

Calibrated Coarsening: Designing Information for AI-Assisted Decisions

Bnaya Dreyfuss
Harvard University

Ruru Hoong*
Harvard Business School

Job Market Paper

This version: July 8, 2025
[Click here for latest version](#)

Abstract

Artificial intelligence (AI) signals are increasingly deployed as human decision-making aids across many critical applications, but human cognitive biases can prevent them from improving outcomes. We propose *calibrated coarsening*—partitioning the signal space into fewer cells at optimised thresholds—as a way to improve decision-making outcomes while (i) keeping humans in the loop, (ii) modifying signals without deception, and (iii) adapting flexibly to various cognitive biases and decision-making contexts. Within an optimal information disclosure framework, we derive the approximately-optimal universal coarsened policy for settings where the designer does not observe the decision-maker’s information. We then empirically demonstrate in a randomised experiment involving loan specialists that coarsening AI signals at the theory-derived threshold significantly improves decision-making outcomes, over both the human-only (based solely on the loan application) and continuous AI (assisted with uncoarsened AI risk-score) benchmarks. We uncover substantial decision heterogeneity amongst loan officers, and use a Bayesian hierarchical model to personalise coarsening policies, which can further improve outcomes as past data become available.

*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 John Beshears, Alex Chan, Benjamin Enke, Christine Exley, Kris Ferreira, Matthew Gentzkow, Yannai A. Gonczarowski, Josh Schwartzstein, 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 department, as well as the 2023/2024 NUS Development Grant. This research is also supported by Singapore’s Social Science Research Council Graduate Research Fellowship (SSRC 2025-004), administered by the Ministry of Education, Singapore (MOE). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) 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

The issue lies not in the AI’s predictive performance, but 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 can offer valuable input, its impact on performance ultimately depends on how humans process and integrate it into their decision making.

Behavioural science has long established that humans often 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 can fail to improve outcomes. Since removing humans from the process is not often an option, a critical challenge arises: how can we improve AI-assisted decision-making outcomes? Crucially, we want to do so while retaining humans’ final decision rights. Moreover, we seek to avoid deception, as manipulative strategies have been shown to erode trust and hinder long-term adoption (Mahmud et al., 2022). Lastly, the solution should be easily tailored to different biases, which likely vary both between contexts and across individuals within a given context.

We propose *calibrated coarsening*—i.e., partitioning information into fewer cells at optimised thresholds—as a broad approach to improving decision-making outcomes that: (i) keeps humans in the loop, (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 to colour-coded risk,⁴ underscoring its in-

¹e.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).

²e.g., Clinicians remain liable in clinical settings (Greenes, 2011; Mezrich, 2022; Cestonaro et al., 2023); employers maintain human involvement to mitigate litigation risks surrounding disparate impact, where a facially neutral procedure disadvantages a protected class per Title VII Civil Rights Act (Kassir et al., 2023); and recent regulations—e.g., EU’s AI Act—mandate human oversight in algorithmic decision-making (European Commission, 2024).

³Leading e-commerce firm Zalando uses human oversight in algorithmic pricing for this reason (Huelsen et al., 2024).

⁴E.g., Pretrial risk is scored on a discrete 1–6 scale (see Figure 2 in Angelova et al. (2023)), and dental platform

tuitive appeal. 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.

A key advantage of calibrated coarsening is its context-agnostic nature: AI predictions can be coarsened through a few simple steps, regardless of the forms of human biases at play. This adaptability is all the more important with rapidly evolving AI technologies: as models evolve, so do information structures and the way humans update from them. Coarsening can be *recalibrated* (i.e., information can be re-partitioned) and therefore seamlessly adjust to these shifts, making it resilient to inevitable change.

In this paper, we develop a theoretical framework to derive an approximately optimal coarsening policy, and empirically validate it in a randomised experiment with 150 professional loan officers, underwriters, and processors (henceforth, “loan specialists”). 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. It also reduces time-to-decision, and many participants prefer—and are willing to pay for—coarsened over full signals. Moreover, we document substantial heterogeneity in specialist behaviour, suggesting potential further gains from personalisation. To that end, we estimate a hierarchical Bayesian model to target optimal coarsening policies to individuals based on past decision data.

How can coarsening signals—making them strictly *less* informative—lead to better decisions? Intuitively, when specialists update their beliefs in ways that deviate from full rationality, they may make different decisions than a fully Bayesian specialist (representing the optimal benchmark) would, even when presented with the same information. We refer to these divergences as “disagreement regions.” By *coarsening* information, we can shrink these regions by aligning the non-Bayesian’s actions more closely with the Bayesian specialist’s actions. 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 ones where both agree to approve, pushing the non-Bayesian towards approval. The same logic applies in the converse. In both cases, coarsening effectively pools regions of disagreement with adjacent areas of agreement, leading non-Bayesian specialists to make the same decision as Bayesians more often.

VideaAI colour-codes disease risk into 3 bands instead of providing the raw probabilities (see Appendix Figure 11).

Our theoretical framework generalises this intuition. 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 first forms beliefs about the state based on a privately observed *human signal* (e.g., by reviewing a loan application) and a message from the Sender (e.g., an AI designer or bank manager). The Sender publicly commits to a *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. While the Sender’s and Receiver’s incentives are aligned, their beliefs may not be. In particular, the Sender’s objective function is consistent with the normative benchmark, i.e., Bayesian updating, whereas the Receiver’s beliefs may deviate arbitrarily from Bayes, subject only to mild continuity and monotonicity conditions. This belief misalignment results in distorted choices for the Receiver. Therefore, when choosing a disclosure mechanism, the Sender’s goal is to align the Receiver’s behavior as closely as possible with the normative Bayesian benchmark.

We first show that when the Sender can condition the disclosure mechanism on the Receiver’s signal, the optimal mechanism is a binary coarsening of the AI signal that always yields higher expected utility than full disclosure. We then consider the more realistic case where the Sender cannot condition the mechanism on the Receiver’s signal. Focusing on binary coarsenings, we show that the optimal mechanism is approximately the *AI-only boundary*—the threshold above which a Bayesian decision-maker relying solely on the AI signal would choose to approve. The approximation error depends on (i) how different the “adhering types” (those who follow the AI’s implied recommendation) are from the general population, (ii) how sensitive marginal decision makers are to small changes around the threshold, and (iii) how likely those marginal decision makers are to exist (i.e., the density at the margins).

Our first experiment demonstrates the real-world applicability of this approach by empirically testing whether coarsening signals at the AI-only boundary improves decision-making outcomes. We ask 150 loan specialists to evaluate 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 their own beliefs about the probability of repayment (human signal). In addition, 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 varying 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]”). The binary thresholds—*Low* (30%), *Medium* (50%), and *High* (70%)—include one theory-predicted optimal threshold (50%) and two placebo conditions, enabling us to evaluate the value of theoretically-informed coarsening over arbitrary signal discretisation. Our primary outcome is loan approval accuracy, since participants are incentivised with a \$0.10 bonus for each correct decision (approving a loan that is repaid, denying a loan that is not repaid). In this experimental setting, the human has no residual information beyond the AI; that is, conditional on the AI signal, the human signal is not predictive of repayment. This reflects a growing trend in many real-world decision contexts: as data-rich, multi-modal AI systems proliferate, it is increasingly the case that humans have no systematic informational advantage over AI.⁵ Given our incentive structure and setting, the *Binary Medium* threshold—the AI-only boundary—serves as the approximately-optimal benchmark.

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 more accurate than under *Probability*—closing 40% of the gap to the Bayesian benchmark—and 6 points 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 reduces decision time—a critical factor in high-throughput settings like loan underwriting. Participants in *Binary Medium* made decisions approximately 7% faster than those in the *Probability* condition, likely due to the reduced cognitive effort required to interpret a binary recommendation versus a continuous score. Importantly, many participants preferred coarsened signals: in an incentivised choice task, over half opted for either a binary signal or no AI assistance over a full probability score. This preference is reflected in their willingness to pay: 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 and cognitive processing costs (Guan et al. (2023); Exley and Kessler (2023)). They also echo interviews we conducted with senior lending professionals, who acknowledged the value of AI assessments but found AI tools “challenging to use” and easily “overridden”. Taken together, these findings suggest that simpler, coarsened signals may reduce cognitive burden and encourage

⁵e.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 health-care, 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.

adoption in practice.

Performance gains, however, are not uniform across all specialists, offering insight into the mechanisms that might be driving the outcome gap and corresponding gains from coarsening. Gender emerges as a striking moderator: men benefit substantially more from AI assistance, with improvements nearly four times as large as those for women. This divergence is not due to baseline differences—both genders perform similarly without AI—but reflects gendered differences in willingness to adopt and rely on AI advice. A similar pattern holds for experience: specialists with less industry tenure perform worse unassisted, but improve more with AI, suggesting they are more open to relying on AI recommendations when lacking effective internal heuristics. Perceived private information also shapes outcomes—those who believe they possess less insight (over the AI) are both less accurate at baseline and more responsive to AI, particularly when signals are coarsened. Finally, consistent with theories of automation neglect,⁶ users who exhibit automation neglect on a standard bias elicitation fail to incorporate AI input as effectively. These findings provide strong evidence of heterogeneity, underscoring the potential of leveraging personalisation (e.g., based on demographic and psychological differences) to further improve outcomes.

To that end, our second set of empirical results documents significant potential gains from tailoring coarsening policies to individual decision-makers. We estimate a hierarchical Bayesian model that leverages historical data from an initial experiment to predict each specialist’s decision accuracy across alternative coarsening policies, allowing us to identify optimal personalised treatments. Implementing such a personalised approach would reallocate approximately 45% of specialists to different coarsening thresholds compared to a universal policy assigning the AI-only boundary (*Binary Medium*) to the population, yielding a predicted 2.5 percentage point improvement in decision-making accuracy on average.

Related Literature

Calibrated coarsening is not the only way to address biases in human decision-making, though to our knowledge it is the only one that does so whilst fulfilling the desiderata laid out above. “Debiasing” in the broader decision-making context has traditionally taken three forms ([Roy and Lerch, 1996](#)): (i) modifying information presentation, (ii) replacing the individual with a model, and (iii) training people to use the appropriate information processing strategies. Coarsening falls

⁶The tendency to underweight information that is algorithmic in nature, often modeled as under-responsiveness to signals (e.g., [Grether \(1992\)](#))

within the first. The second is not feasible as it is precisely the constraint we are operating with: in many situations we cannot replace the human entirely with a model. Indeed, one popular approach proposes delegation—assigning different cases to either an AI or unassisted human (Raghu et al., 2019; Mozannar and Sontag, 2020; Athey et al., 2020; Bansal et al., 2021). However, this solution is infeasible in scenarios where humans *must* make the final decision, and may furthermore fall short in fully leveraging the collective information offered by both humans and AI.

The last of these debiasing approaches, training, has been the subject of much literature, most of which suggesting that it is ineffective unless very extensive (e.g., administering statistic courses) and that effects—if any—remain domain-specific (Tversky et al., 1982; Fong et al., 1986; Fong and Nisbett, 1991; Milkman et al., 2009). Several have designed interventions that counter behavioural biases like confirmation bias (Morewedge et al., 2015; Sellier et al., 2019), but these solutions are typically context- and bias-specific, limiting their systematic implementation. Determining what kinds of training are effective is itself costly and challenging—and often yields limited results, as in the case of correlation neglect (Enke and Zimmermann, 2019). Coarsening, on the other hand, is easily adaptable, low-cost, and can be integrated into the system regardless of context or bias—eliminating the need to continually retrain individuals as conditions change.

Our work relates to several fields, including a managerial prediction literature that forecasts the positive impact of AI on organisational performance and managerial decision-making (Brynjolfsson and McAfee, 2014; Davenport and Kirby, 2016; Daugherty and Wilson, 2018; Choudhary et al., 2025), where even modest gains in decision accuracy can yield substantial economic returns (Cockburn et al., 2018; Agrawal et al., 2019). Notably, Agrawal et al. (2022) discuss how AI advancements may force a decoupling of the fundamental components of decision-making—(AI) prediction and (human) judgment⁷—thereby transforming managerial and organisational structures. We suggest that a more complete characterisation of this shift must also consider the cognitive biases that shape how humans interpret and act on AI predictions,⁸ and the organisational tools—e.g., coarsening—available to mitigate them. AI’s purported transformative potential may therefore be more challenging to realise than appears; our paper offers a practical solution to these barriers, irrespective of the specific deviations from Bayesian decision-making that underlie them.

Recent work in marketing also shows that hybrid human-AI approaches can be effective in pricing, hiring and ideation (Karlinsky-Shichor and Netzer, 2024; Chakraborty et al., 2025; Bell et al.,

⁷Per Agrawal et al. (2022), “judgment” here means human preferences or utility over different outcomes.

⁸Even individuals who are well-calibrated in their beliefs may nonetheless select actions that misalign with their beliefs and preferences, resulting in suboptimal decisions.

