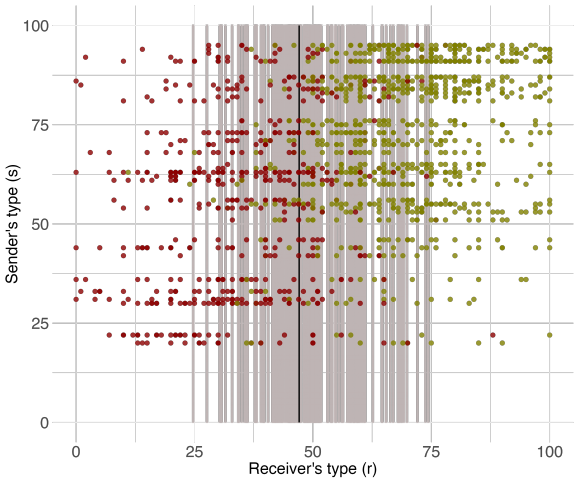
Using data from the main experiment, we estimate the model and draw $S = 1000$ posterior samples of the model parameters. For each specialist-treatment pair, we compute and plot the *median* decision boundary in the (r, s)-plane. Specifically, for the *Probability* treatment, we use the median slope and intercept; for other treatments, we take the median of each relevant point of the adherence curve (e.g., $r_0(t), r_1(t)$, etc.). Each grey line in Figure 11 represents one specialist's median boundary under a given treatment, while the solid black line indicates the median boundary across specialists in that treatment.

Visualisations of the individual decision-making boundaries reveal considerable heterogeneity among specialists. For instance, under *Probability* (Figure 11a), we observe significant variation

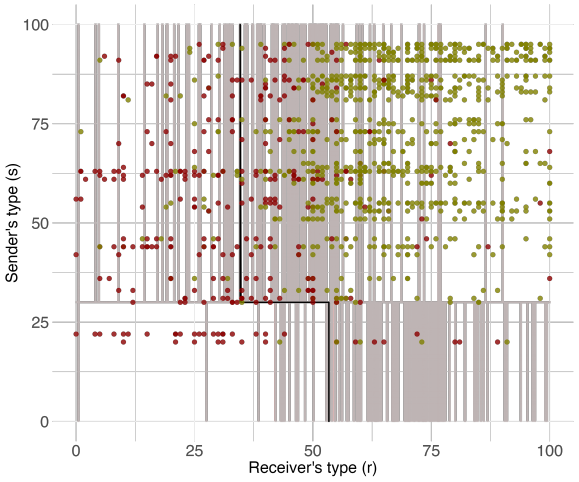
Figure 11: Heterogeneity in decision-boundaries under different treatments



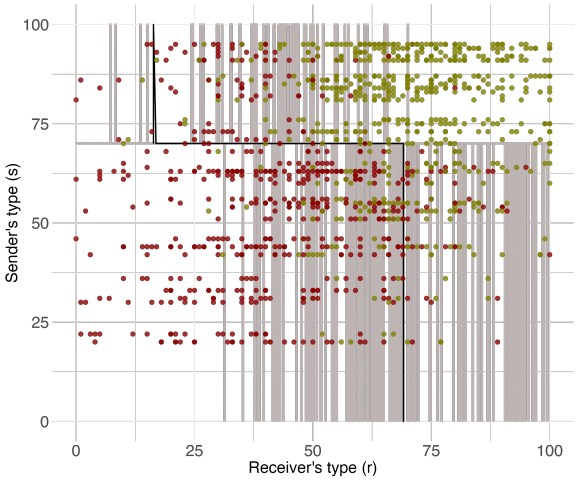
(a) Probability



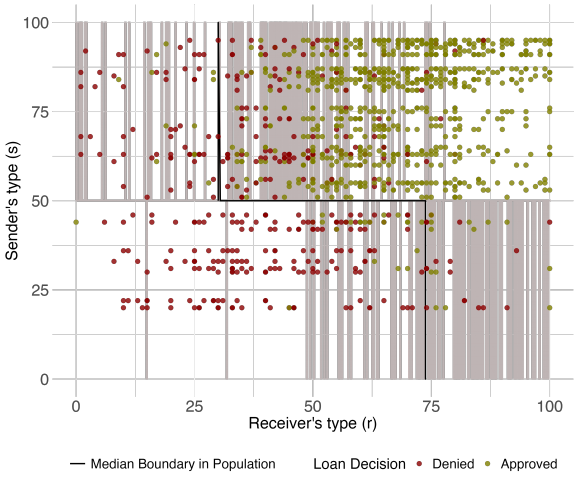
(b) Human-Only



(c) Binary Low (30)



(d) Binary High (70)



(e) Binary Medium (50)

Note: Each panel shows the estimated median decision boundary for individual specialists under different treatments, 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 across specialists. Background dots represent observed decisions, coloured by action.

in slopes and intercepts. Specialists with flatter boundaries and intercepts around $s = 50$ behave closer to the Bayesian benchmark (fully flat); in contrast, steeper boundaries reflect greater weight on the human's private signal or underweighting of the AI's signal. There is also considerable heterogeneity in how people respond to treatments—see for example the wide range in adherence under *Binary Medium* in Figure 11e—suggesting potential for improvement via personalisation.

6.5 Personalising policies using decision boundary estimates

Using the specialist-level parameters estimated as described above, we can then 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 prior in the population.

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 maximises this expected accuracy:

$$t_i^{\text{opt}} = \arg\max_t \text{Accuracy}_{i,T}$$

Assigning each loan specialist their predicted optimal coarsened treatment results in 22% allocated to *Binary High (70)*, 5% allocated to *Binary Low (30)*, and 70% retaining the *Binary Medium (50)* policy. This approach is predicted to yield a 1.5 percentage point improvement over the generalised policy—a modest gain that highlights the potential for further improving decision-making outcomes by calibrating not just thresholds, but treatments to individual decision-makers. Our second experiment tests this in practice by deploying personalised policies in a two-stage experiment.

7 Estimating the Gains from Personalised Policies

Our second experiment recruits a new set of loan specialists to make loan approval decisions, aiming to test whether personalised coarsening policies can improve decision-making outcomes further relative to universal benchmarks. In this experiment, we also extend the consideration set to trinary coarsening policies.

While the model in Section 6 predicts modest average gains from personalisation, even these modest gains do not materialise in practice. We find no average improvement, reflecting both the limited size of the predicted gains and behavioural nonstationarity: personalisation helps (marginally significantly) only when individual decision patterns remain stable. These findings underscore the practical challenges of implementing personalisation in dynamic environments and suggest that a well-calibrated universal coarsening rule may already capture much of the available benefit—offering a simple, robust, and effective way to improve outcomes in AI-assisted decisions.

7.1 Survey overview

Our experiment took place between May 17 and June 24, 2025. We recruited 218 US, UK, Canada, Australia, and New Zealand-based participants aged 18 or older over Prolific. Participants underwent rigorous screening consistent with the procedure described in Section 4, with one modification: instead of requiring participants to report their workplace branch *or* demonstrate familiarity with loan specialist responsibilities, we relax this criterion to scale recruitment. As an additional safeguard, however, we filter out participants after the first survey if they get fewer than the prior probability of loan approvals correct, with some buffer.[40] This exclusion criterion is applied before our key randomisation occurs in the second survey, ensuring it does not introduce selection bias across treatment arms or compromise the internal validity of the treatment comparisons.

Qualified participants were directed to complete the first online survey (Stage 1), hosted on the oTree platform via Heroku. This survey took approximately 60 minutes and involved making 50 loan application decisions in addition to 10 practice questions, as well as answering additional questions (e.g., biases, demographics) detailed in Appendix 9.3.2. We oversampled some types of loan applications for power. Participants are also given optional loan specialist training (in the form of a short < 2 min video) at the beginning of the practice questions.

[40] The prior probability in the population is 66%, so if a participant approved every loan, their accuracy would be 66%. To account for decision noise, we set the filtering benchmark slightly lower than this.

Those who fulfill the selection criteria are also invited back within a month to complete a second survey (Stage 2), taking 40 minutes on average. Participants received a base payment of $16 for the first survey and $10 for the second, with an average additional bonus of about $6.

Table 4 presents the demographic characteristics of the 173 participants who completed both surveys. On average, participants reported having 6 years of experience as loan specialists. Additionally, more than half indicated prior experience working with supervised AI models.

Table 4: Participant Demographics in Personalisation Experiment

	Sample Mean or Percentage
Age	38
Female	49.71%
Bachelor's Degree or Above	81.50%
Income	104205
Loan Specialist Experience	100.00%
Loan Specialist Years	6
Bank / Credit Union Experience	95.38%
Applied to Loans Before	95.38%
Loan Amount	207102
Used Supervised AI Before	50.29%
Never Used any AI	7.51%

Figure 12 summarises the design of our experiment. In Stage 1, decision blocks appear in randomised order for the following conditions: {*Binary Low (30), Binary Medium (50), Binary High (70), Trinary Low (30, 50), Trinary High (50,70)*}.

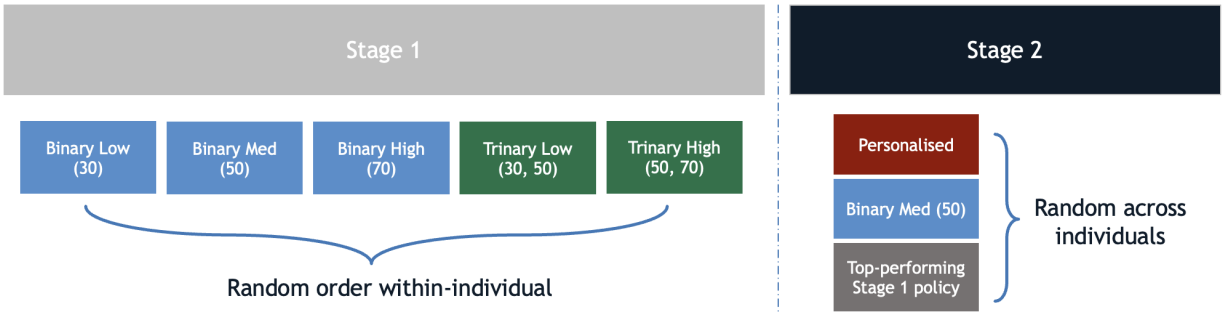
Loan application decisions are made in the same manner as in the main experiment, described in 4.2. The primary difference lies in the presentation of the AI signals, which are now shown on a dial with accompanying text (e.g., "This applicant's score is between 0–50"). This format is intended to minimise the difference in visual design across the binary and trinary conditions (see Appendix Figure 30-31), so that the only variation reflects the coarseness of the AI signal.

7.2 Treatment conditions

Stage 2 forms the basis of our main treatment randomisation. Each participant is shown 30 new loan applications, each randomly assigned to one of the following arms:

- **Personalisation (50%):** AI signal shown using the individual's personalised assignment, based on out-of-sample predictions from model in Section 6, estimated using Stage 1 data

Figure 12: Personalisation experiment survey flow



- **Binary Medium (25%):** AI signal shown using a binary treatment with a threshold of 50—the generalised universal policy used in our main experiment
- **Top-performing (25%):** AI signal shown using the best-performing condition, on average, across the Stage 1 population (in this case, *Trinary High*)

This design allows us to assess the benefits of individualised policy assignment versus two potential benchmarks of universal "off-the-shelf" policies: the model-implied best binary treatment as tested in the main experiment, or the top-performing non-personalised policy in the population.

7.3 Empirical strategy

Our main specification follows that of Equation 2, with the one difference being that the set of treatment conditions $T \in \{\textit{Binary Medium, Top Performing, Personalisation}\}$. We set the *Personalisation* treatment as the reference base level. Again, identification is established by the random assignment of the treatment. Our main objects of interest are the β_ts—in particular we test the following hypotheses, pre-registed in our pre-analysis plan (AEARCTR-0015758):

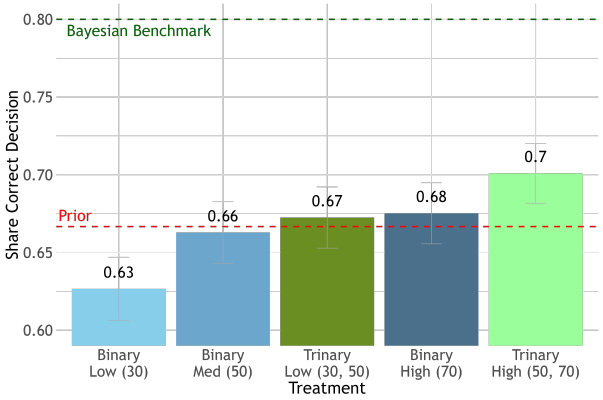
- Whether the *Personalisation* arm performs differently from *Binary Medium* treatment, i.e., test the null that $\beta_{\text{binary med}} = 0$
- Whether the *Personalisation* arm performs differently from *Top-performing* treatment, i.e., test the null that $\beta_{\text{top-perform}} = 0$

7.4 Main results

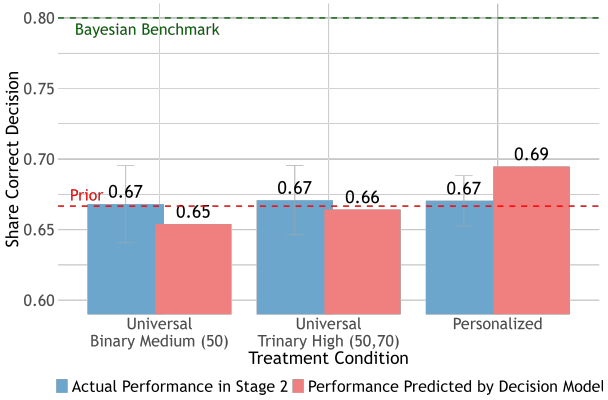
Figure 13 presents the main results from Stage 1 (left) and Stage 2 (right) of the personalisation experiment.[41] In Stage 1, the *Trinary High (50, 70)* condition slightly outperforms the other coarsened policies and is therefore used as the top-performing benchmark going into Stage 2.

Average performance across conditions is low and not significantly different from the prior, even after excluding participants who scored below 60% (as pre-registered). The corresponding regressions can be found in Appendix 31. These raw accuracy levels reflect the fact that we (i) recruited only Prolific participants in this experiment, and (ii) intentionally oversampled harder cases. When we reweight the data to reflect the original case distribution, performance levels are higher across the board (see Appendix Figures 33a and 33b). 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. That said, *Trinary High* remains (marginally) the best-performing policy and is retained as the "top-performing" universal benchmark going into Stage 2.

Figure 13: Stage 1 and Stage 2 results, personalisation 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.** Blue bars represent the share of correct decisions. Red bars indicate the performance *predicted* by the decision model in Section 6, estimated on Stage 1 data.

In Stage 2, we observe no evidence of treatment effects: whether assigned to a personalised policy or not, participants' average performance stays flat at the prior. The red bars represent performance predicted by the decision model introduced in Section 6; while the model suggests gains from personalisation, these gains are not realised in practice.[42]

[41] We present further breakdown of Stage 2 outcomes by specific personalised assignments in Appendix 9.7.1.

[42] Model predictions (red) include only participants who complete Stage 2, so differential attrition is not at play here.

7.5 Model validation

7.5.1 Is the model overpredicting gains to personalisation?

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

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

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 of the analysis using posterior draws from the hierarchical model. For each posterior draw s, we compute:

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

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

7.5.2 Testing Stage 1 model fit

Recall that the model is fit to Stage 1 data from the same participants who are called back in Stage 2; moreover, it predicts accuracy for treatments it has already seen in Stage 1. Thus, the overprediction of gains must stem either from (i) poor model fit to Stage 1 data, or (ii) changes in participant behaviour between Stage 1 and Stage 2 due to forces not captured in the model. We begin by testing for Stage 1 fit.[43]

We find strong evidence of model fit in Stage 1. Out-of-sample, the model accurately predicts held-out approve/deny decisions with 77–84% of the time. In-sample, specialists perform significantly better in their model-predicted optimal treatment over other treatment conditions (Appendix 9.7.7). Given this, we next examine whether shifts in participant behaviour between stages may

Attrition analysis is available in Appendix 9.7.3.

[43] 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 personalisation benefits (Appendix 9.7.4). Second, the model passes a placebo test (Appendix 9.7.5) randomly reshuffling treatment assignments—under which no gains are predicted—suggesting that the model's predictions are not simply an artefact of noise. Third, re-running the model using observed $\hat{r}$ values in Stage 2 (rather than treating r as latent) yields similar results, and excluding participants whose treatments differ under the plug-in $\hat{r}$ approach yields no improvement (Appendix 9.7.6), indicating that model misprediction of r is not the primary issue.

explain why personalisation delivers smaller gains than expected.

7.6 Why is personalisation delivering smaller gains than predicted?

7.6.1 Changes in participant behaviour

One explanation is that participant behaviour shifted between Stage 1 and Stage 2 due to factors not captured by the model—such as fatigue, forgetting, or context shifts. To test for this, we re-estimate the decision model in Section 6 using only Stage 2 data. Since each participant only receives one treatment in Stage 2, we estimate a posterior distribution over a single set of decision parameters per participant-treatment pair.

Instead, we observe substantial deviations across participants (Appendix Figure 35), indicating behavioural shifts not captured by the original model. For some individuals, these shifts are large: a 20 percentage point difference implies, for example, that a participant who approved loans above a 50% AI score in Stage 1 may only approve above 70% in Stage 2. These discrepancies need not be statistically significant on average to be consequential; what matters is the wide dispersion, with many participants far from zero.

Behavioural instability may be partly attributable to the participant population. In Experiment 2, all participants were sourced through Prolific. While they reported prior experience with loan decisions, they had fewer years of relevant experience than the field-recruited professionals in our main experiment, and were likely less familiar with the task overall. This instability complicates the transferability of personalised recommendations, and motivates our use of adherence shifts as a moderating variable.

7.6.2 Using differences in S1 and S2 adherence curve estimates as baseline moderators

We explore whether the effectiveness of personalisation depends on the stability of participants' decision strategies between Stage 1 and Stage 2. 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). When we split the sample at the median, we find that participants with below-median adherence shifts exhibit positive gains on the order of 2 percentage points (though we are not powered for statistical significance) from personalisation, whereas those with larger shifts see no benefit (Appendix Table 37).

We formally test this moderation in an interaction model and find that when adherence curve

differences are near zero—i.e., when participant decision strategies remain stable—the effect of personalisation is positive and statistically significant at the 10% level (Appendix Table 38), suggesting that stable individual behaviour is necessary for realising the gains that the model predicts.

Taken together, these findings highlight why the model-predicted gains from personalisation did not materialise in practice. First, the predicted gains were modest to begin with—suggesting that at least in this setting, personalisation offers only incremental improvements over a universal coarsening. Second, behavioural instability between estimation and deployment further limits effectiveness: when individuals shift how they use information over time, past data becomes a poor guide for tailoring future interventions.

These results underscore the practical challenges of implementing personalised information policies, particularly in environments where inconsistency and behavioural shifts are common, or where decision-makers are unfamiliar with the task. In such settings, a well-calibrated universal coarsening rule captures much of the available benefit and provides a simple and effective alternative. While personalisation may prove more valuable in institutional contexts with long-term behavioural data, our findings suggest that universal calibration is a strong and practical default.

8 Conclusion and Discussion

We show in this paper—both theoretically and empirically—that coarsening AI signals at calibrated thresholds can significantly improve decision-making outcomes, including in high-stakes contexts like loan approvals. Our study offers one of the first empirical demonstrations of how information design can address human information aggregation errors, a key barrier to the effective use of AI predictions in practice. Moreover, our results are demonstrated with professional loan specialists recruited through industry channels, 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 generalise. To address this, we are conducting further experiments in which humans possess private information (e.g., interviews unavailable to the AI). In such settings, personalised 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 personalised rules may

be fragile in the face of context shifts or behavioural drift: changes in decision-maker behaviour or environment can limit the gains to personalisation, as our second experiment suggests.

Calibrated coarsening remains generalisable 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 organisational change. Additionally, we are exploring empirical extensions to other domains, including hiring decisions, to assess the broader applicability of calibrated coarsening in different decision-making contexts.

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 misalignments, richer extensions—such as dynamic policies without commitment—introduce new challenges, including strategic behaviour 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 TP/TN payoffs, which can be accommodated in our current framework by assigning case-specific thresholds, and aggregating across the varying values of $v(r, t)$ implied by heterogeneous TP/TN structures. 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.[44]

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 behaviour, 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

[44]Our framework already extends to any continuous or discrete state space, provided the action space remains binary.

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. (2023). Algorithmic recommendations and human discretion. Technical report, National Bureau of Economic Research.

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. (2024). Optimal disclosure on crowdfunding platforms. Working Paper, SSRN 5003864.

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. (2021). Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454*, 5.

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. (2024). Does AI help humans make better decisions? a statistical evaluation framework for experimental and observational studies. *arXiv preprint arXiv:2403.12108*.

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. (2024). The rapid adoption of generative ai. Technical report, National Bureau of Economic Research.

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. (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.

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. (2019). How do humans interact with algorithms? experimental evidence from health insurance. Technical report, National Bureau of Economic Research.

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. (2024). The abc's of who benefits from working with AI: Ability, beliefs, and calibration. Technical report, National Bureau of Economic Research.

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.

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. (2024). Using ai as gatekeeper or second opinion: Designing patient pathways for ai-augmented healthcare. *Available at SSRN 5055325*.

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., 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.

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. (2011). *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*.

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.

Huelden, T., Jascisens, V., Roemheld, L., and Werner, T. (2024). Human-machine interactions in pricing: Evidence from two large-scale field experiments. Working paper, Zalando.

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.

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. (2022). Learning when to advise human decision makers. *arXiv preprint arXiv:2209.13578*.

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. (2024). Failing to correctly aggregate signals. *Available at SSRN 4795431*.

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. (2024). Algorithm reliance, fast and slow. *Fast and Slow (May 31, 2024)*.

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., Wiely, 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. (2019). Algorithmic risk assessment in the hands of humans.

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

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

Xiang, J. (2024). Physicians as persuaders: Evidence from hospitals in china. *Working Paper (Job Market Paper)*.

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*.

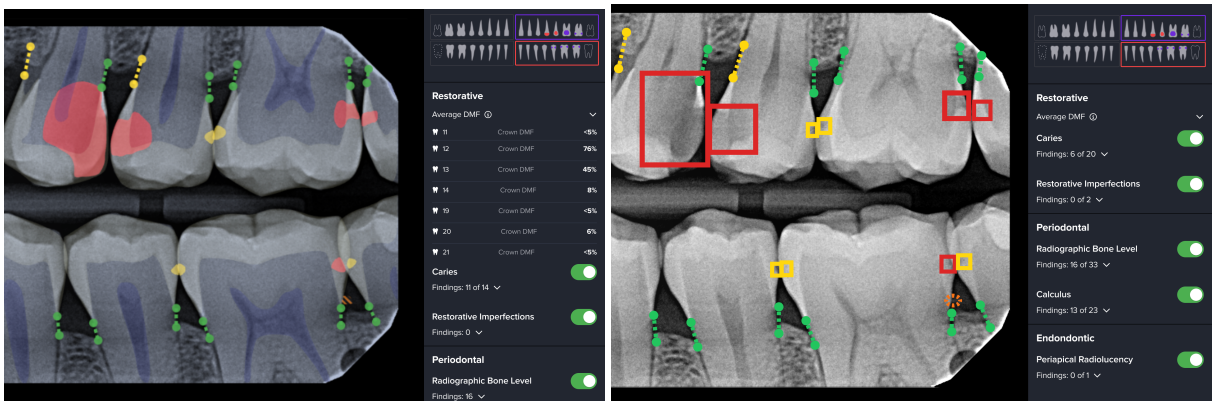
Zhang, W. and Misra, S. (2024). Coarse personalization. In *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 1206–1208.

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.

9 Appendix

9.1 Real-world Examples of Coarsened AI signals

Figure 14: VideaAI Dental Interface (left: probability, right: binary)



9.2 Model

9.2.1 Optimal coarsening under Bayesian Receiver

Proposition 1. *Under Assumption 1, let $\tilde{\pi} = \pi$. Then the Sender's optimal disclosure mechanism is full revelation, Φ_s.*

This result maps to Proposition 6(i) in Kolotilin (2018) and KG, who show that full disclosure is optimal when the Sender's payoff is convex in Receiver's beliefs (as with fully aligned preferences and beliefs).

Proof. Fix any information policy Φ and any decision rule $a = a(r, \Phi(s))$ that is implementable under Φ. Under full revelation Φ_s, the Receiver observes s (and r) and can compute $\Phi(s)$, hence can implement the same rule $a(r, \Phi(s))$.

Therefore,

$$\begin{aligned}
\mathbb{E}[a(r, \Phi(s))\, U(\omega)] &= \mathbb{E}[\mathbb{E}[a(r, \Phi(s))\, U(\omega) \mid r, s]] \\
&= \mathbb{E}[a(r, \Phi(s))\, \mathbb{E}[U(\omega) \mid r, s]] \\
&\leq \mathbb{E}[\max\{\mathbb{E}[U(\omega) \mid r, s]\,,\, 0\}]\,,
\end{aligned}$$

where the inequality holds pointwise because $a(r, \Phi(s)) \in \{0, 1\}$. The upper bound is attained under full revelation by the Receiver's optimal choice

$$a^*(r, s) = \mathbf{1}\Big\{\mathbb{E}[U(\omega) \mid r, s] \geq 0\Big\}.$$

Hence full revelation weakly dominates any policy Φ. □

9.2.2 Remark on relation of AI misweighting and Grether's framework

Remark (Relation to Grether 1992)**.** *When $\rho = 0$, the AI misweighting formulation is isomorphic to the Grether (1992) framework, under which the DM forms a posterior proportional to $g_\omega(r) \cdot f_\omega(s \mid r)^\alpha$, underweighting ($\alpha < 1$) or overweighting ($\alpha > 1$) the evidential contribution of the AI signal. With correlated signals, that approach faces two difficulties. First, it requires arbitrarily designating one signal as the base: writing $f_\omega(s)^\alpha \cdot g_\omega(r \mid s)$ would yield a different posterior. Second, it generates an internal inconsistency whereby the DM is biased about what s reveals about ω, but can correctly account for what s reveals about r (which in turn determines what r reveals about ω), despite both inferences requiring the same understanding of the information structure. Our parametrization avoids both problems: α propagates coherently through the joint likelihood regardless of how we write it; moreover, the distortions to the direct weight on s and the cross-signal adjustment move together in a disciplined way.*

9.2.3 Binary coarsening with known *r*

To establish this next result, we first impose that the Receiver's posterior is a well-behaved function of the interval consistent with the message:

Assumption 2 (Interval-based posterior)**.** *There exists a function $\hat{\pi}(r; [\underline{s}, \bar{s}])$ such that for any monotonic disclosure mechanism Φ and any message k with pre-image $[\underline{s}(k), \bar{s}(k)]$,*

$$\tilde{\pi}_\Phi(r, k) = \hat{\pi}\big(r; [\underline{s}(k), \bar{s}(k)]\big).$$

Moreover, $\hat{\pi}(r; [\underline{s}, \bar{s}]) \in [\tilde{\pi}(r, \underline{s}), \tilde{\pi}(r, \bar{s})]$, and $\hat{\pi}(r; [\underline{s}, \bar{s}])$ is continuous and strictly increasing in both $\underline{s}$ and $\bar{s}$.