2024). However, behavioral frictions like algorithm aversion (Dietvorst et al., 2015, 2018; Logg and Schlund, 2024) often hinder effective implementation. Calibrated coarsening addresses this by discretising the AI signal space. Discretisation has also been useful in marketing, where restricting decision spaces—effectively a form of coarsening—has been shown to improve managerial creativity and decision quality (Sellier and Dahl, 2011), and to benefit personalized marketing; for example, Zhang and Misra (2024) demonstrate that firms can achieve near-optimal personalisation with only a limited set of treatment options. We show that optimally coarsening AI outputs can improve real-world decisions, in a framework that also holds promise for improving outcomes in contexts like pricing and hiring.

We also contribute to a large body of experiments that investigate the human use of AI—or more broadly, algorithmic—signals on overall performance and speed (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; Grimon and Mills, 2025; De-Arteaga and Chouldechova, 2020; Lakkaraju and Farronato, 2022; Imai et al., 2023; Ben-Michael et al., 2024; Kim et al., 2024; Snyder et al., 2024), as well as identifying who benefits most from their provision (Caplin et al., 2024). A growing subset of this literature pertains to applied AI alignment—how to structure AI inputs and interfaces so that human users act on them in ways that improve outcomes in practice. This includes work on behavioural frictions that may impede human-AI performance, including algorithmic appreciation or over-adherence (Logg et al., 2019; Banker and Khetani, 2019; Bai et al., 2021; Buçinca et al., 2021; Fügner et al., 2021), rational inattention (Boyacı et al., 2024), prospect theory (Ye et al., 2022), and correlation neglect (Agarwal et al., 2023). Given the multitude of cognitive biases at play, a key advantage of coarsening is its ability to simultaneously address many of them. Rather than seek to isolate and address any single bias, our approach remains agnostic, increasing the ease with which it can be applied and tailored across contexts and individuals.

There is also an emerging literature that proposes methods to improve the outcomes of collaborative human-AI design, including providing explanations and information about when AI errs and its uncertainty (Bansal et al., 2019; Green and Chen, 2019; McGrath et al., 2020; Lakkaraju and Bastani, 2020; Taudien et al., 2022), or how to select the best AI model to complement human decisions (Guo et al., 2025). Several propose using adaptive techniques to accommodate human information, increase compliance, and optimise when to provide advice (Sun et al., 2022; Caro and de Tejada Cuenca, 2023; Noti and Chen, 2022; Ibrahim et al., 2021; McLaughlin and Spiess, 2024). We relate most closely to two papers that apply information design to improve human-

AI collaboration. [McLaughlin and Spiess \(2022\)](#) (henceforth MS) study a principal-agent model where agents incur additional penalties for making an error when deviating from the AI recommendation. MS find that withholding uncertain signals can improve outcomes—effectively a form of coarsening where extreme signals are pooled into a recommendation while the center is fully disclosed without a recommendation. In a different model that accounts for a broad range of belief updating biases, we find that calibrated coarsening can improve outcomes, even when the agent’s payoff depends only on true and false positives and negatives, rather than on compliance with AI.

Most relevant to our context is concurrent work by [Agarwal et al. \(2025\)](#) (henceforth AMW), who propose a framework for how and when to automate or assist decisions with AI—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 continuity and monotonicity of actions in signals (i.e., higher signals increase likelihood of Receiver action), whereas AMW do not impose structure on this mapping. Instead, AMW assume that human judgments depend on the AI prediction only via the resulting posterior on the state. This assumption 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 in a study of Prolific participants performing fact-checking tasks that coarsening AI is not beneficial relative to full disclosure. By contrast, we find in a study of loan specialists with prior industry experience that coarsening improves decision-making accuracy relative to full disclosure.

Our work also relates to a broader literature in organisational behaviour, psychology, sociology and other fields on collaboration and the transformation of knowledge ([Carlile, 2004](#); [Hardy et al., 2005](#); [Levina, 2005](#)). There is a burgeoning corpus exploring the role of AI technologies in generating and evaluating knowledge claims in human-AI collaboration ([Faraj et al., 2018](#); [Von Krogh, 2018](#); [Rai et al., 2019](#); [Kellogg et al., 2020](#); [Anthony et al., 2023](#)), and in particular on how individual users engage with AI tools in practice ([Christin, 2020](#); [Pachidi et al., 2021](#); [Lebovitz et al., 2022](#)), as well as in contexts where fairness and discrimination play critical roles ([Kelley et al., 2022](#)). One strand of this literature also focuses more specifically how individuals seek and

⁹Examples 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](#)).

utilise advice (Yaniv, 1997; Yaniv and Kleinberger, 2000; Yaniv, 2004; Surowiecki, 2005; Soll and Larrick, 2009; Sah and Loewenstein, 2015), where second opinions are shown to be able to improve judgments as well as outcomes like cancer diagnosis (Staradub et al., 2002; Taylor and Potts, 2008). AI or algorithmic inputs can also be seen as a form of second opinion. In fact, our framework broadly applies to any form of human decision-making involving the aggregation of multiple signals; coarsening any form of second opinion can improve outcomes. We focus specifically 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.

Finally, we also rely upon the literature on (non-)Bayesian persuasion and information design (Kamenica and Gentzkow, 2011; Alonso and Câmara, 2016; Kamenica, 2019; Levy et al., 2022; de Clippel and Zhang, 2022), as well as optimal information disclosure (Kolotilin, 2018). Like us, Aybas and Turkel (2019) study coarse communication, but in their framework, coarsening is a *constraint* on the signal space that limits the Sender persuasion. In our framework, calibrated coarsening is a *tool* that can improve welfare by pooling signals to counteract biased belief updating. More broadly, the persuasion and disclosure literature has been largely theoretical; only a few studies have applied informational design in empirical settings (Decker, 2022; Xiang, 2024), fewer still in the context of AI (MS, AMW). Our work is the first to provide experimental evidence that the implementation of coarsening to address biases indeed improves outcomes.

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 our first experiment, which tests a universal coarsening policy with loan specialists, and Section 5 presents the results. Section 6 introduces a Bayesian hierarchical model of decision-making and outlines a method for personalising policies. Section 7 concludes.

2 Model

2.1 Basic framework

We focus on binary classification problems where a *Receiver* (e.g., loan specialist) must choose between two actions $a \in \{0, 1\}$ (e.g., denying or approving the loan). There is a binary state $\omega \in \Omega = \{0, 1\}$, which, in our context, reflects the repayment outcome—whether the loan applicant

will repay the loan on time.

The Receiver has type $r \in R \subseteq \mathbb{R}$, which can be interpreted as the *human* signal the loan specialist receives based on their review of the loan application. The *Sender* (e.g., a bank manager who controls the output of an AI system or the designer/engineer of the system) has type $s \in S \subseteq \mathbb{R}$ corresponding to the AI *signal* or risk score of the applicant, which is distributed according to $F_\omega(s)$. We allow the signals to be correlated even conditional on the state ω : conditional on s and ω , r is distributed according to $G_\omega(r|s)$. Both marginals are continuously differentiable and have strictly positive densities $f_\omega(s)$ and $g_\omega(r|s)$. For notational convenience, we denote distributions integrating over the state by omitting the subscript ω , e.g., $g(r|s) = \int_\Omega g_\omega(r|s)P(\omega) d\omega$.

We assume that the Sender and Receiver are incentive-aligned, and share the same payoff function conditional on the state, though the framework—and all of the results—still apply if we relax this assumption and allow the Sender and Receiver to have misaligned utilities. However, we would then have to handle the Receiver’s potential strategic considerations. Without loss of generality, we normalize the payoff from inaction ($a = 0$) to 0. The payoff from action ($a = 1$), conditional on the state ω , is given by:¹⁰

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

where c_{FP} captures the cost of a wrongly-approved loan ($a = 1, \omega = 0$), and c_{TP} benefits of correctly approving an applicant who repays the loan on time ($a = 1, \omega = 1$).

While payoffs are identical, we allow the Sender’s and Receiver’s *beliefs* to diverge, which can in turn generate a wedge between their (subjective) optimal actions. Although the Sender cannot take actions directly, they can influence the Receiver by publicly choosing a fixed, deterministic and monotonic *information disclosure mechanism* $\Phi : S \rightarrow \mathbb{R}$.

Let $\pi(r, s)$ denote the Bayesian posterior over ω given the pair of signals r, s . We denote the posterior induced by a signal r and a message $\Phi(s)$ (generated by the disclosure mechanism Φ) by $\pi_\Phi(r, \Phi(s))$.¹¹ The Sender’s objective function aligns with Bayesian updating: given r , their

¹⁰This equation is derived by normalising inaction without loss of generality from the utility function given by: $\hat{u}(a|\omega) = -c_{FP}\mathbb{I}(a = 1, \omega = 0) + c_{TP}\mathbb{I}(a = 1, \omega = 1) - c_{FN}\mathbb{I}(a = 0, \omega = 1) + c_{TN}\mathbb{I}(a = 0, \omega = 0)$

¹¹Formally, the Bayesian posterior given a pair of raw signals is $\pi(r, s) = \frac{p \cdot g_1(r|s)f_1(s)}{g(r|s)f(s)}$, and the posterior given a disclosure mechanism Φ , a signal r and a message $\Phi(s) = k$ is

$$\pi_\Phi(r, k) = \frac{\int_{s:\Phi(s)=k} \pi(r, s)g(r|s)f(s)ds}{\int_{s:\Phi(s)=k} g(r|s)f(s)ds}.$$

posterior belief would be $\pi(r, s)$. The Sender is the normative benchmark: Bayesian updating is the best one can perform given the information available. We assume that the joint distribution of s and r satisfies the multivariate monotone likelihood ratio property (MLRP) in ω . Given that Φ is restricted to be monotonic, this implies the posterior $\pi_\Phi(r, \Phi(s))$ is monotonically increasing in both $\Phi(s)$ and r , and continuously differentiable in r , as well as in $\Phi(s)$ whenever Φ is continuous.

The Receiver’s posterior—given a disclosure mechanism Φ —on the other hand, is denoted by $\tilde{\pi}_\Phi(r, \Phi(s))$. In the special case where $\Phi(s) = s$ it is denoted by $\tilde{\pi}$. In general, we allow $\tilde{\pi}_\Phi$ to deviate from Bayesian updating, but remain agnostic about the exact nature of this deviation, whether driven by belief misspecification or any particular underlying cognitive bias. The only assumption we impose on $\tilde{\pi}_\Phi$ is the following:

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, e.g., updating in the wrong direction, interpreting a higher signal as indicative of a *lower* likelihood of $\omega = 1$. Most commonly studied models of biased belief updating, such as Grether’s (1992) under- or over-reaction and correlation neglect (Enke and Zimmermann, 2019), do not violate Assumption 1.

The Sender’s and Receiver’s mappings from signals to posteriors are allowed to depend on the information disclosure mechanism Φ chosen by the Sender (in addition to taking $\Phi(s)$ as an argument). Importantly, notice that many standard models of deviations from Bayesian updating—such as confirmation bias or base-rate neglect—generate a class of posteriors $\tilde{\pi}_\Phi$ that are all well defined and internally consistent for any Φ .

The timing of the communication game proceeds as follows: the Sender first publicly selects a disclosure mechanism Φ . Next, the state ω is realized, and the signals r and s are drawn according to F_ω and G_ω . The Receiver then observes the pair $(\Phi(s), r)$ and takes an action a . Finally, both Sender’s and Receiver’s utilities are realised.

We denote the true expected utility given r and s —which also reflects the Sender’s subjective expected utility—by $v(s, r) = E_\pi(U(\omega))$. In contrast, given a mapping Φ , and a pair of observed signals $\Phi(s), r$, the Receiver chooses $a = 1$ iff

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

By Assumption 1, we get the following result:

Result 1 (*Single crossing*). *Let Φ be a monotonic information disclosure mechanism. Then there exists a **decision boundary** $b_{\tilde{\pi}_\Phi}(\Phi(s))$ such that the Receiver chooses $a = 1$ if and only if $r \geq b_{\tilde{\pi}_\Phi}(\Phi(s))$, where the boundary $b_{\tilde{\pi}_\Phi}(\Phi(s))$ is decreasing in $\Phi(s)$.*

The decision boundary $b_{\tilde{\pi}_\Phi}$ is a (decision) function induced by a (posterior) function. A Bayesian Receiver's decision boundary is denoted by b_π . Since in general, $b_{\tilde{\pi}} \neq b_\pi$, the Receiver's decisions as a function of the signals deviates from the Bayes-optimal benchmark. This discrepancy is central to our framework: due to cognitive limitations or misspecified beliefs, the Receiver makes suboptimal choices. Because decisions based on b_π yield the first-best outcomes from a Bayesian perspective—that is, they are optimal given the (full) available information—the Sender seeks to design a disclosure mechanism Φ such that $b_{\tilde{\pi}_\Phi}$ is as close as possible to b_π . We say that the Sender can *fully implement* a decision boundary b if there exists a disclosure mechanism Φ such that $b_{\tilde{\pi}_\Phi}(\Phi(s)) = b(s)$ for all $s \in S$. In other words, for any pair (s, r) , Φ induces the Receiver to make the same decision as that implied by the boundary b .