Assumption 2 requires that the posterior induced by a message depends only on the interval of signals consistent with it, ruling out, e.g. framing effects. The posterior must also lie between

the posteriors generated by the highest and lowest signals in that interval, ruling out beliefs more extreme than any constituent posterior (while allowing arbitrary aggregation within those bounds). Finally, continuity and monotonicity in both endpoints ensure that as lower signals are removed or higher signals are added, the Receiver's posterior continuously increases.

Under this additional assumption, binary coarsening is the optimal disclosure mechanism whenever r is known:

Proposition 2. *Let $r \in R$. Assume the Receiver's beliefs deviate from Bayes, i.e., $\tilde{\pi} \neq \pi$ subject to Assumptions 1 and 2. If Φ is chosen when r is known then:*

1. *The Sender's optimal disclosure mechanism is a binary coarsening Φ_{t^*}.*
2. *Φ_{t^*} strictly increases the Sender's expected utility relative to full revelation.*
3. *If $s^*(r) > \tilde{s}(r)$ and $r < r(k; \tilde{\pi}_{\Phi_\emptyset})$, or if $s^*(r) < \tilde{s}(r)$ and $r \geq r(k; \tilde{\pi}_{\Phi_\emptyset})$, the Sender can fully implement $s^*(r)$.*

Knowing r (or equivalently, when knowing r is immaterial), the Sender can always construct a binary coarsening that strictly improves on full revelation. Moreover, (i) this is the optimal mechanism, and (ii) depending on the direction of the deviation and the Receiver's default action, full implementation of the Bayesian benchmark may be possible. The result has a simple intuition: binary coarsening converts the continuous signal into a recommendation. Knowing r, the Sender can ensure that the recommendation is incentive compatible and will be followed. (In the irrelevant-r interpretation, since r affects neither the optimal action nor adherence, the Sender can construct the same binary coarsening without conditioning on r).

Proof. Fix $r \in \mathcal{R}$. Without loss of generality, assume $s^*(r) > \tilde{s}(r)$; the opposite case is symmetric throughout. Let $\bar{\pi} \equiv c_{FP}/(c_{FP} + c_{TP})$ denote the action threshold.

For any monotonic disclosure mechanism Φ, Assumption 1 ensures $\tilde{\pi}_\Phi(r, \Phi(s))$ is weakly increasing in s, so the Receiver's behavior takes a threshold form. Define the *effective threshold*

$$s_{\text{eff}}(\Phi) \equiv \begin{cases} \sup S & \text{if } \tilde{\pi}_\Phi(r, \Phi(s)) < \bar{\pi} \text{ for all } s \in S, \\ \inf S & \text{if } \tilde{\pi}_\Phi(r, \Phi(s)) \geq \bar{\pi} \text{ for all } s \in S, \\ \inf\{s \in S : \tilde{\pi}_\Phi(r, \Phi(s)) \geq \bar{\pi}\} & \text{otherwise,} \end{cases}$$

so the Receiver acts if and only if $s \geq s_{\text{eff}}(\Phi)$, with $s_{\text{eff}} = \sup S$ meaning the Receiver never acts and $s_{\text{eff}} = \inf S$ meaning the Receiver always acts. Full revelation has $s_{\text{eff}} = \tilde{s}(r)$. For a binary coarsening Φ_t to which the Receiver adheres, $s_{\text{eff}}(\Phi_t) = t$. Define

$$t^* \equiv \min\Big\{ \sup\big\{t : \text{the Receiver adheres to } \Phi_t\big\},\ s^*(r)\Big\}.$$

Step 1 (The Receiver adheres at t^*; $t^* > \tilde{s}(r)$.). At $t = \tilde{s}(r)$, every s in the "No" pre-image $[\inf S,\ t)$ satisfies $\tilde{\pi}(r,s) < \tilde{\pi}(r,t) = \bar{\pi}$ by strict monotonicity of $\tilde{\pi}$ in s (Assumption 1). Betweenness gives $\tilde{\pi}_t(r,0) \leq \sup_{s<t} \tilde{\pi}(r,s) < \bar{\pi}$, where the strict inequality holds because the supremum is not attained in the pre-image. By continuity of $\tilde{\pi}_t(r,0)$ in t (Assumption 2.3), the strict inequality $\tilde{\pi}_t(r,0) < \bar{\pi}$ persists for t in an open neighborhood above $\tilde{s}(r)$. For "Yes" at any such $t > \tilde{s}(r)$: every $s \geq t > \tilde{s}(r)$ has $\tilde{\pi}(r,s) > \bar{\pi}$, so betweenness gives $\tilde{\pi}_t(r,1) \geq \tilde{\pi}(r,t) > \bar{\pi}$. Hence the Receiver adheres at some $t > \tilde{s}(r)$, establishing $t^* > \tilde{s}(r)$.

Step 2 (Φ_{t^*} strictly improves on full revelation). Under full revelation, the Receiver acts iff $s\tilde{s}(r)$; under Φ_{t^*} with adherence, iff $s \geq t^*$. Since $\tilde{s}(r) < t^* \leq s^*(r)$, actions differ exactly on $s \in [\tilde{s}(r),\ t^*)$, a set of positive measure. For each such s, $s < s^*(r)$ implies that the Bayesian's expected utility from acting is negative. Full revelation incorrectly approves; Φ_{t^*} correctly denies. Therefore, the Sender's expected utility is strictly higher under Φ_{t^*} compared to full revelation.

Step 3 (Φ_{t^*} is optimal among all monotonic mechanisms). It suffices to show $s_{\text{eff}}(\Phi) \leq t^*$ for every monotonic Φ, since any mechanism with $s_{\text{eff}} \leq t^* \leq s^*(r)$ is weakly dominated by Φ_{t^*}.

Suppose for contradiction that $s_{\text{eff}}(\Phi) > t^*$ for some monotonic Φ. Since $s_{\text{eff}}(\Phi) > t^* > \tilde{s}(r) > \inf S$, the Receiver does not always act under Φ, meaning there exists a message k inducing denial ($\tilde{\pi}_\Phi(r,m) < \bar{\pi}$) whose pre-image has an upper endpoint $\bar{s}(k) \geq s_{\text{eff}}(\Phi) > t^*$. Consider the binary coarsening $\Phi_{\bar{s}(k)}$. Its "No" pre-image $[\inf S,\ \bar{s}(k))$ has the same upper endpoint as m's pre-image and a weakly lower lower endpoint ($\inf S \leq \underline{s}(k)$). By interval sufficiency and strict monotonicity in $\underline{s}$ (Assumption 2.2–3):

$$\tilde{\pi}_{\bar{s}(k)}(r,0) \ \leq\ \tilde{\pi}_\Phi(r,k) \ <\ \bar{\pi}.$$

"Yes" adherence follows since $\bar{s}(k) > \tilde{s}(r)$ (as in Step 1). Hence the Receiver adheres to a binary coarsening at threshold $\bar{s}(k) > t^*$, contradicting the definition of t^*.

Step 4 (Full implementation when $r < r(k;\tilde{\pi}_{\Phi_\emptyset})$). We show the Receiver adheres to $\Phi_{s^*(r)}$. If

the signal is "Yes" the pre-image is $\{s \geq s^*(r)\}$. Since $s^*(r) > \tilde{s}(r)$, every s in this set has $\tilde{\pi}(r,s) > \bar{\pi}$. Betweenness gives $\tilde{\pi}_{\Phi_{s^*(r)}}(r,1) \geq \tilde{\pi}(r, s^*(r)) > \bar{\pi}$. If the signal is "No," The pre-image is $[\inf S,\ s^*(r))$. Full censorship pools all of S, which has the same lower endpoint and weakly higher upper endpoint. By interval sufficiency and monotonicity in $\bar{s}$ (Assumption 2), $\tilde{\pi}_{s^*(r)}(r,0) \leq \tilde{\pi}_{\Phi_\emptyset}(r) < \bar{\pi}$, where the last inequality holds because $r < r(k; \tilde{\pi}_{\Phi_\emptyset})$. Hence $\Phi_{s^*(r)}$ fully implements $s^*(r)$: the Receiver acts if and only if $s \geq s^*(r)$. □

To gain further intuition, we apply this result in the case of AI misweighting.

Example (AI misweighting (cont.))**.** *In what follows, we assume $r < r(k; \tilde{\pi}_{\Phi_\emptyset})$, i.e., the "default action" the Receiver would take if they did not have any information on s is $a = 0$ (the opposite case is symmetric). Fix r, and denote the Sender's inverse boundary by $s^* = s^*(r)$. Conditional on r, the Sender would choose $a = 1$ iff $s \geq s^*$. Notice a Receiver who over- (under-) weights the AI signal has a lower (higher) inverse boundary $\tilde{s}(r)$. Applying Proposition 2, we obtain:*

Result 2. *Let $\alpha > 0$ be the misweighting parameter characterizing the Receiver's behavior. Then:*

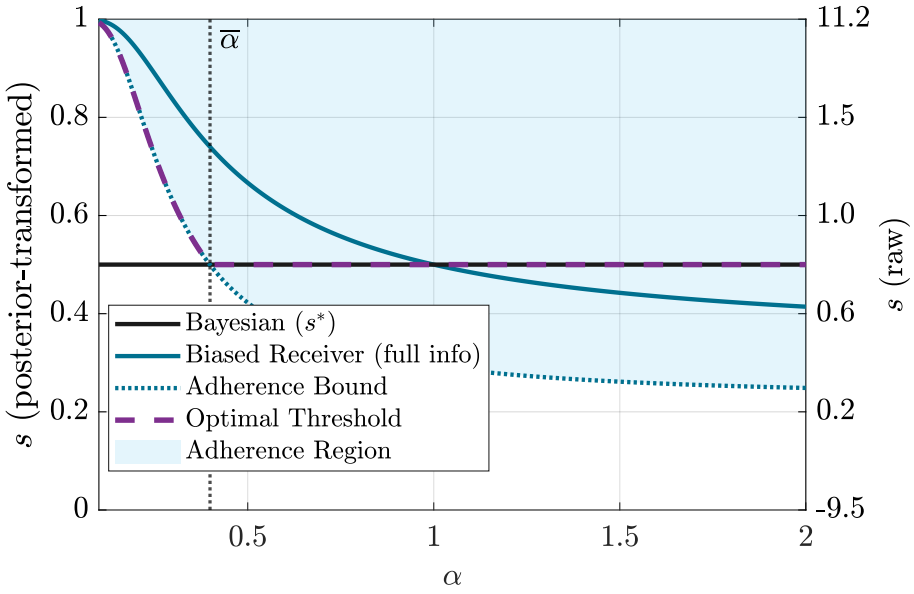
1. *The optimal disclosure mechanism is a binary coarsening.*
2. *For any $\alpha > 0$, the optimal binary coarsening strictly improves the Sender's utility over full revelation.*
3. *There exists $\overline{\alpha} < 1$ such that if $\alpha \geq \overline{\alpha}$, the optimal binary coarsening fully implements $s^*(r)$.*

In other words, when the Receiver's default is inaction, under overweighting *($\alpha > 1$), binary coarsening can always fully implement the full-revelation Bayesian benchmark. When $\alpha < 1$ (underweighting), binary coarsening can always achieve higher utility than full revelation, but full implementation is possible only when α is sufficiently high.*

For intuition, Figure 15 plots the AI signal s against the AI misweighting parameter α and illustrates how the Receiver's adherence to the binary message varies with their bias and the threshold at which the message is coarsened.

The solid teal line shows the Receiver's full-information boundary: above it, they act. This intersects the Sender-optimal (black solid) boundary at $\alpha = 1$, when the Receiver is unbiased and the Sender's inverse boundary is s^* (the optimal binary coarsening for an unbiased agent). A binary coarsening at threshold t divides the AI signal space into two messages: 1 ("Yes") and 0 ("No"). The blue dotted line marks the lowest threshold t that induces action $a = 1$ with a "Yes"

Figure 15: AI misweighting when r is known: graphical illustration



Note: Boundaries, adherence, and optimal threshold with known r, under different values of AI misweighting. The parameters used in this illustration: $p = \frac{1}{3}$, $s \sim N(\omega, 1)$ (r is a constant and is therefore omitted). The y-axis is labeled both in the units of the raw AI signal (right) and in the units of the AI signal transformed into (Bayesian) posterior units (left).

message—since the default is inaction, the message must be sufficiently high to induce action. The light blue shaded region shows all (α, t) pairs where the Receiver adheres to the Sender's coarsened message.

Since non-biased agents ($\alpha = 1$) already adhere to the binary message coarsened at s^*, agents with $\alpha > 1$—who place greater weight on the AI signal—are even more *likely to follow it. This corresponds to the last case in Proposition 2: the default action is 0, and the Sender's inverse boundary is higher, enabling full implementation.*

By contrast, $\alpha < 1$ agents underweight the AI signal and require a stronger message to act, so their red line lies above the Sender's. Sufficiently underweighting agents (i.e., $\alpha < \overline{\alpha}$) require the Sender to raise the threshold higher than s^*, though doing so still improves Sender utility relative to full revelation. The vertical black dotted line at $\overline{\alpha}$ denotes the lowest α for which the Receiver strictly adheres to the Sender's optimal boundary, allowing the Sender to fully implement their preferences. The optimal threshold (purple dashed line) is thus given by the upper envelope of the dotted blue line and the solid orange Sender-optimal boundary.

9.3 Main Experiment Design

9.3.1 Main Experiment Instructions

Figure 16: Instructions to participants to highlight prior and AI details

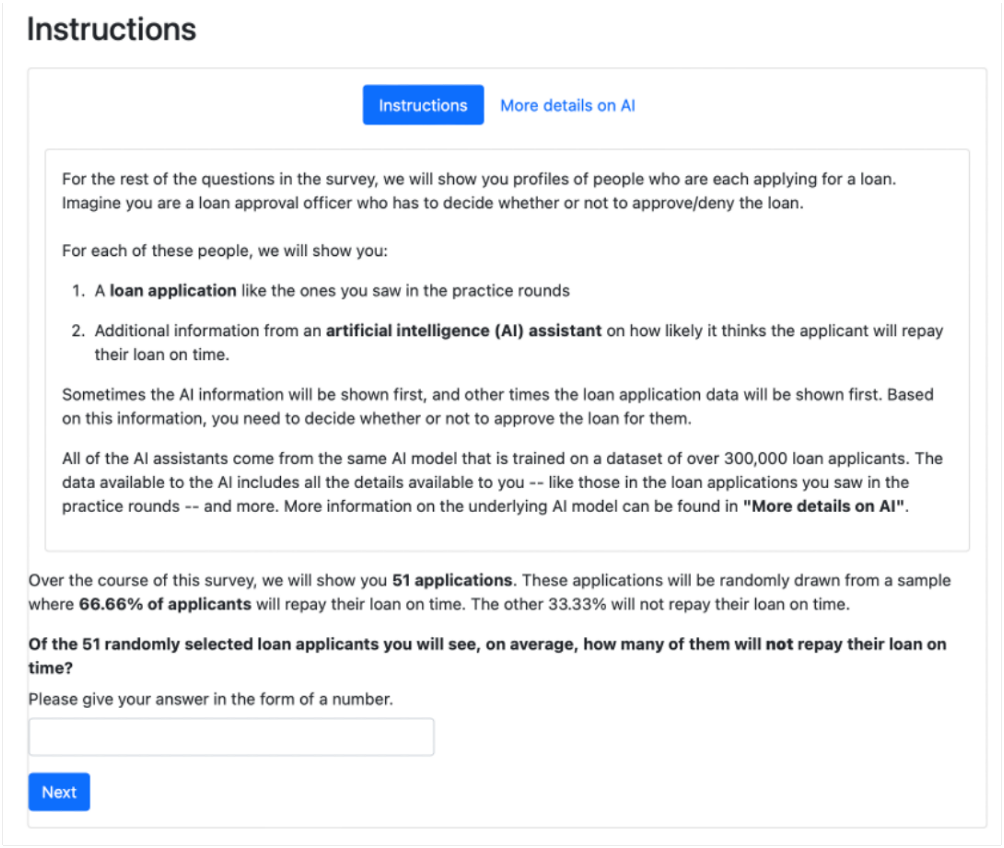


Figure 17: Detailed instructions give “More details on AI”



Figure 18: Instructions for each treatment condition highlight AI performance (*Binary Medium (50)*) condition shown here as example

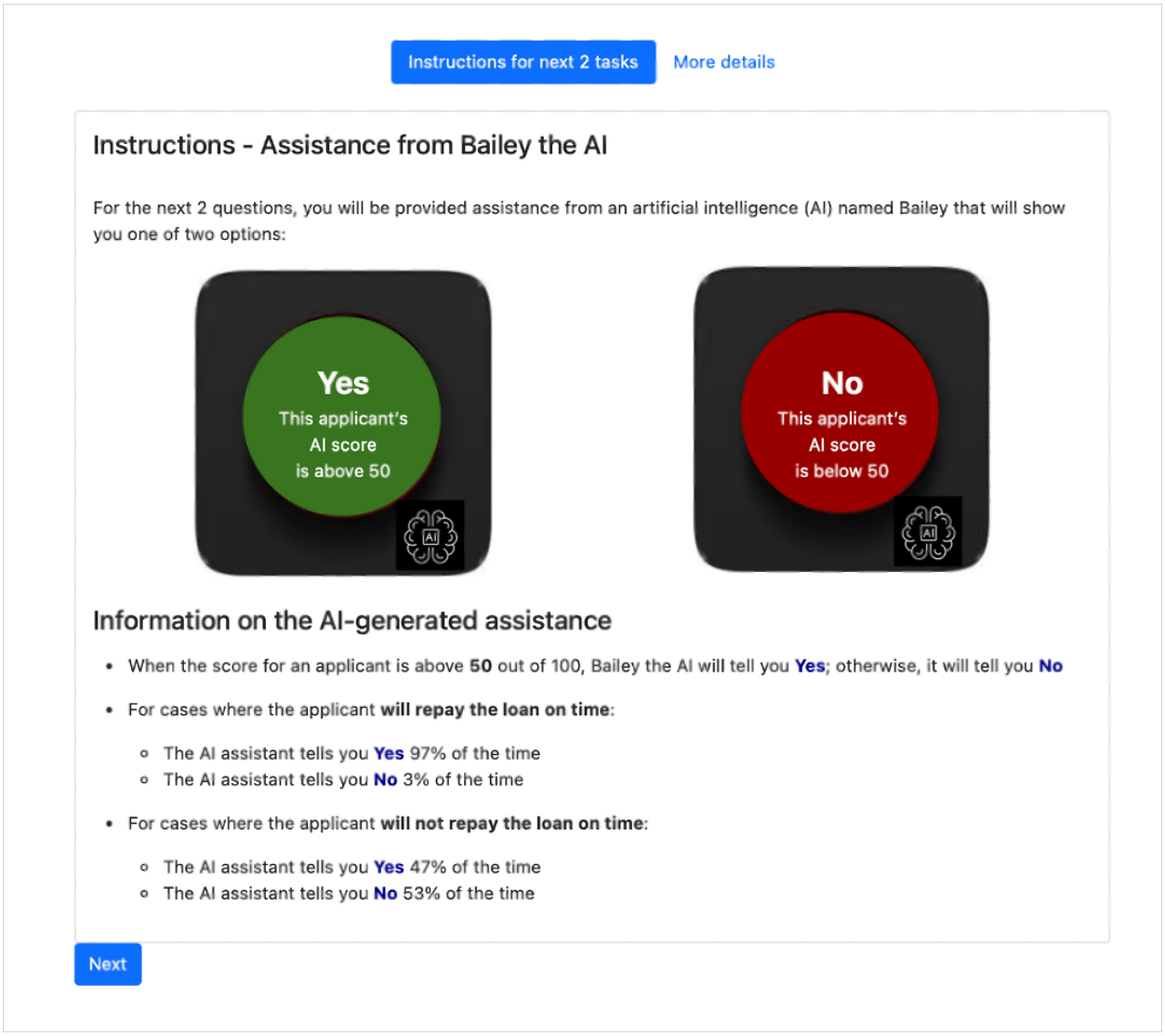


Figure 19: Detailed instructions for each treatment condition (*Binary Medium (50)*) condition shown here as example

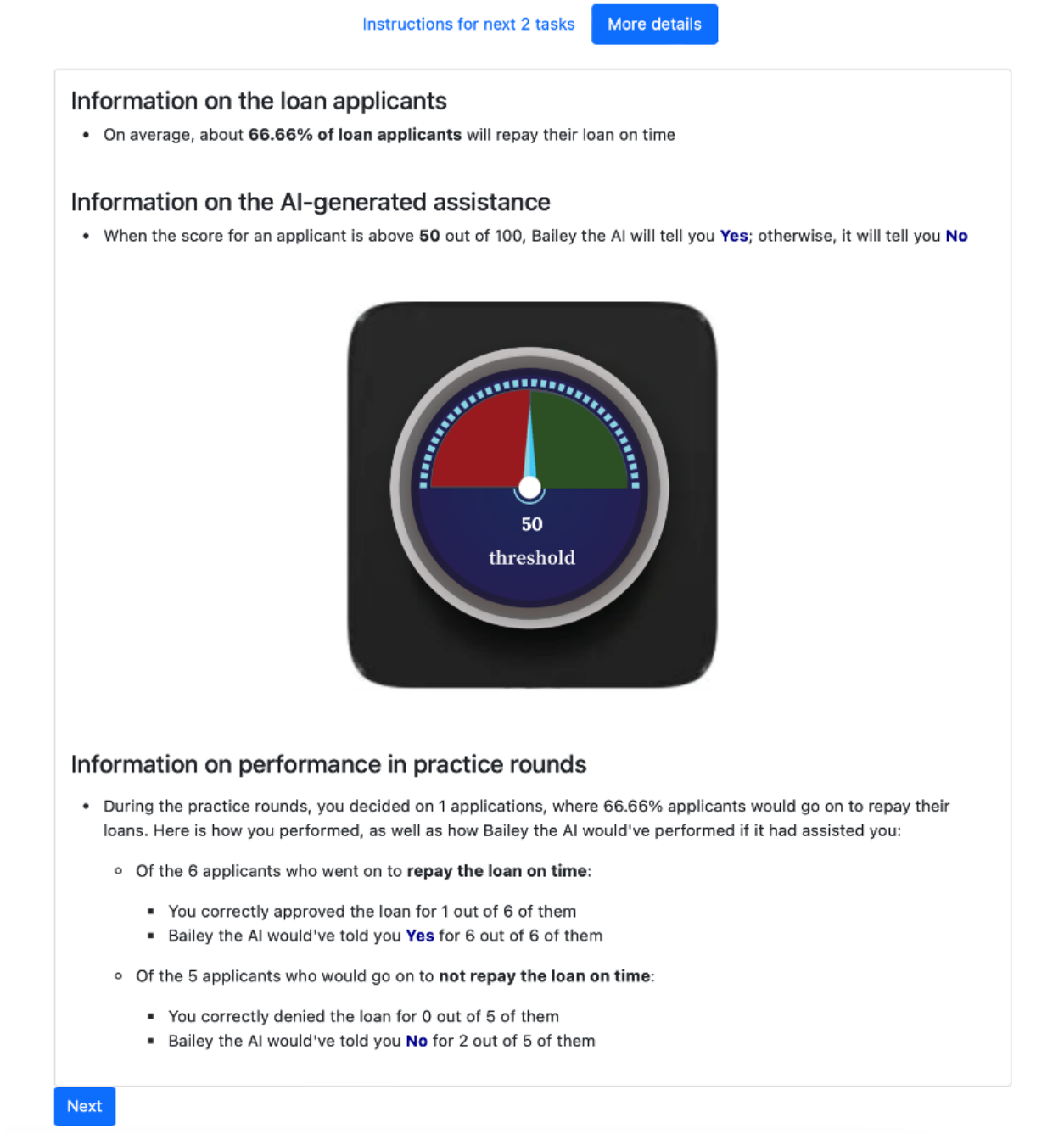
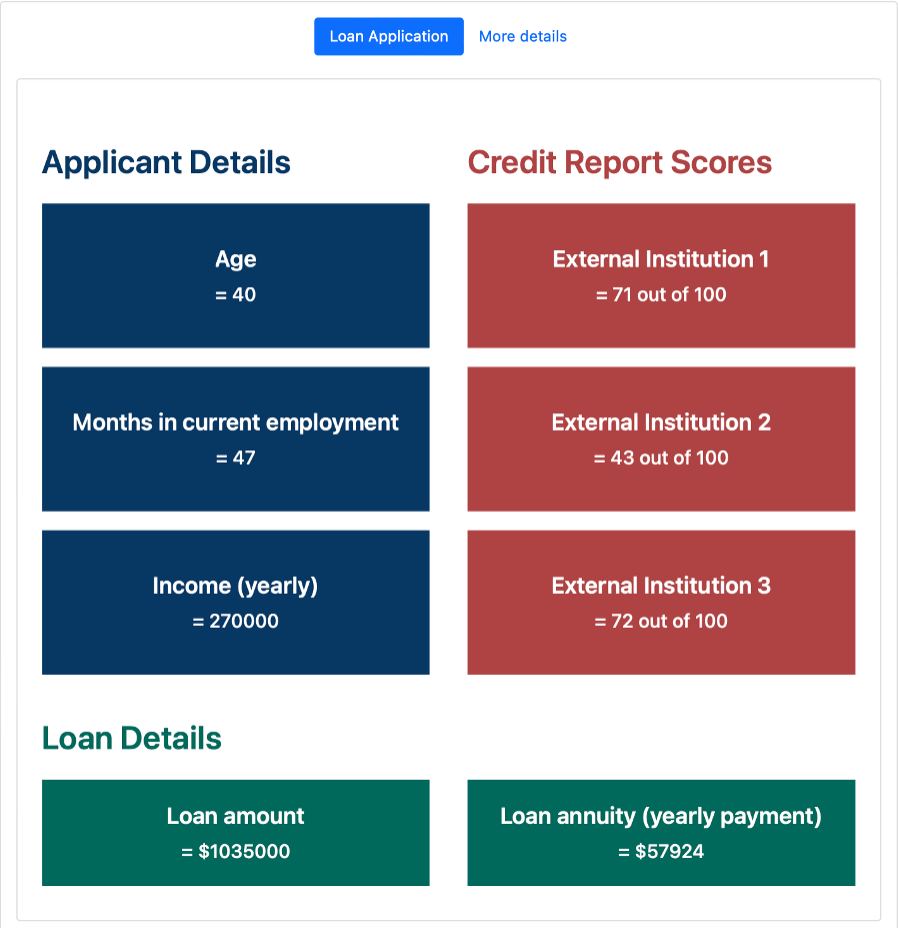