2.2 Disclosure mechanisms

In what follows, we focus on the following disclosure mechanisms:

- A *K-coarsening at a set of thresholds T*, denoted by Φ_T , is a mechanism that generates K messages $\mathcal{M} = \{0, \dots, K-1\}$. Each $s \in S$ is mapped to a natural number k according to:

$$\Phi_T(s) = k \quad \text{for all } s \in [t_k, t_{k+1}), k \in \{0, \dots, K-1\}$$

where $t_k < t_{k+1}$ for all k , $t_0 = \underline{s}$, and $t_K = \bar{s}$. We denote the $K-1$ thresholds chosen by the Sender to coarsen the signal space by $T = \{t_1, \dots, t_{K-1}\}$.

- A *full revelation mechanism*, denoted by Φ_s , is a limit case of a *K-coarsening* that generates a different message for all $s \in S$, i.e., $K = |S|$.
- A *full censorship mechanism*, i.e. a *1-coarsening*, denoted by Φ_\emptyset , is one that generates the same message $\Phi(s) = m$ for all $s \in S$.

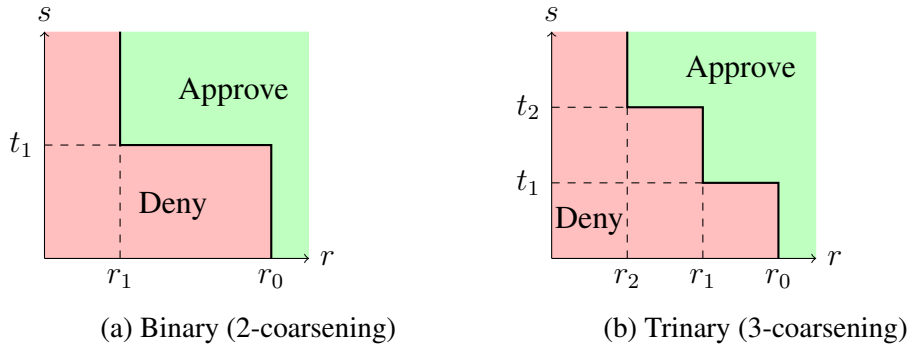
To simplify notation, from this point on when referring to posteriors, we will omit Φ from the subscript (and as noted above, when referring to the posterior under full revelation, we omit the

subscript altogether). For example, $\tilde{\pi}_{\Phi_T}$ becomes $\tilde{\pi}_T$, and $\tilde{\pi}_{\Phi_s}$ is simply $\tilde{\pi}$. Additionally, we refer to the action taken by a type- r Receiver under full censorship—i.e., in the absence of any information about s —as their *default action*, denoted by $a_\emptyset(r)$.¹²

For any K -coarsening, a message $k \in K$ induces type- r Receivers to choose $a = 1$ iff $r \geq r_k \equiv b_{\tilde{\pi}_T}(k)$. We illustrate these below for 2 and 3-coarsening in Figures 1a and 1b. Consider for example the 2-coarsening—or henceforth, *binary coarsening*—at threshold t_1 , illustrated in Figure 1a. This disclosure mechanism generates two messages, 0 for all $s \in [\underline{s}, t_1)$, and another message 1 for all $s \in [t_1, \bar{s}]$. This disclosure mechanism pools together all signals below the threshold t_1 , and all signals above the threshold. A Receiver with $r \in [r_1, r_0)$ is said to *adhere* to the message, since they choose $a = 1$ if and only if $\Phi(s) = 1$; if $r < r_1$, they will always deny ($a = 0$), if $r \geq r_0$, they will always approve ($a = 1$). Figure 1b has a similar illustration for $K = 3$.

Importantly, for any given Φ , the set of $K - 1$ thresholds T and the corresponding set of K boundaries $\{r_0, \dots, r_{K-1}\}$ are sufficient to fully characterise the actions of the Receiver, while imposing minimal structure on the precise updating bias. This later becomes useful in our empirical setting, where we can directly estimate $\{r_0, \dots, r_{K-1}\}$ from observed Receiver decisions.

Figure 1: Two examples of K -coarsening mechanisms



2.3 Optimal mechanisms

2.3.1 Optimal disclosure when $\tilde{\pi} = \pi$

When the Receiver's and Sender's full-revelation beliefs are aligned—that is, when the Receiver acts as a Bayesian updater under full revelation, the Sender's optimal mechanism becomes straightforward: fully reveal the signal, effectively disclosing all the information they possess.

¹²Formally, $a_\emptyset(r) = 1$ iff $E_{\tilde{\pi}_\emptyset}(U(\omega) \mid r, m) \geq 0$. Notice that the default action implicitly depends on $\tilde{\pi}_\emptyset$.

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

Proof in [Appendix](#).

2.3.2 Optimal disclosure with known r

We now extend to the case where $\tilde{\pi} \neq \pi$, but where the Sender knows r when choosing Φ . We therefore treat r as a known constant for the rest of this sub-section. We denote the *inverse boundary*, $\hat{b}(r) \equiv b^{-1}(r)$, i.e., the lowest s that induces $a = 1$, given r .

We impose the following structure on the Receiver's posterior (note that results in further sub-sections do *not* depend on this assumption):

Assumption 2. *For any K -coarsened disclosure mechanism Φ_T and integer $k \leq K - 1$:*

1. *The posterior $\tilde{\pi}_T(k, r)$ is a weighted average of the posteriors in the pre-image of k , $\{\tilde{\pi}(s, r) : s \in [t_k, t_{k+1})\}$.*
2. *The posterior $\tilde{\pi}_T(k, r)$ is continuous and strictly monotonically increasing in the thresholds t_k and t_{k+1} .*

Part 1 of Assumption 2 restricts the Receiver's posterior to respect a weak version of the Law of Total Probability: for any r , when the Receiver learns that s lies on a given segment, the resulting posterior is *some* weighted average of the posteriors from each s on that segment. The weights can be misspecified, as long as they are non-negative and sum to 1. This requires some level of consistency, and precludes the Receiver from adopting a posterior belief that lies outside the implied support of s given a coarsened signal. Part 2 ensures that $\tilde{\pi}_T$ is well-behaved: it prevents the weights (and hence the posterior) from changing discontinuously in response to local changes in the thresholds, and ensures that an increase in the segment's endpoints leads to a higher posterior. As with Assumption 1, we are not aware of a commonly used model that violates Assumption 2.

We show that when beliefs diverge, the Sender can always do better than fully revealing s .

Proposition 2. *Let $r \in R$. If Receiver's beliefs deviate from Bayes, i.e., $\tilde{\pi} \neq \pi$, 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 the Sender's inverse boundary is higher (i.e., $\hat{b}_\pi(r) > \hat{b}_{\tilde{\pi}}(r)$) and the default is inaction (i.e., $a_\emptyset(r) = 0$) the Sender can fully implement b_π . Similarly, if the Sender's inverse boundary is lower (i.e., $\hat{b}_\pi(r) < \hat{b}_{\tilde{\pi}}(r)$) and the default is to act (i.e., $a_\emptyset(r) = 1$), the Sender can fully implement b_π .

Proof in [Appendix](#). In other words, contingent on knowing r , the Sender can always construct a binary coarsening that strictly improves outcomes over 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.

Example: Automation Bias (or Neglect) To gain intuition behind this result, we apply it in a specific, well-documented behavioral bias: automation bias (or neglect), i.e., the tendency to overweight (or underweight) signals from AI relative to one's own information, a bias thought to be prevalent in human-computer interaction decision-making contexts ([Alberdi et al., 2009](#); [Agarwal et al., 2023](#)). It is often modeled using the [Grether \(1992\)](#) framework:

$$\tilde{\pi}(s, r; \alpha) = \frac{p \cdot g_1(r' \mid s)[f_1(s)]^\alpha}{p \cdot g_1(r' \mid s)[f_1(s)]^\alpha + (1 - p)g_0(r' \mid s)[f_0(s)]^\alpha},$$

where $\alpha > 0$ captures the subjective weight placed on the AI signal. $\alpha > 1$ implies automation *bias*, over-updating in the direction of the AI signal, whereas $\alpha < 1$ captures automation *neglect*, capturing under-reaction to the AI signal.¹³ In this example, we model the Receiver's posterior in the coarsened case, $\tilde{\pi}_T$, as a (correctly) weighted average of (biased) posteriors:

$$\tilde{\pi}_T(k, r; \alpha) = \int_{s \in S} \tilde{\pi}(r, s; \alpha) f(s \mid r, k) ds,$$

where $f(\cdot \mid r, k)$ is shorthand for the density of s conditional on r and $\Phi(s) = k$.

In what follows, we assume $a_\emptyset(r) = 0$ (the opposite case is symmetric). Denote the Sender's inverse boundary by $\hat{b}_\pi \equiv s^*$. In words, conditional on r , the Sender would choose $a = 1$ iff $s \geq s^*$. Notice that an automation biased (neglectful) Receiver has a lower (higher) inverse boundary $\hat{b}_{\tilde{\pi}(r, s; \alpha)}$. Applying Proposition 2 to our case, we attain the following result:

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

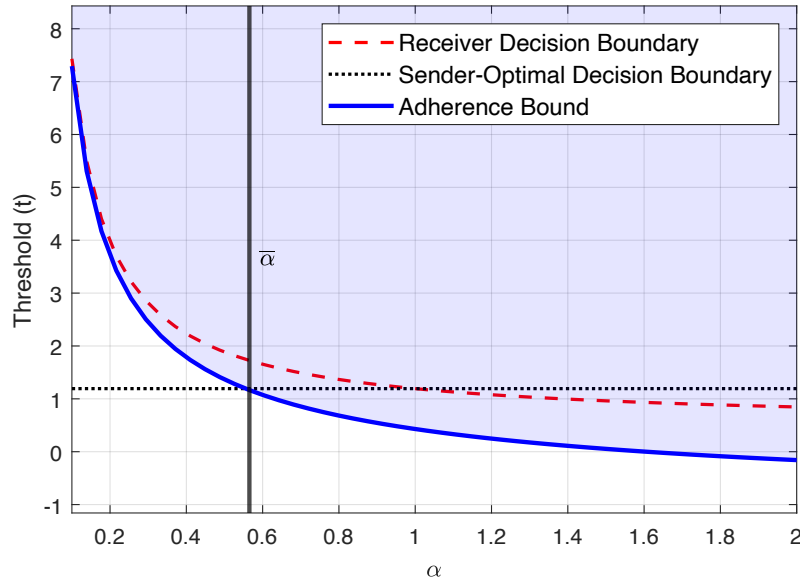
¹³When s and r are jointly normal, automation bias (neglect) is isomorphic to over- (under-) estimating the variance of the AI signal, treating it as more (less) precise than it actually is.

1. For any $\alpha > 0$, the optimal binary coarsening strictly improves the Sender's utility over full revelation.
2. There exists $\bar{\alpha} < 1$ such that if $\alpha \geq \bar{\alpha}$, optimal binary coarsening fully implements b_{π} .

In other words, under *automation bias* ($\alpha > 1$), binary coarsening can always fully implement the full-revelation Bayesian benchmark. Under *automation neglect* ($\alpha < 1$), binary coarsening can always achieve higher utility than full revelation, but full implementation is possible only when α is sufficiently high. For intuition, since the Receiver chooses $a = 0$ in absence of AI information (i.e., $a_0(r) = 0$), a “No” message ($\Phi_t(s) = 0$), which implies lower posterior over s (and thus ω) must also induce $a = 0$. Moreover, since non-biased ($\alpha = 1$) already adhere to the binary message coarsened at s^* , automation-biased agents—who place even greater weight on AI—are *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. In contrast, for sufficiently automation-neglectful agents (i.e., $\alpha < \bar{\alpha}$), the Receiver's signal under-weighting forces the Sender to pick a threshold higher than s^* to induce action, but still improves Sender's utility over full revelation.

Figure 2 plots the threshold t against α , in a case where s and r are jointly normal.

Figure 2: Automation neglect when r is known: graphical illustration



Note: The parameters used in this simulation: $p = \frac{1}{3}$, $s \sim N(\omega, 1)$.

The blue shaded region represents (α, t) pairs where the Receiver would adhere to a coarsened binary signal at threshold t . Since the default action is $a = 0$, t must be sufficiently high to induce action. The red dotted line shows the Receiver's full-information boundary—above it, they choose to act. This intersects the Sender-optimal (black dotted) boundary at $\alpha = 1$, when the Receiver is unbiased. For $\alpha < 1$ (automation neglect), the Receiver underweights s , requiring a stronger signal to act, so their red line lies above the Sender's (and vice-versa under $\alpha > 1$, automation bias). The optimal threshold is given by the upper envelope of the solid blue line and the Sender-optimal black-dotted boundary, and always yields higher Sender utility than full revelation. Finally, the vertical black line at $\bar{\alpha}$ denotes the lowest α where the Receiver strictly adheres to the Sender's optimal boundary, allowing the Sender to fully implement their preferences. When $\alpha < \bar{\alpha}$, the Sender must raise the threshold above s^* to induce the Receiver to act.

2.3.3 Optimal binary coarsening with unknown r

We have previously shown that contingent on knowing r , the Sender can always construct a binary coarsening that improves outcomes. However, eliciting r and tailoring the threshold for every case might be impractical. Moreover, measuring r in practice may be noisy, distorted, or biased—even in absence of strategic considerations. What happens, then, if the Sender does not know r ?

We first consider the optimal binary coarsening $\Phi_{\{t\}}$. Recall that the Sender's subjective expected utility is given by $v(s, r) = E_{\pi}(U(\omega))$. The Sender then solves:

$$\max_{t \in \mathbb{R}} V(t) = \int_{-\infty}^t \int_{r_0(t)}^{\infty} v(s, r) g(r | s) dr f(s) ds + \int_t^{\infty} \int_{r_1(t)}^{\infty} v(s, r) g(r | s) dr f(s) ds$$

where the first term represents the Sender's utility from Receiver types who choose to approve the loan ($a = 1$) despite receiving a “No” ($\Phi_t(s) = 0$) message, while the second term captures utility from Receivers who approve after receiving a “Yes” ($\Phi_t(s) = 1$) message.¹⁴

Taking the first-order condition with respect to t , we obtain:

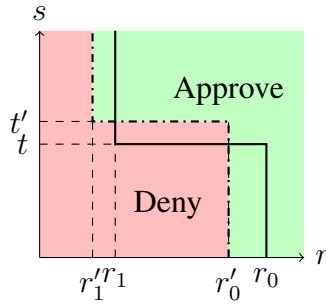
$$\left. \begin{aligned} & r'_0(t) \int_{-\infty}^t v(s, r_0(t)) g(r_0(t) | s) f(s) ds \\ & + r'_1(t) \int_t^{\infty} v(s, r_1(t)) g(r_1(t) | s) f(s) ds \end{aligned} \right\} \text{Adherence selection}$$

$$+ \int_{r_1(t)}^{r_0(t)} v(t, r) g(r | t) dr f(t) = 0 \left. \right\} \text{Recommendation}$$

¹⁴Graphically, these regions are captured in Figure 1a. The former is denoted by the region to right of $r_0(t_1)$ and below threshold t_1 ; the latter is the area right of $r_1(t_1)$, and above threshold t_1 .

Intuitively, this optimality condition shows two simultaneous effects when the threshold t changes, as illustrated in Figure 3. The first is an *adherence selection* effect: as the threshold changes from t to t' , the informational content of the binary signal changes and therefore the set of adhering types $[r_1(t), r_0(t)]$ shifts to $[r_1(t'), r_0(t')]$, captured by the horizontal shift from the solid to the dashed vertical lines. The second is a *recommendation* effect: as the threshold changes, the recommendations—and hence, choices—change, captured by the vertical shift from the solid to the dashed horizontal line. In other words, shifting the thresholds affects both *who* follows the recommendation, and *when* an action is recommended.

Figure 3: Binary (2-coarsening) when t increases to t'



Let \hat{t} denote the *AI-only boundary*, defined by condition $E_r[v(r, \hat{t})] = 0$. \hat{t} is the Bayesian-optimal inverse boundary absent any information on r , i.e., the threshold for s above which the Sender would choose $a = 1$, if they knew only s . The following proposition shows that under the certain conditions described below, the AI-only boundary is approximately the optimal threshold.

Proposition 3. Suppose $V(t)$ is uniformly concave,¹⁵ twice continuously differentiable and that there exists ϵ such that for $t = \hat{t}$,

1. $|r'_j(t)| \times g(r_j(t)) \leq \epsilon, j \in \{0, 1\}$ (slow-moving adherence OR wide adherence).
2. $|E_r[v(r, t) \mid r \in [r_1(t), r_0(t)]] - E_r[v(r, t)]| \leq \epsilon$ (symmetric adherence).

Then there exists a constant $C > 0$ such that t^* satisfies

$$|t^* - \hat{t}| \leq C\epsilon.$$

Consequently, as $\epsilon \rightarrow 0$, t^* converges to the AI-only boundary \hat{t} .

¹⁵Concavity of $V(t)$ is assumed to ensure an internal solution for t (a corner solution would imply that full censorship is optimal).

Proof in [Appendix](#). Proposition 3 shows that the Sender-optimal threshold equals the AI-only boundary, plus a deviation term whose size depends on two key factors, corresponding to the two effects discussed above: first, it depends on how sensitive Receiver adherence is to changes in the threshold, weighted by the density around the margin. If shifting the threshold has little effect on the composition of adhering types—either because marginal Receiver types are insensitive, or because the density of these types is small—the returns to deviating from \hat{t} to increase adherence are small. Second, it depends on how representative the adhering Receiver types under $\Phi_{\hat{t}}$ are compared to the overall population. Intuitively, the quality of the recommendation depends on how similar the adhering set is to the overall population; when they are similar, recommending the Receiver to act when $s > \hat{t}$ is close to optimal.

Both conditions of Proposition 3 are likely to hold in many real-world settings, making \hat{t} a good approximation for the optimal binary coarsening. First, in practice, adherence is unlikely to be sensitive to small perturbations in the threshold (e.g., Receivers are unlikely to be sensitive to a 0.50 vs. 0.52 threshold). Second, non-adhering Receivers likely come from the tails of the distribution—those with sufficiently strong human signals to overrule the binary message. If the distribution of r is not heavily skewed, the selection effect is likely small: symmetrically excluding the extremes typically leaves the mean stable. The recommended action for the set of adhering types is therefore likely similar to that for the full population. Together, these suggest that the AI-only boundary is a good approximation for the optimal binary signal in many contexts.¹⁶

2.3.4 Optimal binary coarsening with unknown r , and no residual information over AI

Consider the case where the Receiver has *no residual information* over the AI—that is, the human signal r provides no additional information about the state beyond the AI signal s , i.e., $\pi(s, r) = \pi(s, r')$ for all r, r' . This is an important case to address in a world of big-data and multi-modal models: in many decision-making contexts, 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.

In the absence of residual human information, the subjective expected utility function simplifies to $v(s, r) = v(s)$, which in turn simplifies the problem in two ways: (i) the inequality in the second condition of Proposition 3 holds with equality at $\epsilon = 0$, and (ii) the AI-only boundary is the threshold \hat{t} such that $v(\hat{t}) = 0$.

¹⁶Our simulations support this: across a wide range of parameters and behavioural biases, the optimal binary threshold typically lies at or near the AI-only boundary.

Therefore, as long as the first condition of Proposition 3 holds—which is likely in many settings for the reasons described above—the AI-only boundary is characterised by \hat{t} satisfying $v(\hat{t}) = 0$. Intuitively, if the Receiver’s human signal has no informational content conditional on the AI signal, then Bayes-optimal decisions should rely entirely on the AI. In this case, whenever the AI signal suggests a sufficiently high likelihood of $\omega = 1$, the Sender wants to shift the Receiver’s $a = 0$ (“Deny”) decision to $a = 1$ (“Approve”). If adherence doesn’t increase substantially when deviating from the AI-only boundary, it emerges naturally as the optimal decision rule.

Example: Correlation Neglect We illustrate the no residual information case with a simple example in which r and s are drawn from a highly correlated bivariate normal distribution conditional on the state—reflecting common settings where AI and human experts have substantial informational overlap. The Receiver is *correlation neglectful*: although the conditional correlation is $\rho = 0.8$, the Receiver perceives $\tilde{\rho} = 0$, treating the signals as conditionally independent.

Figure 4a visualises this example.¹⁷ The shading represents the joint signal distribution, and the grid visualises key benchmarks. Figure 4b shows the human-only case: the Receiver approves ($a = 1$) when $r > 0.5$, and denies ($a = 0$) otherwise. The lines show (inverse) decision boundaries in (r, s) space: the blue line (“biased”) shows the correlation neglectful Receiver’s boundary $b_{\tilde{\pi}}$, while the orange line (“Bayes”) shows the Bayesian boundary b_{π} . For convenience, the axes are rescaled to reflect posterior probabilities.¹⁸

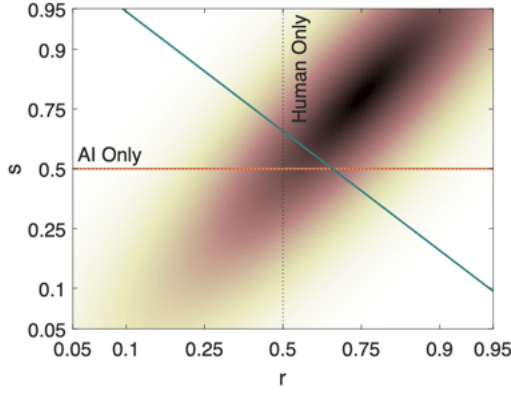
Since the Receiver has no residual information, the Bayes decision boundary is horizontal, approving whenever the AI signal implies a posterior above 0.5. In contrast, the biased boundary is steeper: the Receiver mistakenly treats r and s as independent, effectively double-counting their own information and placing more weight on it than a Bayesian would. The gap between Bayesian and biased boundaries, then, is where there is potential for improvement over full revelation. Figure 4c highlights in blue the areas where the biased Receiver disagrees with the Bayesian—denying when a Bayesian would approve (left) or approving when they would deny (right).

The treatment effect of the binary coarsening depends on the *adherence bounds*, $r_1(t)$ and $r_0(t)$, shown as the dotted curves in Figure 4d. For any binary threshold t , a Receiver with $r \in [r_1(t), r_0(t)]$ adheres to the message, and takes their default action otherwise. Figure 4e illustrates the coarsened policy at the AI-only boundary, $\hat{t} = 0.5$: a Receiver with r within the adherence

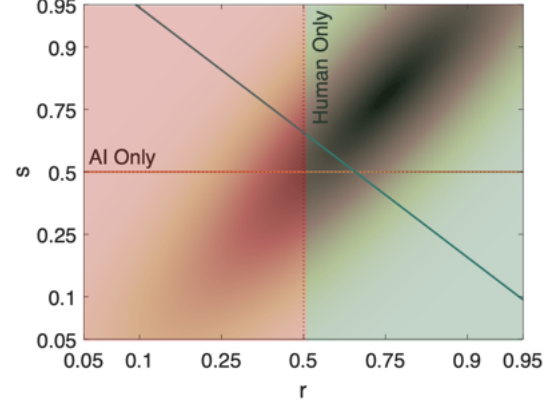
¹⁷It is worth noting that Figures 4a-4f, though simulated using an example of correlation neglect, can also be produced by a range of other biases, e.g., automation neglect.

¹⁸For instance, $r = 0.5$ means the human signal alone would imply a 0.5 posterior probability.

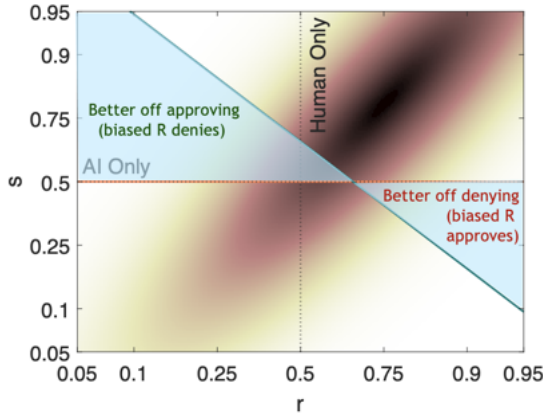
Figure 4: Decision boundaries, under no residual information and correlation neglect



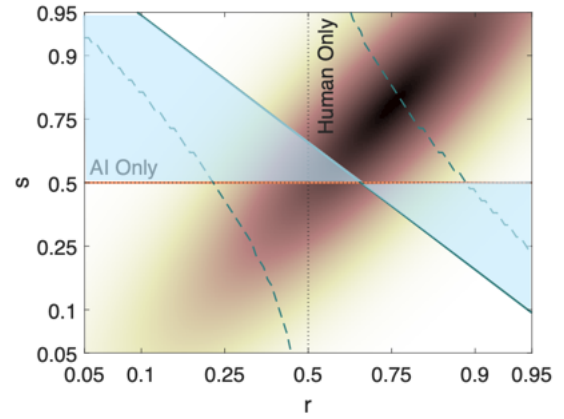
(a) Decision boundaries



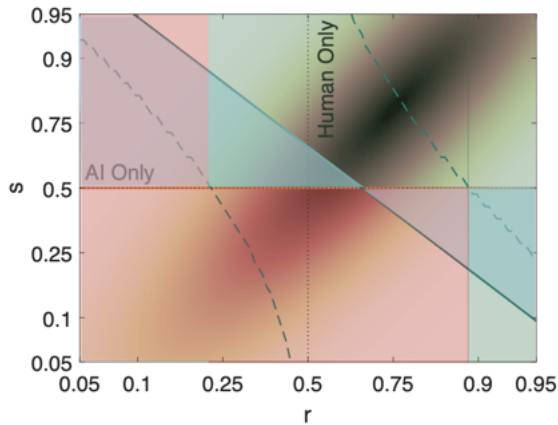
(b) Human-only Benchmark



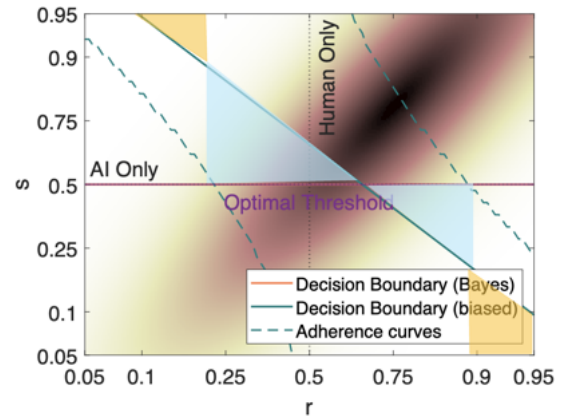
(c) Scope for Outcome Improvement



(d) Adherence Bounds (dotted)



(e) Coarsening at Decision Threshold ($s = 0.5$)



(f) Net Treatment Effect of Coarsening at Decision Threshold

Note: Biased and Bayesian boundaries in the r, s plane, with axes rescaled to posterior probabilities. The Bayesian boundary (which coincides with the AI-only boundary) is illustrated in orange (black dashed). The biased boundary is illustrated in blue. The vertical black dashed line represents the human-only boundary. The joint density is illustrated in the shading. The parameters used for the illustration are: $p = \frac{2}{3}$, $s \sim N(\mu_\omega, 1)$, $s \sim N(\mu_\omega, 1.25)$, $\mu_1 = 1$, $\mu_0 = -1$, $\rho = 0.8$. Receiver's belief $\tilde{\pi}$ is constructed under a misperceived correlation $\tilde{\rho} = 0$.

bounds approves whenever $s > 0.5$, always denies when r is below the leftmost adherence bound ($r < r_1(0.5)$), and always approves when r is above the rightmost adherence bound ($r \geq r_0(0.5)$).