Figure 20: Example of first signal (in this case, loan application)



- **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:** Normalised credit score from external data source #1-3 (e.g., FICO, VantageScore), ranging from 0-100, with 100 being the most credit-worthy.

Figure 21: Example of first “posterior” elicitation



Figure 22: Example of second signal (in this case, AI with *Binary Medium (50)* condition shown here as example)

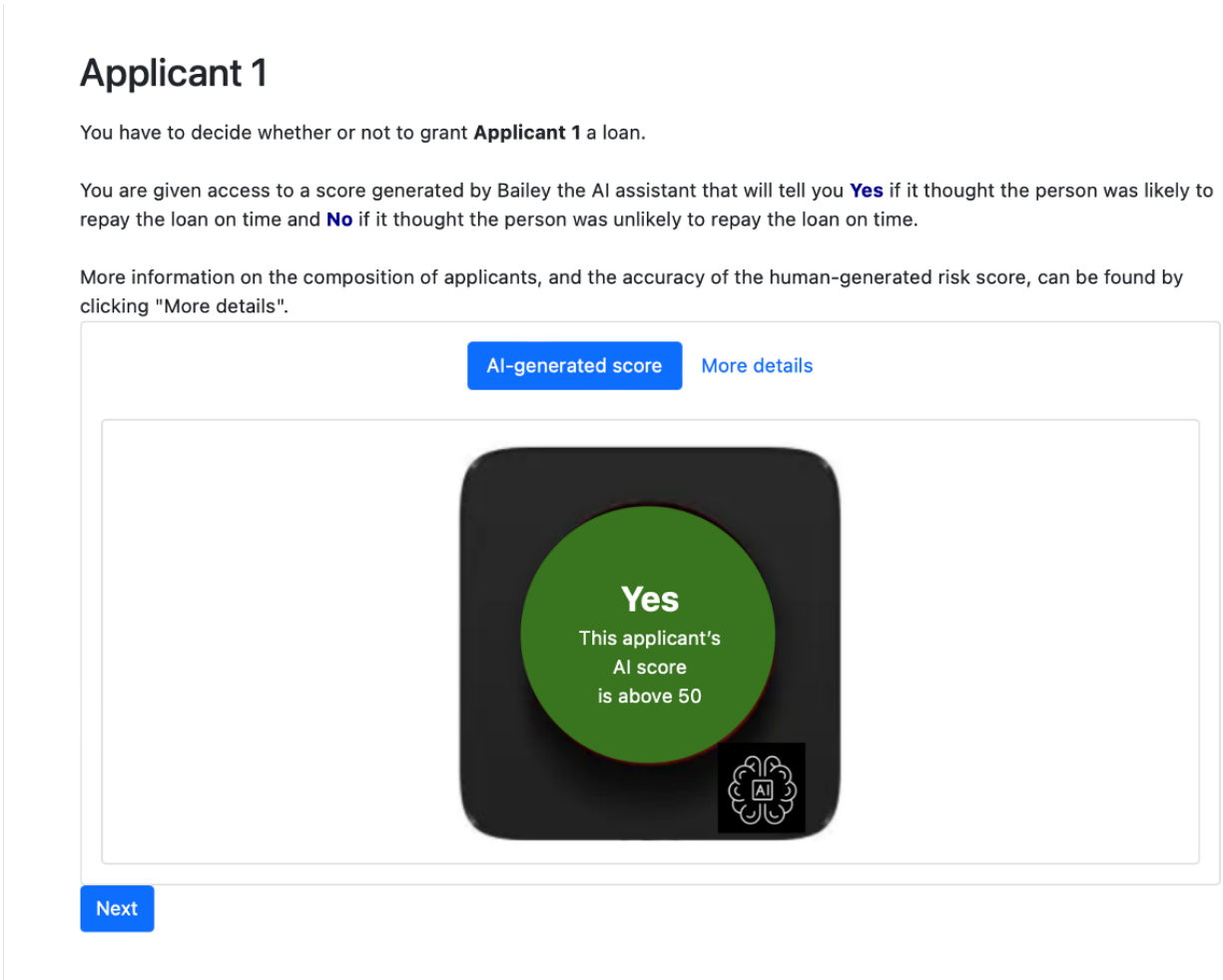
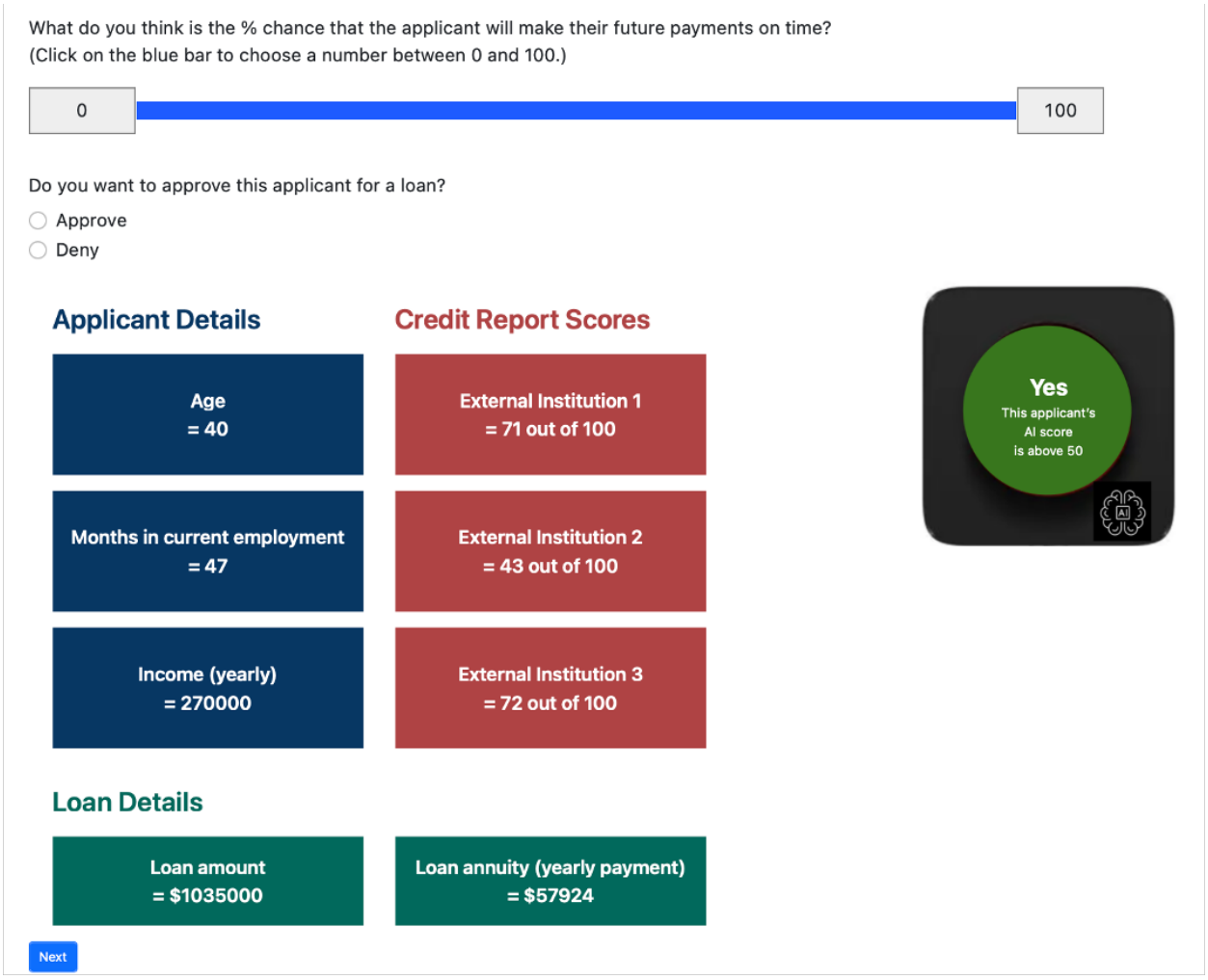


Figure 23: Approve/deny screen displays both signals again for easy decision-making



9.3.2 Additional survey questions

Demand for AI – At the end of the rounds, we ask participants to choose between having AI (either in probability or binary form) assist them, or not having AI assistance in two remaining decisions. For the binary form, they also get to choose their preferred threshold. We then additionally ask for their willingness to pay (in cents) for different types of AI assistance, in the form of a multiple price list ranging up to 10 cents, the reward for a correct answer, and randomly implement one row of the multiple price list for the final two decisions.

Perceived private information – At the end of the probability treatment, participants estimate how many of 10 decisions they believe they will get right, both with and without access to the AI score alongside the loan application information, as well as how many they believe they will get right if they only had access to the AI. The difference between these estimates can give us some insight into the participant's perceived private 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.

Cost ratio – In the experiment, the cost of a true/false positive/negative is given by the incentive structure. To understand the real-world cost ratios that loan specialists might face, at the end of the experiment we ask participants to imagine they could revise a decision: either denying a loan that would have been repaid on time, or approving one that would not have. Participants are asked how much they would hypothetically pay to change their decision.

Biases – Participants answer a series of cognitive questions documented below to establish a baseline measure of susceptibility to behavioral biases. These questions are adapted from established

literature and cover questions related to base-rate neglect, correlation neglect, AI underweighting, balls-and-urns belief updating, and more general signal aggregation (Enke et al., 2023; Piccione and Rubinstein, 2024).

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.

9.3.3 Bias elicitations

In all of the bias elicitation questions below, we additionally tell participants how their decision is incentivised (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 utilise 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 Tversky 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:

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

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 Overweighting We adapt this question from the base-rate neglect cognitive task in Enke et al. (2023). AI overweighting (AO) is calculated as the ratio of the response to the AI overweighting question (see below) to the response to the base-rate neglect question (see above):

$$AO = \frac{\text{AIOverweight}}{\text{baseRateNeglect}}$$

An AO value of 1 indicates no AI overweighting. If $AO > 1$, it suggests AI overweighting. If $AO < 1$, it indicates AI underweighting.

> 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 utilise 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. Because he has never seen the bucket, Charlie does not see it either, but **you** are asked to produce an estimate of its weight. Now you talk to Ann and Charlie. They 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 utilise 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 colour 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 (2024) 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 realises 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?**

9.4 Empirical Evidence—Main Experiment

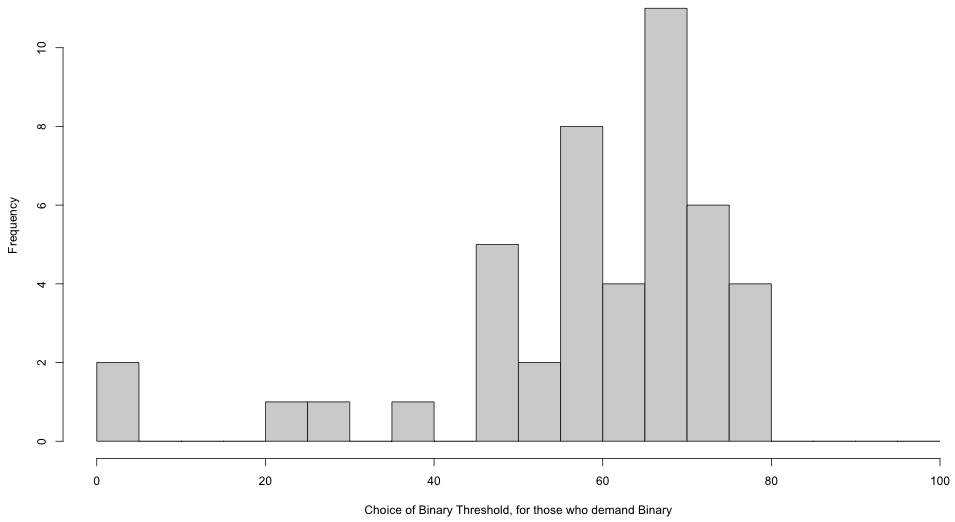
9.4.1 Demand

Table 5: Participant Demographics by AI Assistant Choice

	Binary	No AI	Probability
Age (years)	41.51	47.31	43.87
Female (%)	64.44	62.86	52.17
Bank/Credit Union Experience (%)	37.78	31.43	33.33
Applied to Loans Before (%)	37.78	31.43	34.78
Loan Amount (USD$)	179365	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.

Figure 24: Chosen binary threshold among participants who opted for a binary AI signal



Note: Histogram shows the distribution of the binary threshold values selected by participants who opted for a binary AI signal.

9.4.2 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 8) and the "Overall Accuracy" columns of these true/false positive and negative summary tables (e.g., Table 9) arise because the former incorporate fixed effects, while the latter report simple means.

Gender

Table 6: 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 7: Share of True/False Positives and Negatives by Treatment and Gender

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% 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 8: 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	-0.061^{***}	-0.058^{***}	-0.003
	(0.022)	(0.020)	(0.030)
Probability	-0.033	-0.028	0.0002
	(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
	(0.020)	(0.020)	(0.027)
Observations	3,950	3,500	7,450
R^2	0.306	0.323	0.283
Adjusted R^2	0.281	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 9: Share of True/False Positives and Negatives by Treatment and Age

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% 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	48.10	19.37	14.56	17.97	67.47	790
Binary Med (50)	53.11	19.25	15.22	12.42	72.36	644
Probability	48.82	20.06	14.85	16.27	68.88	633

Years of Loan Specialist Experience

Table 10: 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 11: Share of True/False Positives and Negatives by Treatment and Years of Loan Experience

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% 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 Private Information

Table 12: 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	-0.074^{***}	-0.042^{*}	0.035
	(0.019)	(0.024)	(0.030)
Probability	-0.038^{*}	-0.023	0.020
	(0.021)	(0.026)	(0.033)
Binary High (70)	-0.053^{***}	-0.063^{**}	-0.006
	(0.020)	(0.025)	(0.032)
Binary Low (30)	-0.044^{**}	-0.038^{**}	0.011
	(0.021)	(0.019)	(0.027)
Observations	3,800	3,650	7,450
R^2	0.314	0.309	0.283
Adjusted R^2	0.289	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 13: Share of True/False Positives and Negatives by Treatment and Perceived Private Info

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% 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	48.95	18.82	15.26	16.97	67.76	760
Binary Med (50)	56.98	19.32	12.50	11.20	76.30	616
Probability	50.33	19.57	15.13	14.97	69.90	608

Baseline Practice Accuracy

Table 14: 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	-0.064^{***}	-0.046^{*}	0.015
	(0.020)	(0.024)	(0.030)
Probability	-0.034	-0.027	0.013
	(0.021)	(0.027)	(0.033)
Binary High (70)	-0.063^{***}	-0.057^{**}	0.013
	(0.020)	(0.026)	(0.032)
Binary Low (30)	-0.055^{***}	-0.015	0.047^{*}
	(0.017)	(0.022)	(0.027)
Observations	5,100	2,350	7,450
R^2	0.300	0.336	0.283
Adjusted R^2	0.278	0.306	0.277
Note:	Asymptotic standard errors clustered at the individual level in parentheses.		
			*p<0.1; **p<0.05; ***p<0.01

Table 15: Share of True/False Positives and Negatives by Treatment and Practice Accuracy

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% of total)	Overall Accuracy (%)	Total Count
Prac Acc (Below Median)						
Human Only	46.96	20.49	15.78	16.76	67.45	1020
Binary Med (50)	56.67	18.20	12.62	12.50	74.88	824
Probability	50.86	18.64	14.94	15.56	69.51	810
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 16: 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 17: Share of True/False Positives and Negatives by Treatment and Rubinstein

Treatment	TP	TN	FP	FN	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 18: 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	-0.065^{***}	-0.028	0.032
	(0.018)	(0.038)	(0.043)
Probability	-0.027	0.003	0.023
	(0.021)	(0.036)	(0.041)
Binary High (70)	-0.075^{***}	-0.031	0.040
	(0.020)	(0.038)	(0.042)
Binary Low (30)	-0.038^{**}	-0.024	0.009
	(0.018)	(0.030)	(0.034)
Observations	4,300	1,900	6,200
R^2	0.308	0.341	0.289
Adjusted R^2	0.285	0.307	0.282

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

Table 19: Share of True/False Positives and Negatives by Treatment and Balls Urns

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% of total)	Overall Accuracy (%)	Total Count
BU (Below Median)						
Human Only	47.44	19.30	15.93	17.33	66.74	860
Binary Med (50)	56.01	18.25	14.00	11.74	74.26	707
Probability	53.17	17.97	13.99	14.87	71.13	679
BU (Above Median)						
Human Only	50.79	19.21	13.16	16.84	70.00	380
Binary Med (50)	57.81	19.27	10.63	12.29	77.08	301
Probability	49.03	22.08	14.94	13.96	71.10	308

AI Overweighting (AO)

Table 20: 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		

Table 21: Share of True/False Positives and Negatives by Treatment and Automation Bias

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% of total)	Overall Accuracy (%)	Total Count
AO (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
AO (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 22: 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 23: Share of True/False Positives and Negatives by Treatment and Base Rate Neglect

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% 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 24: 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 25: Share of True/False Positives and Negatives by Treatment and Correlation Neglect

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% 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

9.4.3 Effective action thresholds

9.4.4 Algorithmic adherence

Figure 27 illustrates the share of loans approved across different treatments and thresholds. On average, loan officers approve 67% of loans, as indicated by the dashed horizontal line. 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.

Figure 27 confirms this expectation, as evident in the upward-sloping blue 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 (30)* threshold, the figure plots the approval rate when the AI score exceeds 30. The purple bar represents the average loan approval rate across the full *Probability* condition.

These findings confirm that participants respond to AI-calibrated thresholds. At the same time, the fact that approval rates for *Binary No* (when the binary AI recommends "No") and *Binary Yes* (when the binary AI recommends "Yes") conditions are not 0% and 100%, respectively, shows that loan specialists do not blindly follow AI recommendations.

9.4.5 Where loan specialists are performing well

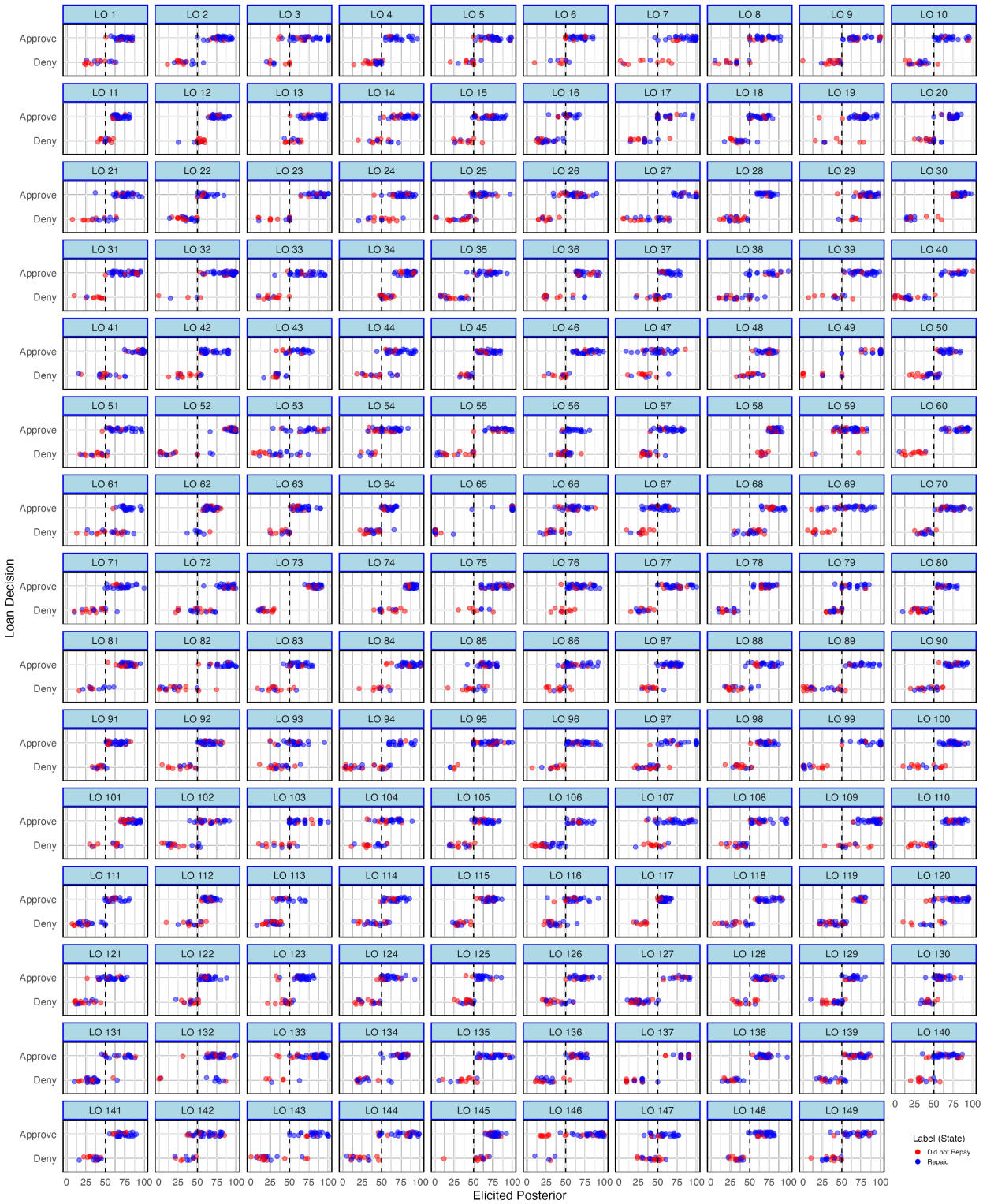
Table 26-28 present the shares of true/false positives and negatives across treatments. These results are broken down by the main treatment arms *Binary Medium*, *Human-only*, and *Probability*, and further categorised by buckets of AI signals s and human-elicited signals r (where available), offering insight into how treatments influence decision-making accuracy across different levels of signal strength.

Table 26: 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)	56.08	18.86	13.32	11.75	74.94
Human Only	48.59	19.26	15.37	16.78	67.85
Probability	50.88	19.00	14.90	15.23	69.87

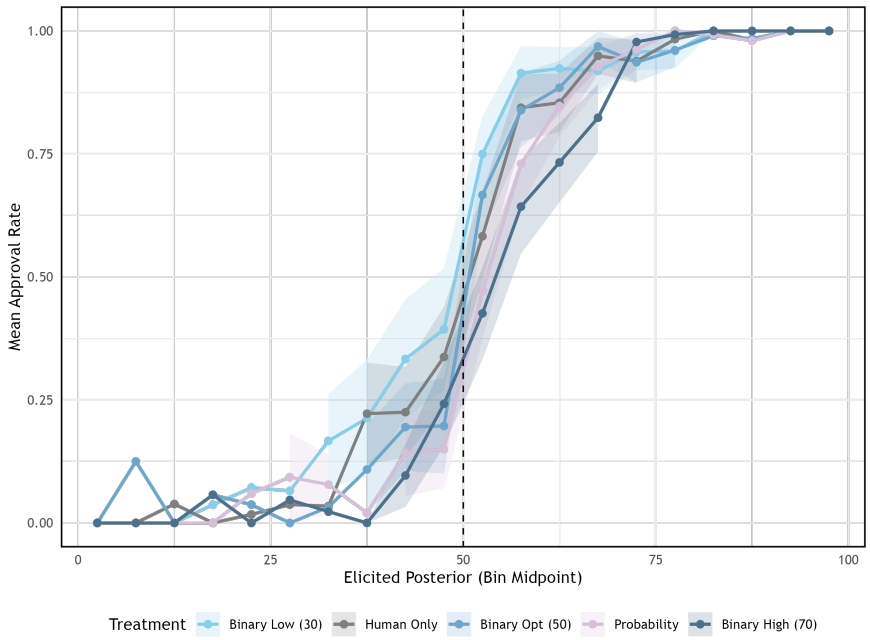
As seen in Table 26, regardless of treatment, loan specialists are much better at approving