We can see in the illustrations that the conditions in Proposition 3 are met in this example. First, the adherence bounds at $t = 0.5$ lie outside the central mass of the joint distribution of s and r . This means that a marginal change in t only affects the tails and has a negligible impact on adherence, satisfying the first part of the proposition. Second, because the Receiver has no residual information—as reflected in the flat Bayesian boundary at $s = 0.5$ —selection into adherence does not affect the AI-only recommended action, fulfilling the second condition in Proposition 3.

The blue areas in Figure 4f indicate gains from coarsening at the AI-only boundary (relative to full revelation); the yellow areas indicate losses. The *net effect* of coarsening at the optimal binary threshold \hat{t} , then, depends on the size of these regions, weighted by the density of the joint distribution of s and r . Whether binary coarsening is indeed the optimal disclosure policy is therefore an empirical question.

As seen in the following sections, we empirically test this in an experiment with loan specialists. We compare decision-making accuracy under binary coarsening at $t = \hat{t}$ (*medium*) against the following benchmarks: full revelation (full access to the AI risk score), human-only (decisions made without any AI input), and two other binary coarsenings—*high* ($t \gg \hat{t}$) and *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 multi-national financial institution Home Credit, which includes a wide range of relevant 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. One advantage of this particular context is that ground truth

exists: the dataset includes labels for whether or not an applicant repaid their loans on time. Ground truth for the direct outcome of interest is often unavailable in many other contexts; for example, clinical outcomes in radiology are hard to come by, and are instead usually defined by averaging the predictions of “top experts” (Agarwal et al., 2023). Here, our diagnostic standard exactly reflects the state(s) of the world 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 available only for a subset of the population. Specifically, ground truth is observed only for applicants approved for loans in the original Home Credit setting. Our AI model is trained on this subset (with repayment as the outcome). However, this is not a significant concern in our experiment, as participants evaluate a random sample from the Home Credit dataset with an application distribution that matches that of the AI training data,¹⁹ mitigating traditional selective label concerns. Instead, the primary consideration is of 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. While possible, this is unlikely. Moreover, firms deploying calibrated coarsening typically will 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 in-

¹⁹Of course, these cases are excluded from the training of the AI model.

tegrating 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 triary 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 that 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 and verify 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.²⁰

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) significantly constrain the full automation of consumer loan approvals.²¹ These structures create

²⁰This 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.

²¹Increased adoption of advanced computational methods—including AI/ML—has attracted much regulatory atten-

strong incentives for lenders to preserve human involvement in underwriting. While firms could, in theory, assume liability while still delegating decisions entirely to AI, doing so exposes them to greater legal risk. Black-box algorithms make it difficult to explain, defend, or audit credit decisions, particularly when challenged by regulators or plaintiffs. Without the ability to produce a transparent, human-defensible rationale, institutions face higher barriers to defending against claims of discrimination or unfairness. Thus, in practice, maintaining human oversight remains the prevailing—and risk-minimising—design.

For instance, Wells Fargo’s Enhanced Credit Score (ECS) algorithm has faced continued scrutiny through a consolidated class-action lawsuit alleging ECOA and FHA violations (*Williams et al. v. Wells Fargo Bank, N.A., Case No. 3:22-cv-00990-JD (N.D. Cal.)*). In fact, Wells Fargo defends ECS as an internal workflow tool that sorts applicants based on credit risk, assigning higher-risk applicants to more experienced specialists, thus serving a complementary rather than automated decision-making role (Bloomberg, 2024). Continued issues of legal liability strongly suggest that this practice—using algorithmic inputs predominantly in a human-in-the-loop manner—will continue, especially in a consumer loans context,²² preserving the central role of human underwriters.

4 Experiment I: Design

In our first experiment, we recruit 149 loan specialists to make hypothetical loan approval decisions, 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

tion. 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).

²²Non-consumer loans like small business loans, on the other hand, receive oversight primarily from the Small Business Administration (for SBA-backed loans) or the Federal Trade Commission (FTC), and are subject to less protection. Even so, The FDIC’s Small Business Lending Survey shows that while automation and financial technology (FinTech) are increasingly present in small business lending, they primarily serve to augment rather than replace human decision-making. Approximately 31% of banks reported using FinTech in at least one step of the small business loan process, with an additional 22% considering its adoption (FDIC, 2024).

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. A subset of the loan specialists were recruited through custom screening on Prolific, with stringent screening processes implemented to filter out scammers.²³ 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 the initial 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 8.5.1 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 and NAMP contacted via their LinkedIn groups. All participants—regardless of recruitment channel—have experience as loan specialists, with an average of 12 years in the field. Additionally, most reported prior exposure to artificial intelligence, though only a fifth had experience using supervised AI models, such as those commonly employed in loan approval predictions.

Once through with screening, qualified participants were directed to complete an online survey (taking 60 minutes on average) hosted on the Otree platform via Heroku, completing 60 loan application decisions, as well as answering additional questions (e.g., demand for AI signals, biases, demographics, amongst others) detailed in Appendix 8.3.3.

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

²³Prolific 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.

Table 1: Participant Demographics

	Sample Mean or Percentage
Age	44
Female	58.39%
Bachelor’s Degree or Above	61.07%
Income	107836
Loan Specialist Experience	100.00%
Loan Specialist Years	12
Bank / Credit Union Experience	34.23%
Applied to Loans Before	34.90%
Loan Amount	193615
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%

this information, they are directed to a landing page where they respond to two questions:

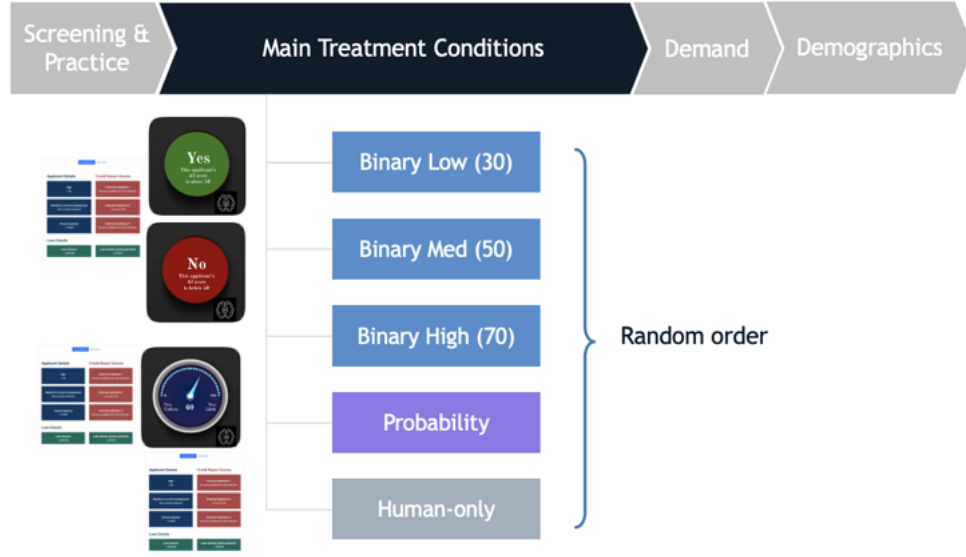
- “What do you think is the % chance that the applicant will make their future payments on time? [Click on the blue bar to choose a number between 0 and 100.]”
- “Do you want to approve this applicant for a loan? [Approve / Deny]”

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 [8.3.1](#).

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

Figure 5: Experiment I survey flow



approved (for on-time payers). Equalising the costs of false positives and negatives is equivalent to wanting to maximise accuracy, so for the rest of this section we focus mainly on diagnostic accuracy as our main result. Elicited posteriors are additionally incentivized via a binarized scoring rule (Hossain and Okui, 2013). Participants are expected to earn \$3-4 in bonuses on average, in addition to a base rate which varied from \$16 (Prolific) to \$40 per survey (Indeed, LinkedIn).

4.2.1 Loan application details

Participants are shown the following variables as part of a loan application (see Figure 16 in the Appendix for an example). These key variables were identified through expert interviews with loan specialists and results from pilot studies, with a focus on the debt-to-income ratio, months in employment (relative to age), as well as credit scores from external sources. These are the primary factors that loan specialists use to assess and make decisions on applications. Variables include: loan amount, loan annuity, age, months in current employment, income (yearly), and normalized credit scores from three sources, each ranging from 0–100, with higher values indicating greater creditworthiness.

4.2.2 AI model details

We train one of the top-performing AI models from a public competition on the Home Credit dataset of 300,000 loan applicants and their repayment outcomes, as described in 3. The AI model

has access to all details in the loan application made available to human participants, as is explicitly made clear in the experimental instructions (see Appendix Figure 12).

The model employs Bayesian optimisation to automate the tuning of hyper-parameters, resulting in a cross-validation AUC-ROC score of 0.79. 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 information 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 it thinks 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],” 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 specific loan applications within each block are also randomised, and thus typically differ across treatment conditions for different individuals. As a robustness check, we report block-level results in Appendix 8.5.3 and show that *Binary Medium* (50) consistently outperforms *Probability* across all five blocks.

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 “human signal” in cases where: (i) the posterior is elicited both before and after the second signal is shown, 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 bias our estimates of how coarsening affects belief updating and decision accuracy. Moreover, since we later use elicited posteriors to visualise how people update from their own signal vs. the AI’s, it is important to know how sensitive these posteriors are to the order in which information is presented. In Appendix 8.5.2, 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 mainly on loan approval accuracy as our main result, given that we incentivise for accuracy. Our main specification is as follows:

$$Y_{ilt} = \beta_0 + \sum_t \beta_t \text{Treatment}_{il}^t + \epsilon_{ilt} \quad (1)$$

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 (we set the *Binary Medium* treatment as the reference base level). We estimate the regression with and without fixed effects (e.g., case number fixed effects). Identification is established by the random assignment of the treatment.

Our main objects of interest are the β_t coefficients. In particular, we test the following 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* or *Probability* treatment.

5 Empirical evidence I: Universal calibrated coarsening improves outcomes

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

5.1 Main results

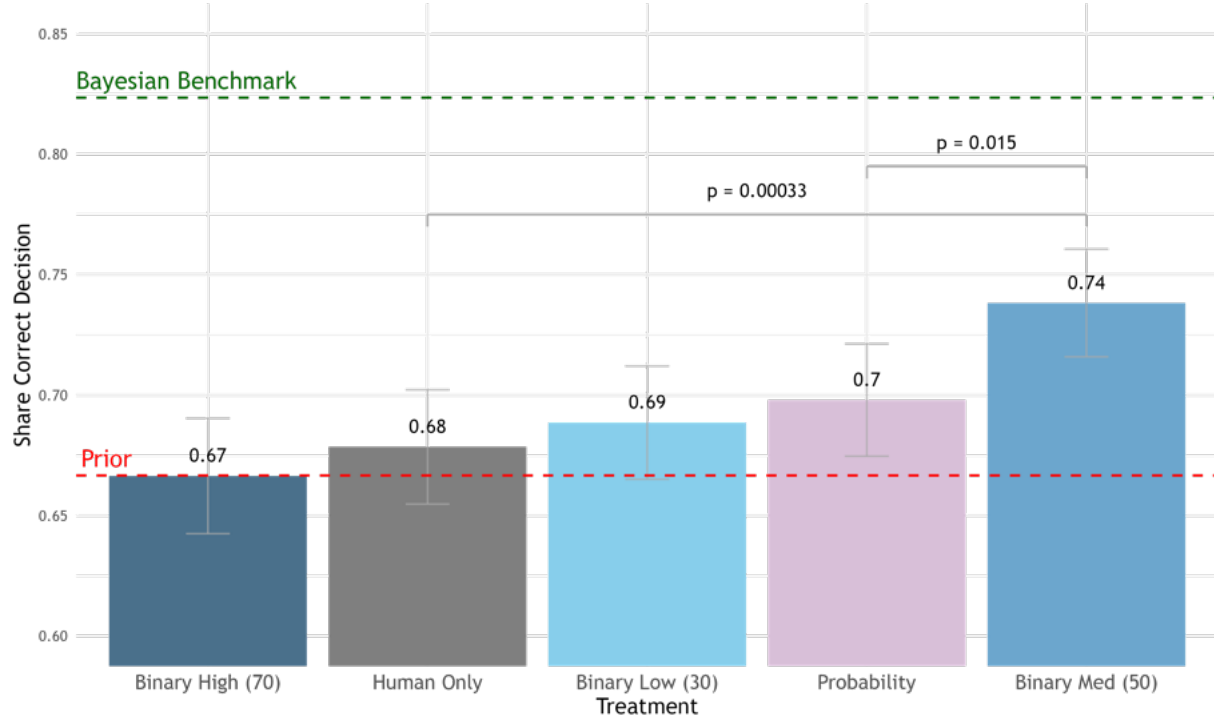
5.1.1 Impact of coarsening on decision-making outcomes

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 1, using the *Binary Medium* (50) treatment as the reference level. Loan specialists in this condition made, on average, 4 percentage points more correct decisions than those in the *Probability* condition (Column 1). This effect remains significant even after controlling for loan application (case-level) 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).²⁴ Moreover, while providing participants with the full AI score in the *Probability* condition

²⁴Note 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.

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.

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 more than half the gap between the *Human-only* condition and the Bayesian benchmark.

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. This underscores the importance of *calibrated* coarsening: by selecting an optimised threshold, as in the *Binary Medium* (50) condition, decision-making performance improves substantially, bringing participants significantly closer to the Bayesian benchmark.

Table 2: Decision-making outcomes under different treatment conditions

	<i>Dependent variable: Share of Decisions Correct</i>	
	(1) No Fixed Effects	(2) Case-Level Fixed Effects
Human-Only	−0.060*** (0.017)	−0.057*** (0.014)
Probability	−0.040** (0.017)	−0.029** (0.014)
Binary High (70)	−0.072*** (0.017)	−0.058*** (0.014)
Binary Low (30)	−0.050*** (0.017)	−0.040*** (0.014)
Constant	0.738*** (0.012)	0.699*** (0.034)
Observations	7,450	7,450
R ²	0.003	0.282
Adjusted R ²	0.002	0.277

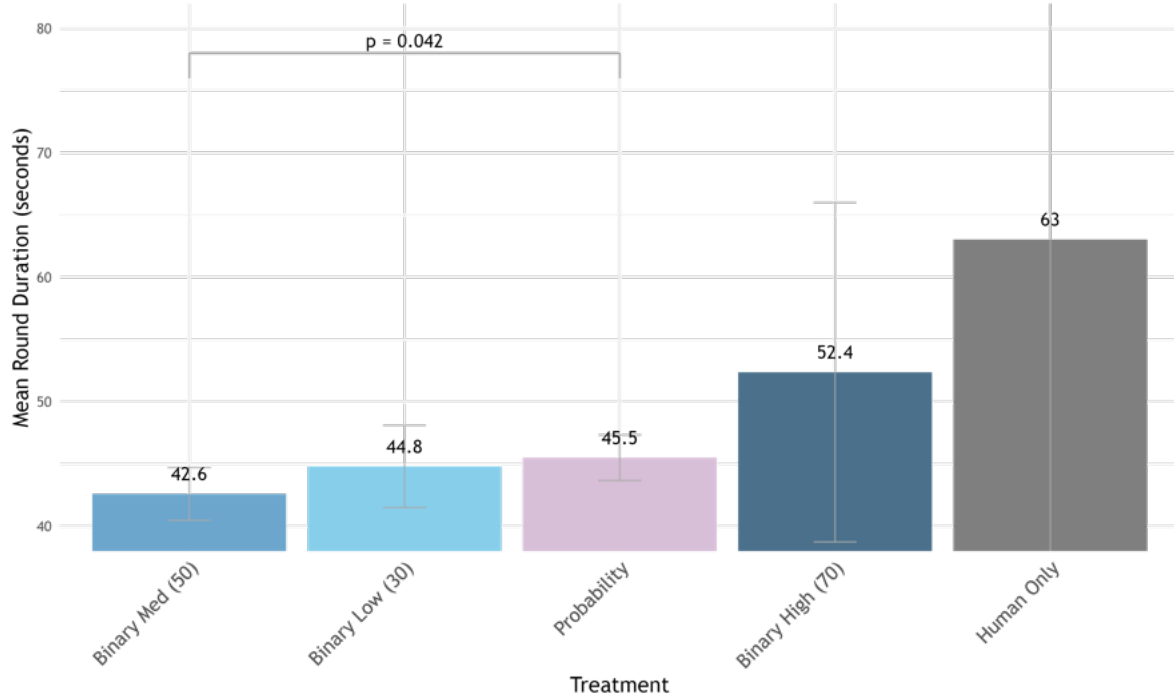
Note: Regression estimates per Equation 1 of loan approval accuracy by treatment condition, with *Binary Medium (50)* as the omitted reference group. Column (1) reports estimates without fixed effects; Column (2) includes application (case-level) fixed effects. Standard errors in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$

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.

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

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.

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 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.

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

Choice of AI	<i>Highest Payoff Treatment</i>				Mean WTP Binary (cents/case)	Mean WTP Prob (cents/case)
	Total (#)	Human Only (% total)	Binary (% total)	Prob (% total)		
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 effective for them. The majority of participants for whom the highest payoff treatment was *Human-only* chose not to use any AI,²⁵ 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 4, 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 22 in the Appendix shows the distribution of these thresholds. Choices are skewed to the right of 50: many partici-

²⁵This may be mechanically correlated, of course, 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 *Human-only* treatment.

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 clearly good applicants (those 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.

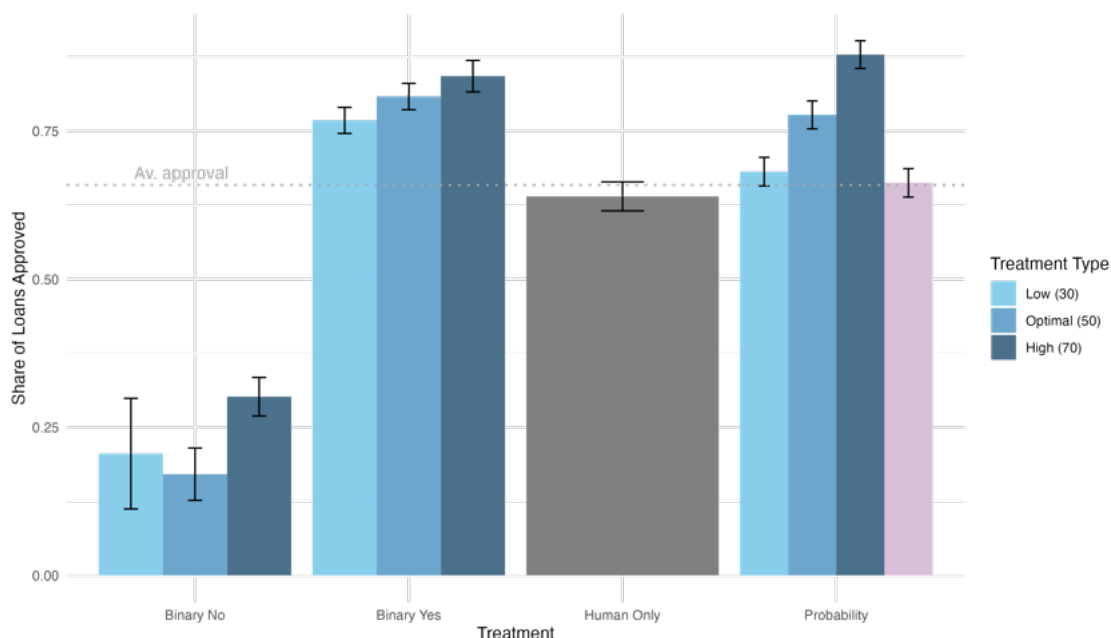
One important aspect we verify is whether loan specialists behave consistently with the decision rule implied by the incentive structure. Since correct approvals and denials are rewarded equally, participants should approve when their posterior exceeds 50% and deny otherwise. We conduct this check to gain insight into whether deviations from the Bayesian benchmark are happening at the level of belief-updating or inconsistent translation of beliefs into actions. Appendix Figure 23 shows that for most participants, approve/deny decisions are cleanly separated around the 50% posterior threshold (“action threshold”)—suggesting that most deviations stem from belief formation rather than decision noise (conditional on posterior) or mis-calibrated action thresholds.

5.2.1 Algorithmic adherence

Having established that participants generally apply consistent action thresholds, we begin by taking a deeper dive into how participants react to AI signals. Figure 8 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. But 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. Conversely, a lower threshold should make a “No” message weaker, leading to fewer denials. Figure 8 confirms this expectation: as the threshold of the binary AI signal increases, loan specialists approve more loans when the AI recommends “Yes” and deny fewer when it recommends “No”. This trend is 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.

Figure 8: Loan Approval Rates, by Treatment and (implied) AI Recommendation



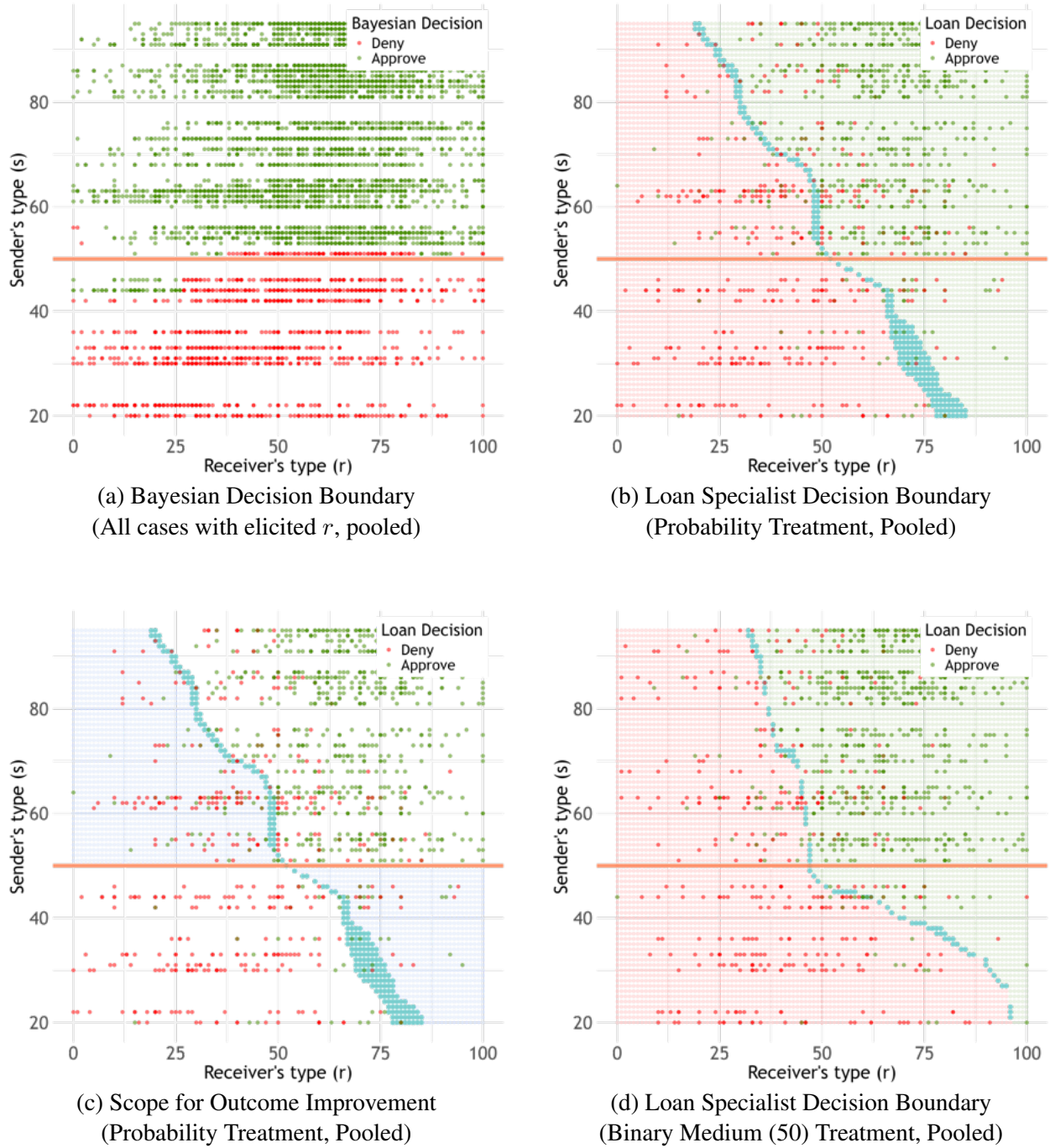
Note: Bars show loan approval rates by treatment. *Binary* conditions are split by whether the AI recommends “Yes” or “No” at thresholds of 30%, 50%, or 70%. *Probability* is split by whether the AI score exceeds each threshold; the purple bar shows the overall average. Error bars indicate 95% confidence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

These findings confirm that participants respond to AI-calibrated thresholds rather than treating all AI recommendations equally. 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 do not fall at 0% and 100%, respectively, shows that loan specialists do not blindly follow AI recommendations. This imperfect adherence suggests that participants are exercising their own judgment—albeit erroneously in this case, where they don’t have any residual information over the AI—rather than following the AI’s suggestions.