Figure 25: Loan Specialists' Effective Action Thresholds



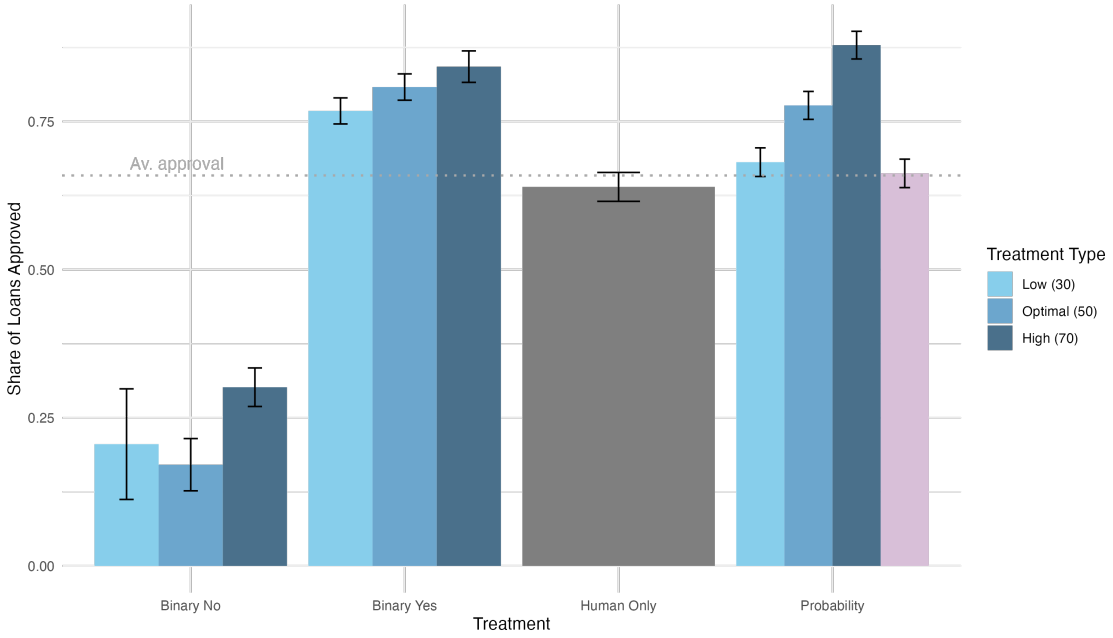
Note: This figure shows the relationship between the elicited posterior belief and the loan decision (approve/deny) for each loan specialist. The blue dots represent applications that were repaid, and the red dots represent those that were not repaid. For most participants, decisions are clearly separated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

Figure 26: 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.

Figure 27: Loan Approval Rates, by Treatment and (implied) AI Recommendation



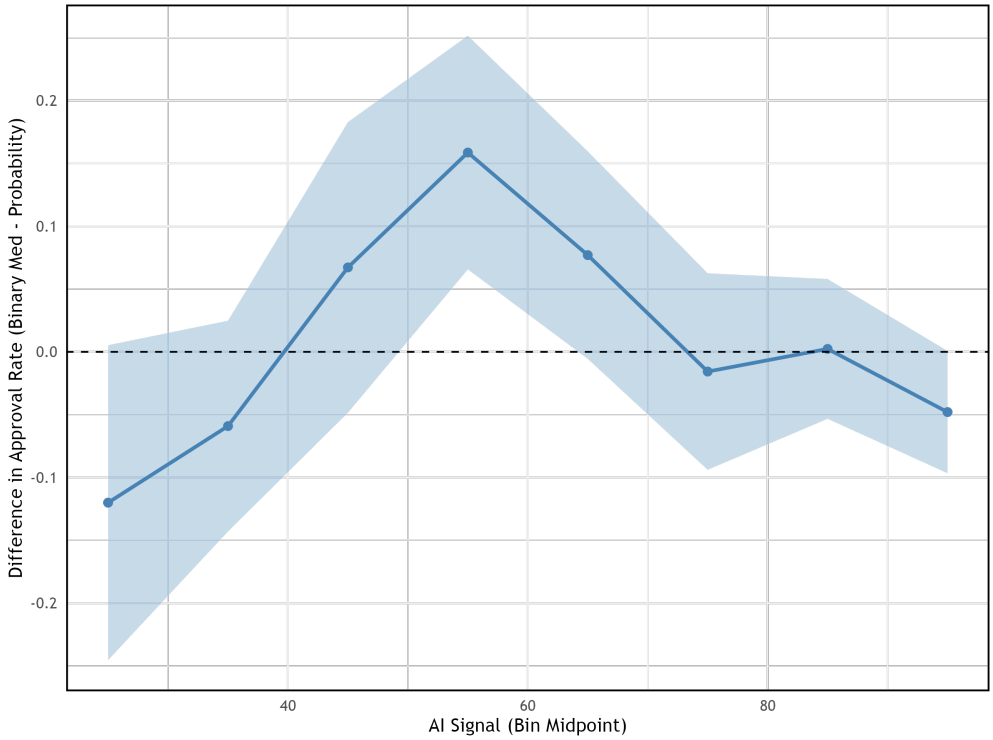
Note: Bars show loan approval rates by treatment. *Binary* conditions are split by whether the AI recommends "Yes" or "No" at thresholds of 30%, 50%, or 70%. *Probability* is split by whether the AI score exceeds each threshold; the purple bar shows the overall average. Error bars indicate 95% confidence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

correctly[45] than denying correctly,[46] despite the fact that they are incentivised equally for true positives and true negatives. This aligns with anecdotal evidence of real-life loan specialists' behaviour: they are often trained to focus on 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). There is typically less emphasis on systematically identifying individuals to reject, leading to a greater focus on approving "cherries" (strong applicants) 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. This suggests that, in this case, the treatment is particularly effective in helping loan specialists identify and approve creditworthy applicants who might otherwise have been incorrectly denied.

Breaking this down further by the human signal inferred from observing the loan application (where elicited), Table 27 shows that loan specialists perform worst when their own signal is low (0-33), i.e., in the implied "Deny" region, and perform best when their own signal is high (67-100).

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

Figure 28: Difference in Approval Probability: Binary Medium vs. Probability



Note: This figure plots the difference in approval probability between the *Binary Medium (50)* 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).

Table 27: 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 (%)
0-33	Binary Med (50)	7.28	50.99	1.99	39.74	58.28
0-33	Human Only	1.13	57.36	0.75	40.75	58.49
0-33	Probability	9.71	49.14	5.71	35.43	58.86
34-66	Binary Med (50)	51.81	21.00	14.29	12.91	72.81
34-66	Human Only	45.20	19.67	14.86	20.27	64.86
34-66	Probability	44.40	21.83	14.74	19.03	66.23
67-100	Binary Med (50)	76.73	6.08	15.72	1.47	82.81
67-100	Human Only	75.13	0.72	22.90	1.25	75.85
67-100	Probability	72.93	4.96	18.39	3.72	77.89

The AI treatments do not appear to have positive effects on accuracy when the human signal is low (0-33). The *Binary Medium* treatment offers the greatest improvement (6.6pp over *Probability* and 8.0pp over *Human-only*) when the human signal falls in the uncertain range (34-66), suggesting that this treatment is particularly effective in aiding decision-making under uncertainty by reducing the share of cases falsely denied when participants are uncertain.

We can alternatively break this out by the underlying AI risk score of each loan case, as seen in Table 28. 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.[47] The *Binary Medium* treatment shows the greatest improvement over *Probability* when the AI score falls in the ambiguous range (34-66). 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.

9.5 Robustness checks - Main Experiment

9.5.1 Restricting to LinkedIn (NAMU/NAMP inclusive) and Indeed participants

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

[47] 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 28: 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 (%)
0-33	Binary Med (50)	0.00	90.52	9.48	0.00	90.52
0-33	Human Only	0.00	80.72	19.28	0.00	80.72
0-33	Probability	0.00	87.12	12.88	0.00	87.12
34-66	Binary Med (50)	41.63	24.65	18.37	15.35	66.28
34-66	Human Only	38.95	21.92	19.57	19.57	60.87
34-66	Probability	35.76	22.08	16.56	25.61	57.84
67-100	Binary Med (50)	75.26	2.56	10.71	11.46	77.83
67-100	Human Only	65.93	4.15	11.53	18.39	70.08
67-100	Probability	73.11	1.97	14.10	10.82	75.08

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 29: 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 29 presents the results shown in Table 2, 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*.

9.5.2 Whether AI or human signal came first, elicitation of first signal

Recall that we implemented case-level randomisation 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 30, neither the order of the AI signal nor the elicitation of posterior beliefs

significantly influences the outcomes.

Table 30: Robustness Checks

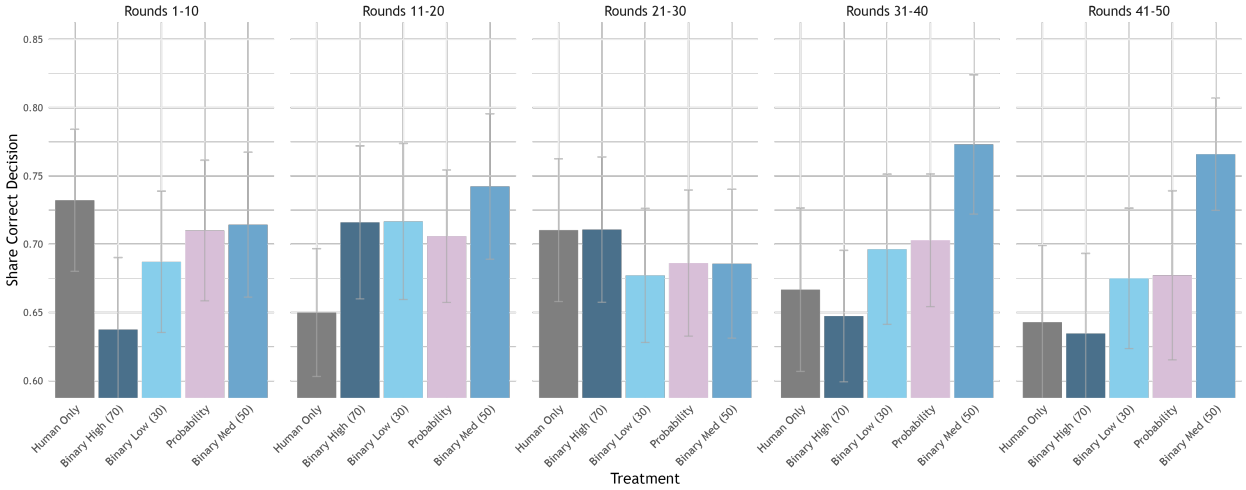
	Dependent variable:			
	Loan Decision	Posterior Belief	Loan Decision	Posterior Belief
	(1)	(2)	(3)	(4)
First Signal	−0.023	−1.081		
	(0.019)	(0.984)		
Elicitation Order			0.001	−0.424
			(0.018)	(0.841)
Constant	0.680***	60.454***	0.658***	59.855***
	(0.020)	(1.112)	(0.020)	(1.066)
Observations	7,450	7,450	7,450	7,450
R^2	0.0002	0.0002	0.00000	0.00003
Adjusted R^2	0.0001	0.0001	-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.

9.5.3 Restricting to each decision block

Figure 29 presents the results shown in Figure 6, breaking down the results by each decision block. *Binary Medium* outperforms *Probability* in all of these decision blocks, indicating that its advantages hold regardless of 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.

Figure 29: 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.

9.6 Personalisation Experiment Design

9.6.1 Personalisation Experiment Additional Instructions

Figure 30: Instructions for *Trinary High (50, 70)* condition in the personalisation 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 31: Instructions for *Binary Medium (50)* condition in the personalisation experiment

Instructions for next 2 tasks More details

Instructions - Assistance from Bailey the AI

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



Information on the AI-generated assistance

- For cases where the applicant **will repay the loan on time**:
  The AI assistant tells you **"This applicant's score is between 50 and 100"** 91% of the time
  The AI assistant tells you **"This 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 **"This applicant's score is between 50 and 100"** 41% of the time
  The AI assistant tells you **"This applicant's score is between 0 and 50"** 59% of the time

Next

9.7 Empirical Evidence—Personalisation Experiment

9.7.1 Main results

Table 31 presents regression results from Equation 2, using the *Personalised* treatment as the reference level; there are no significant differences in approval accuracy between the different treatment conditions.

Table 31: Decision-making outcomes with and without fixed effects

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Binary Medium (Randomised)	−0.002	−0.002
	(0.015)	(0.015)
Trinary High (Randomised)	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

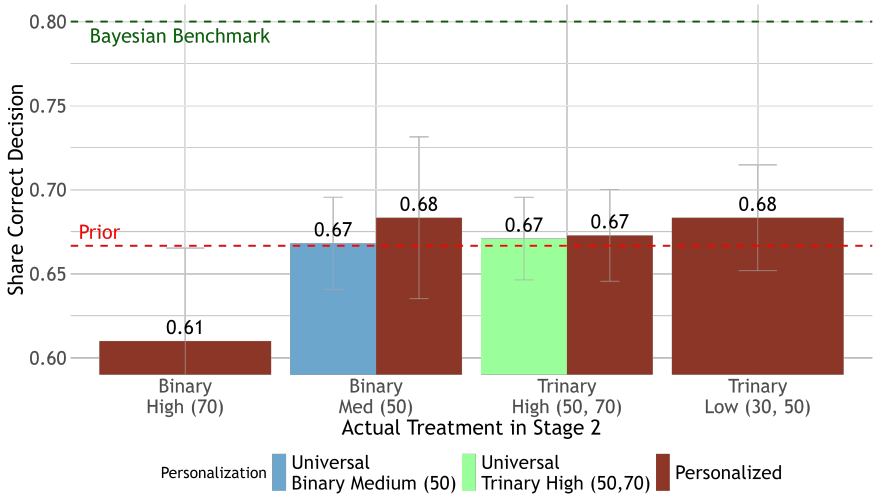
Note: Regression estimates per Equation 2 of loan approval accuracy by treatment condition, with *Personalised* as the omitted reference group. Column (1) reports estimates without fixed effects; Column (2) includes application (case-level) fixed effects. We do not include individual or block order fixed effects because the personalisation 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 13b 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 Personalised condition (model recommending that treatment). If personalisation 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 32 shows slightly higher accuracy in the Personalised 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 personalisation worsens outcomes.

Figure 32: Experiment 2 Results by Actual Treatment



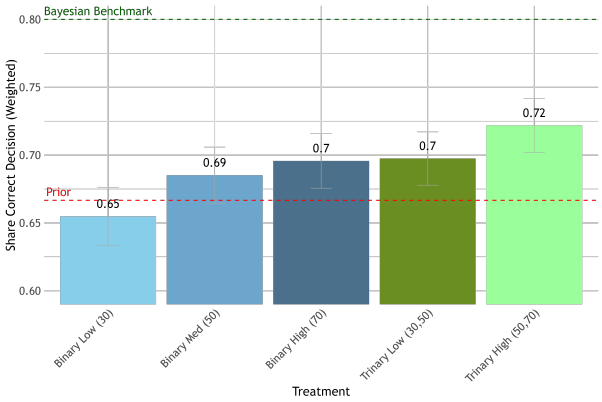
9.7.2 Main results, reweighted

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 (50)* 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.

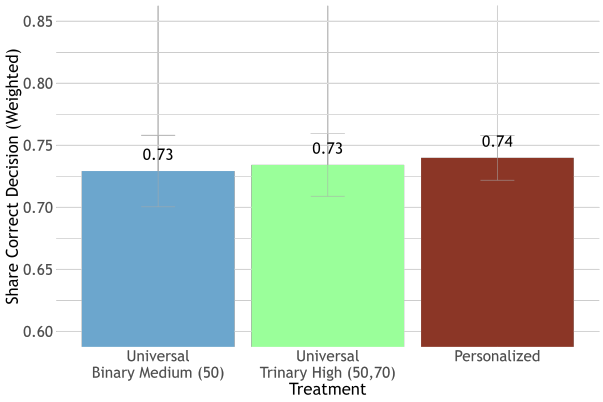
9.7.3 Attrition

Table 32 reports detailed attrition rates by final treatment condition and whether the treatment was assigned via personalisation. Attrition is notably higher in the non-personalised *Trinary High* (23.7%) and *Binary Opt* (25.5%) groups. By contrast, the Personalised 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 Personalised condition may have been systematically weaker, we find no evidence of such bias: Stage 1 performance among attritors is not statistically different from that of participants who remained.

Figure 33: Stage 1 and Stage 2 results (reweighted), personalisation 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 32: 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	45	23.7%
Trinary High	Personalized	45	37	17.8%
Trinary Low	Personalized	39	25	35.9%

To probe this further, Table 33 compares Stage 1 performance across treatment groups. Participants in the Personalised condition have slightly lower Stage 1 accuracy (0.664) than those in the Universal *Trinary High* group (0.668), though the difference is small and not statistically significant (SE = 0.005–0.007).

Table 33: 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)	0.668	0.007	45
Personalized	0.664	0.005	84

Taken together, these patterns suggest that while attrition varies across groups, there is no indication of systematic performance differences among attritors, and the small average differences

in Stage 1 accuracy are insufficient to account for the lack of observed gains from personalisation.

9.7.4 Model calibration

Table 34 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 0.442 to 0.577. This implies that when the model predicts a participant will perform 1 percentage point better under their assigned treatment, they actually perform 0.5 percentage points better on average. These results remain stable when controlling for treatment fixed effects, suggesting that the model captures perfomance variation consistently across treatments.

9.7.5 Placebo test: Does the model perform better than random?

To verify that the model's predictions are not simply an artefact 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 treatments. If the model is working well, it should not predict gains to personalisation 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.

9.7.6 Is the model mispredicting r?

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

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

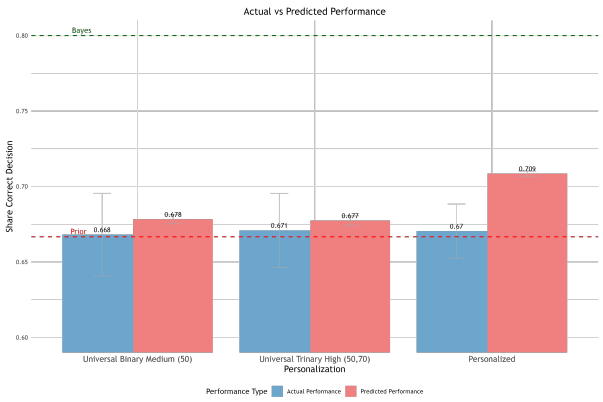
However, since we observe $\hat{r}$ for 90% of the observed cases in Stage 2, we can plug in $\hat{r}$ instead of treat r as latent. Figure 34a shows that even when we plug in the observed $\hat{r}$ directly, the model continues to predict sizable gains from personalisation.

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

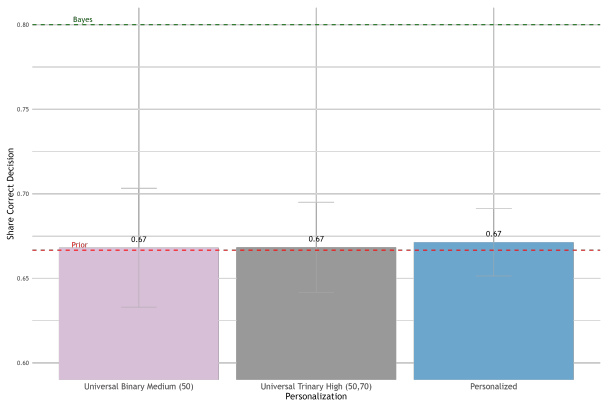
Table 34: Quality of Fit: Stage 2 Accuracy on Model-Predicted Accuracy

	Dependent variable:					
	Stage 2 Accuracy					
	Basic	Pers.	Baseline	Demog.	AI Exp.	Treatment FE
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted Accuracy	0.486***	0.577***	0.442***	0.471***	0.474***	0.474***
	(0.144)	(0.157)	(0.161)	(0.164)	(0.164)	(0.168)
Personalized		−0.019	−0.013	−0.014	−0.014	−0.008
		(0.013)	(0.013)	(0.013)	(0.013)	(0.016)
Baseline Performance			0.075***	0.074***	0.074***	0.068**
			(0.027)	(0.027)	(0.027)	(0.028)
Age				0.001	0.001	0.001*
				(0.001)	(0.001)	(0.001)
Underwriting Years				−0.00004	−0.0001	−0.00004
				(0.001)	(0.001)	(0.001)
Female				−0.002	−0.002	−0.001
				(0.012)	(0.012)	(0.012)
AI Never					0.013	0.011
					(0.024)	(0.025)
Actual Treatment (Binary Med)						0.048
						(0.029)
Actual Treatment (Trinary High)						0.031
						(0.028)
Actual Treatment (Trinary Low)						0.040
						(0.029)
Constant	0.339***	0.287***	0.127	0.079	0.076	0.052
	(0.098)	(0.104)	(0.117)	(0.122)	(0.122)	(0.125)
Observations	167	167	167	167	167	167
R^2	0.065	0.077	0.118	0.133	0.134	0.153
Adjusted R^2	0.059	0.065	0.102	0.100	0.096	0.098

Note: *p<0.1; **p<0.05; ***p<0.01



(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 34: Stage 2 predictions and outcomes using observed $\hat{r}$.

participants whose treatments differ across the two methods. As shown in Figure 34b, this does not improve performance—suggesting that misprediction of r is unlikely to explain the lack of observed gains from personalisation.

9.7.7 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 behaviour and generalises 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 personalised 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 personalisation.

9.7.8 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.[48]

9.7.9 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 35 reports results. The dependent variable is the share of correct decisions at the specialist-treatment level. Column (1) shows that participants performed significantly better in their 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 generalisation. 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.

[48] 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.

Table 35: Decision-making outcomes under different treatment conditions

	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. Case fixed effects included but 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 personalisation, we find no significant evidence of model fit in this setting.

Table 36: 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 personalised.

9.7.10 S1 and S2 adherence as moderators

Figure 35 plots the joint distributions of adherence curve differences across the two stages.[49] 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 behavioural stationarity between stages.

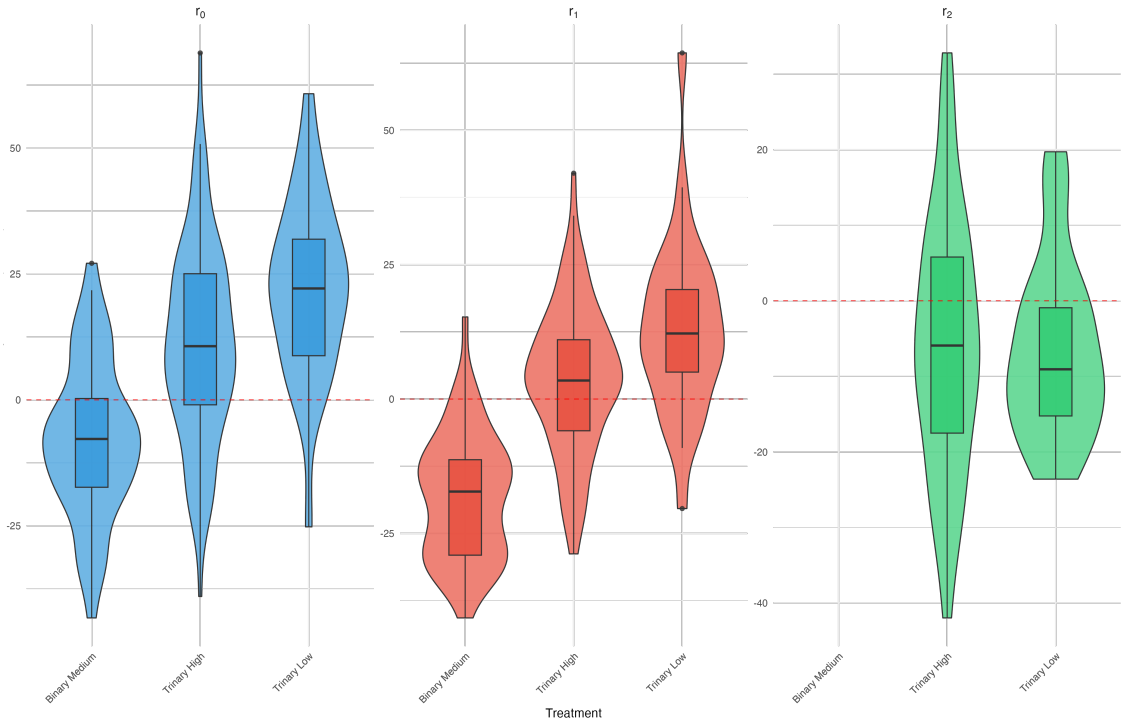


Figure 35: Comparison of r_0, r_1 and r_2 adherence estimated from Stage 1 vs. Stage 2

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 personalisation. Table 37 splits participants by whether their average absolute adherence difference (across r_0, r_1, and r_2) is above or below the median. Those below median exhibit positive gains from personalisation (though not statistically significant), whereas those above median do not.

[49] 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.

Table 37: Decision-making outcomes, without fixed effects

	Dependent variable:	
	Share of Decisions Correct	
	Below Median Adherence Curve Diff	Above Median
	(1)	(2)
Binary Medium (Randomised)	−0.024	−0.000
	(0.023)	(0.026)
Trinary High (Randomised)	−0.018	−0.000
	(0.023)	(0.022)
Constant	0.693***	0.667***
	(0.015)	(0.013)
Observations	2,459	2,430
R^2	0.001	0.000
Adjusted R^2	-0.0003	-0.001

Note: Regression estimates per Equation 2 of loan approval accuracy by treatment condition, with *Personalised* as the omitted reference group. Column (1) reports results for those below median average absolute adherence difference (across r_0, r_1, r_2); 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 38 formally tests this interaction: when the adherence curve difference is zero, the effect of personalisation is positive and significant at the 10% level. This suggests that stable decision strategies are a necessary condition for realising the benefits of personalisation.

Table 38: Decision-making outcomes with fixed effects

	Dependent variable:
	Share of Decisions Correct
Personalised	0.039^{*}
	(0.021)
Adherence Diff (S2-S1)	0.0001
	(0.001)
Personalised x Adherence Diff (S2-S1)	-0.002
	(0.001)
Constant	0.638^{***}
	(0.032)
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 *Personalised* report mean differences in correctness relative to when unpersonalised, 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 personalisation conditions. Case fixed effects included. Asymptotic standard errors (in parentheses) are clustered at the individual level. $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$.