5.2.2 Estimating pooled decision boundaries

If participants are not simply following the AI signal, how are they actually making decisions? To investigate this, Figures 9a-9d present the empirical analogues of the decision boundary graphs in Figures 4a-4f in the model section, estimated using our experimental data.

Figure 9: Pooled Decision boundaries, Visualised Using Experiment Data



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.

First, we verify the Bayesian decision boundary—the best a decision-maker can do given their information—represented by the orange line in Figure 9a. For all cases where we have elicited the human signal over the loan application (r),²⁶ 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. In the figure, green and red dots represent cases where the Bayesian would approve and deny, respectively, and the orange line marks the decision boundary between them. As expected in this setting—where humans have no residual information over the AI—the Bayesian decision boundary is a flat line around $s = 50$. This reflects that a Bayesian decision-maker would perfectly follow the AI signal, approving any case where the AI score exceeds 50 and denying otherwise, serving as a benchmark to evaluate the extent to which participant decisions deviate from this ideal behavior.

Next, we estimate the decision boundary of the loan specialists in our experiment. To do this, 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 9b.

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 9c 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 behavior.

Examining decisions under the *Binary Medium (50)* treatment reveals that coarsening the AI signal can indeed improve decision-making outcomes. Figure 9d illustrates how thresholding the

²⁶Since 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, this results in our observing r for 81% of cases.

AI signal at the model-implied optimised threshold ($s = 50$) helps participants adhere more closely to the AI recommendation within the range of $r = 40$ to $r = 90$. 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.

5.2.3 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, consistent with institutional norms that emphasize approving “cherries” over rejecting “lemons.” 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 25–27.

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), presented with case fixed effects in Appendix 8.4.2.

5.3.1 HTEs by demographics

Gender emerges as a key moderator, with men benefiting significantly more from AI assistance. The *Binary Medium* (50) treatment increases their accuracy by 9.5 percentage points over *Human-only*, compared to just 2.6 percentage points for women. Similarly, men improve by 4.8 percentage points when moving from the *Probability* condition to *Binary Medium* (50), whereas the corresponding gain for women is only 1.4 percentage points. Notably, these differences are not

driven by baseline disparities, as both genders exhibit similar accuracy levels (64–65%) in the *Human-only* condition. Instead, the pronounced gender gap in AI-driven performance gains 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).

Other demographic factors also matter. While age does not meaningfully moderate treatment effects, younger loan specialists—those below the median age—perform better across all treatments, with a 3.9 percentage point advantage in the *Human-only* condition. Experience, on the other hand, plays a more pronounced role. Loan specialists with fewer years of experience perform significantly worse at baseline, with an 8 percentage point accuracy gap in the *Human-only* condition between specialists with low and high years of experience. However, they experience higher treatment effects, closing about half of this initial gap, suggesting that those with less prior domain knowledge are more willing to listen to AI’s recommendations, benefitting disproportionately from AI assistance, at least in this context—a finding supported in the literature (Dell’Acqua et al., 2023). While overall differences in treatment effects by practice accuracy (proxy for baseline performance) are modest, there is some variation: those with below-median practice accuracy benefit slightly more from coarsened signals.

Perceived private information also influences responsiveness to AI. Those with below-median perceived private information had lower baseline accuracy in the *Human-only* condition (3 percentage points lower). However, they benefited more from AI assistance: the *Binary Medium* (50) treatment improved their accuracy by 7.4 percentage points over *Human-only*, while the *Probability* treatment led to a 3.8 percentage point increase (in contrast to 4.2 and 2.0pp respectively for those above the median). 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, regardless of its presentation format.

5.3.2 HTEs by cognitive biases

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 human-AI 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 **automation bias**, baseline accuracy in the *Human-only* condition is similar across groups, but the treatment effects are higher for those who face less automation bias (that is, more automation neglect). In this setting, where there is no residual human information over the AI, there is no scope for automation bias. Instead, the relevant concern is automation neglect, as specialists may rely too heavily on their own information rather than following the AI’s signal, which is the best approach in this context. Our finding that those facing more automation neglect benefit more from coarsening supports this idea, reflecting insights from the simulated example in Section 2. When the loan specialist underweights the AI signal, coarsening can help pool signals together and encourage the specialist to listen more to the AI.

The **balls-urns** measure shows a stark contrast: those with an above-median parameter—who exhibit close-to-Bayesian updating or conservatism bias—have much higher baseline accuracy in the *Human-only* condition (12.8 percentage points 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. **Base-rate neglect** also appears to play a role: those who face above-median base-rate neglect start with lower accuracy in the *Human-only* condition and experience a correspondingly larger treatment effect, which makes sense given that individuals with higher base-rate neglect are discounting priors more, and thus face greater updating bias. By contrast, the **aggregation bias** and **correlation neglect** measures do not moderate either baseline accuracy or treatment effects.

Together, these findings provide suggestive evidence of the sorts of biases (e.g., automation neglect, base-rate neglect) that may be driving suboptimal use of AI assistance, and how calibrated coarsening can help correct these biases.

6 Estimating decision boundaries to personalise policies

The previous section documented substantial heterogeneity in how specialists respond to different types of AI assistance—across both demographic groups and cognitive biases—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 designing AI-assistance policies tailored to each specialist.

In brief, the method specifies a prior distribution for the model parameters and a model of the likelihood function, and applies Bayes' rule to derive the posterior distribution of the parameters, reflecting updated beliefs after seeing the data. 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.

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 $T = \{t_1, \dots, t_{K-1}\}$ 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 $\mu_M(t) = \frac{E[s_n | t_1 \leq s_n < t_2] - E[s_n | s_n < t_1]}{E[s_n | s_n \geq t_2] - E[s_n | s_n < t_1]}$, so that all three support points lie on $[0, 1]$.²⁷ 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

²⁷This ensures that a jump from $0 \rightarrow \mu$ has the same marginal effect on latent approval as from $\mu \rightarrow 1$.

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 $a_{i,T,n} = \mathbb{1}[y_{i,T,n}^* \geq 0]$. The variable $y_{i,T,n}^*$ is given by:

$$y_{i,T,n}^* = \alpha_{i,T} + r_{i,T,n} + \beta_{i,T} \cdot \Phi_T(s_n) + \epsilon_{i,T,n} \quad \epsilon_{i,T,n} \sim \mathcal{N}(0, \sigma_{\epsilon,T}^2)$$

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

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 where data is sparse. 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 $\sigma_{\alpha,T}^2$ and $\sigma_{s,T}^2$ represent residual heterogeneity in intercepts and slopes, respectively. This structure allows specialists with similar characteristics to have similar coefficients while still permitting idiosyncratic variation. We assume that both $\alpha_{i,T}, \beta_{i,T}$ are independent of each other and of r ,²⁸ and normalise the coefficient on the human signal to fix scale.²⁹ The term $\epsilon_{i,T,n}$ captures residual idiosyncratic variation in decision-making not explained by the observed signals, where $\sigma_{\epsilon,T}^2$ 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.

²⁸The 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.

²⁹To 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

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

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

6.2 Decision model estimation

As we wish to retain our flexible conceptual framework and allow people to follow different updating procedures depending on the signal presentation without further structural constraint,³⁰ 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 first experiment, this includes $T \in \{30\}, \{50\}, \{70\}, \emptyset$, and $S \setminus \{\min(S)\}$. The econometrician observes $D = \{a_{i,T,n}, \mathbb{1}_{i,T,n}^{\text{obs}}, \hat{r}_{i,T,n}, s_n, \Phi_T(s_n)\}$, where $\hat{r}_{i,T,n}$ is the loan specialists' human signal *when elicited* and is only observed when $\mathbb{1}_{i,T,n}^{\text{obs}}$ 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:

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

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 $\sigma_{r,T}^2$, 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 $\Theta = \{\alpha_{i,T}, \beta_{i,T}, \lambda, \delta, \sigma_{\epsilon,T}^2, \sigma_{r,T}^2\}$:

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

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 ($\mathbb{1}_{i,T,n}^{\text{obs}} = 1$) we also explicitly model $\hat{r}_{i,T,n} \sim \mathcal{N}(Z_i^\top \delta + X_n^\top \lambda, \sigma_{r,T}^2)$ as part of

³⁰i.e., π_Φ is indexed by Φ as per our conceptual framework.

the joint likelihood, via the *latent signal model*:

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

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

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

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 the human signal \hat{r} is reported, we regress \hat{r} on the specialist covariates Z_i and loan covariates X_n ; the resulting point estimates $(\hat{\delta}, \hat{\lambda})$ and squared standard errors $(\hat{\sigma}_{\delta}^2, \hat{\sigma}_{\lambda}^2)$ become the means and variances of our priors: $\delta \sim \mathcal{N}(\hat{\delta}, \hat{\sigma}_{\delta}^2)$ and $\lambda \sim \mathcal{N}(\hat{\lambda}, \hat{\sigma}_{\lambda}^2)$. The residual signal variance is given a truncated normal prior $\sigma_r \sim \mathcal{N}(\hat{\sigma}_r, \hat{\sigma}_{r,se}^2) \mathbb{I}_{(0,\infty)}$.

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 latent decision boundary is defined as the set of points (r, s) at which their latent utility $y_i^* = 0$. We can express this 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: $r_1 = -\alpha_{i,T} - \beta_{i,T}, r_0 = -\alpha_{i,T}$ 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.

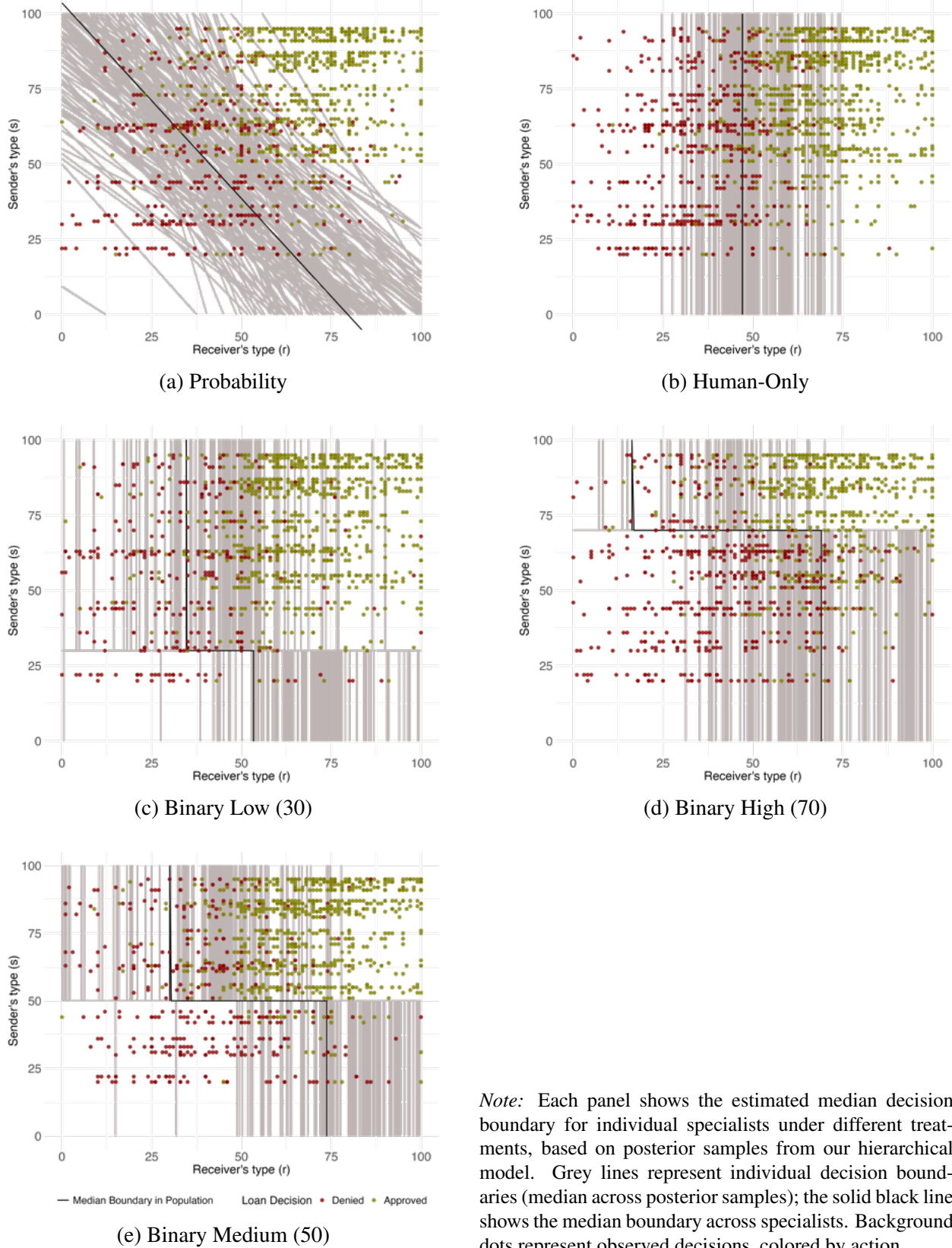
Trinary Treatments: $r_2 = -\alpha_{i,T} - \beta_{i,T}, r_1 = -\alpha_{i,T} - \beta_{i,T} \cdot \mu_M(t), r_0 = -\alpha_{i,T}$ for $t \in \{30, 50\}, \{50, 70\}$, where the actions of the Receivers for each trinary treatment are characterised by $\{t_1, t_2, r_0, r_1, r_2\}$, where t_1, t_2 are the thresholds 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 Experiment I data

Using data from Experiment I, 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 adherence bound (e.g., r_0, r_1 , etc.). Each grey line in Figure 10 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 10a), we observe significant variation in slopes and intercepts. Specialists with flatter boundaries and intercepts around $s = 50$ behave

Figure 10: Heterogeneity in decision-boundaries under different treatments



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, colored by action.

closer to the Bayesian benchmark (fully flat); in contrast, steeper boundaries reflect greater weight on the human’s private signal or under-weighting 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 10e. The observed heterogeneity in treatment responses suggests potential gains from 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 $\hat{a}_{i,T,n}^{(s)} = \mathbb{I}[\alpha_{i,T}^{(s)} + \frac{1}{\sigma_{\epsilon,T}} \cdot r_{i,T,n}^{(s)} + \frac{\beta_{i,T}^{(s)}}{\sigma_{\epsilon,T}} \cdot \Phi_T(s_n) \geq 0]$. For each draw, we compute the predicted accuracy for each treatment by averaging over the evaluation set and posterior draws:

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

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 30% allocated to *Binary High* (70), 15% allocated to *Binary Low* (30), and 55% retaining the *Binary Medium* (50) policy from the generalised assignment. This personalised approach is predicted to yield a 2.5 percentage point improvement over the generalised policy, highlighting room for further improvement of decision-making outcomes by calibrating not just thresholds, but treatments to individual decision-makers.

7 Conclusion and Discussion

We show in this paper—both theoretically and empirically—that coarsening AI signals at optimised thresholds can significantly enhance human-AI collaboration outcomes, including in high-stakes fields such as loan approvals. We propose a framework for improving decision-making that keeps humans in the loop, adapts to various forms of human bias, and is applicable across diverse contexts. This approach addresses the challenge of human cognitive biases that often hinder the effective use of AI predictions in real-world deployments.

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), and empirically testing our personalised thresholds in a two-stage experiment. Additionally, we are exploring extensions to other domains, including hiring decisions, to assess the broader applicability of calibrated coarsening in different decision-making contexts.

In conclusion, calibrated coarsening is a practical and effective approach to improving human-AI decision-making. Future research avenues include scenarios involving strategic considerations, where the Receiver’s objective function may diverge from that of the Sender’s due to misaligned incentives or motivated biases, such as those against gender or race in hiring or loan approvals. Additionally, developing dynamic, adaptive systems that adjust information presentation in real-time based on decision-maker behavior and context may hold significant promise. Moreover, in contexts where humans may not need to have final decision rights but may have useful residual information over the AI, further developing methods to integrate human judgment into automated AI-human decisions could prove an interesting direction. As human-AI collaboration evolves, we believe advancements in 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.
- Alberdi, E., Strigini, L., Povyakalo, A. A., and Ayton, P. (2009). Why are people’s decisions sometimes worse with computer support? In *Computer Safety, Reliability, and Security: 28th International Conference, SAFECOMP 2009, Hamburg, Germany, September 15-18, 2009. Proceedings 28*, pages 18–31. Springer.
- 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.
- 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.
- 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 Fuller, 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.

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. 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-the-Underwriting-Process>. 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.
- 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.
- 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, Maria, F. 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., Kraymer, 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., Lerner, 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.
- 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.

- Guo, Z., Wu, Y., Hartline, J., and Hullman, J. (2025). The value of information in human-AI decision-making. *arXiv preprint arXiv:2502.06152*.
- 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.
- Huellden, 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.
- Lai, V., Chen, C., Liao, Q. V., Smith-Renner, A., and Tan, C. (2021). Towards a science of human-AI decision 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.
- 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.

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.

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-engagement-in-real-time.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.

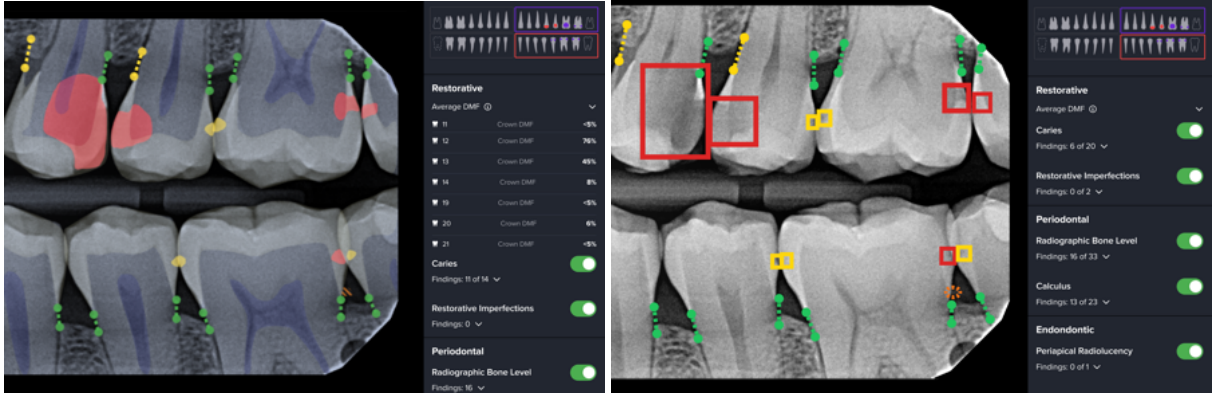
- Snyder, C., Keppler, S., and Leider, S. (2024). Algorithm reliance, fast and slow. *Fast and Slow* (May 31, 2024).
- Soll, J. B. and Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of experimental psychology: Learning, memory, and cognition*, 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.

8 Appendix

8.1 Real-world Examples of Coarsened AI signals

Figure 11: VideAI Dental Interface



8.2 Proofs

Proposition 1

Proof. Fix $u(r, s) = v(r, s)$ for all $r \in R$ and $s \in S$. The Sender's expected utility under Φ is given by:

$$\mathbb{E}_{\Phi}[v] = \mathbb{E}_{\Phi}[u] = \int_{R^* \times S} \left(\int_r^{\bar{r}} \tilde{u}(\tilde{r}, s) d\tilde{r} \right) d\Phi(r, s)$$

since a message k induces the Receiver r to act if and only if $r \geq r_k$. Under the *full revelation mechanism*, Φ_s generates the message $r^*(s)$ for each $s \in S$. Hence the expected utility of the Sender under Φ_s is:

$$\mathbb{E}_{\Phi_s}[v] = \mathbb{E}_{\Phi_s}[u] = \int_S \left(\int_{r^*(s)}^{\bar{r}} \tilde{u}(\tilde{r}, s) d\tilde{r} \right) f(s) ds = \int_{R^* \times S} \left(\int_{r^*(s)}^{\bar{r}} \tilde{u}(\tilde{r}, s) d\tilde{r} \right) d\Phi(r, s),$$

Recall that r^* satisfies $u(r^*(s), s) = 0$. Using Fubini's theorem and the condition $\tilde{u}(r^*(s), s) = 0$, we can express the difference in expected utilities as:

$$\mathbb{E}_{\Phi_s}[v] - \mathbb{E}_{\Phi}[v] = \int_S \int_{r > r^*(s)} \left(\int_{r^*(s)}^r \tilde{u}(\tilde{r}, s) d\tilde{r} \right) d\Phi(r, s) - \int_S \int_{r < r^*(s)} \left(\int_r^{r^*(s)} \tilde{u}(\tilde{r}, s) d\tilde{r} \right) d\Phi(r, s)$$

By Result 1, we know that $\tilde{u}(\tilde{r}, s) > 0$ for $\tilde{r} > r^*(s)$, which implies $\int_{r^*(s)}^r \tilde{u}(\tilde{r}, s) d\tilde{r} > 0$ for $r > r^*(s)$. Any mechanism Φ that differs from Φ_s assigns strictly positive probability to the event $r > r^*(s)$. Otherwise, the term $\int_{R^* \times S} \tilde{u}(r, s) d\Phi(r, s)$ would be strictly negative rather than zero. Therefore, the first integral is strictly positive; conversely, the second integral has to be strictly negative. Therefore, $\mathbb{E}_{\Phi_s}[u] - \mathbb{E}_{\Phi}[u] > 0$ for any Φ that differs from Φ_s , making Φ_s the optimal mechanism that maximises the Sender's expected utility.

□

Proposition 2

Proof. Fix r . Starting with the first part, assume without loss of generality that the Sender's full-revelation (inverse) boundary is higher, i.e., $\hat{b}_\pi(r) > \hat{b}_{\tilde{\pi}}(r)$. Let

$$t^* = \min \left\{ \{t : r_0(t) = r\}, \hat{b}_\pi(r) \right\}$$

The threshold t^* is either Sender's optimal decision boundary, or the highest binary threshold Sender can set such that the Receiver still adheres to the binary signal. We now prove that $t^* \in (\hat{b}_{\tilde{\pi}}(r), \hat{b}_\pi(r)]$.

First, observe that the Receiver will adhere to the binary signal when the Sender sets the threshold t at the Receiver's decision boundary, i.e., $t = \hat{b}_{\tilde{\pi}}(r)$. To see why, notice that by Part 1 of Assumption 2, the Receiver's posterior is a weighted average of posteriors. When $t = \hat{b}_{\tilde{\pi}}(r)$, the coarsened signal $\Phi(s) = 0$ pools together signals s below the boundary, i.e., $\tilde{\pi}(r, s) < \tilde{\pi}(r, \hat{b}_\pi)$. Therefore, any weighted average must satisfy $\tilde{\pi}_\Phi(r, 0) < \tilde{\pi}(r, \hat{b}_\pi)$, and it must be the case that the Receiver chooses $a = 0$. Similarly $\Phi(s) = 1$ pools together all signals above the boundary, which implies that the Receiver chooses $a = 1$. Moreover, by Part 2 of Assumption 2, there exists $\Phi_{t'}$ with t' close enough to the boundary $\hat{b}_{\tilde{\pi}}(r)$ such that the Receiver still adheres to $\Phi_{t'}$. Therefore, $t^* > \hat{b}_{\tilde{\pi}}(r)$.

Notice that this immediately implies that Φ_{t^*} improves Sender's utility over full revelation: their utility over the segment $s \in [\hat{b}_{\tilde{\pi}}(r), t^*]$ is strictly higher relative to full revelation, and it is identical everywhere else.

To see why Φ_{t^*} is optimal, assume by way of contradiction that there exists a monotone coarsening Φ that dominates Φ_{t^*} . In particular, Φ must induce $a = 0$ at some $s' > t^*$, but this immediately leads to a contraction, since by the definition of t^* , no binary coarsening Φ_t can induce $a = 0$

when $\Phi_t(s') = 0$, which then implies that no monotone coarsening can induce $a = 0$ at s' . The proof for the case $\hat{b}_\pi(r) < b_{\bar{\pi}}(r)$ is symmetric.

Moving on to the last part, first notice that the Sender fully implements their boundary iff Receiver follows $\Phi_{\hat{b}_\pi(r)}$. Assume $\hat{b}_\pi(r) > \hat{b}_{\bar{\pi}}(r)$, and that the default action is 0. Clearly, the Receiver chooses $a = 0$ whenever $\Phi_t(s) = 0$, since it induces a posterior strictly lower than the posterior induced by full censorship. Similarly, since $\hat{b}_\pi(r) > \hat{b}_{\bar{\pi}}(r)$, the Receiver learns that $s \geq \hat{b}_\pi(r) > \hat{b}_{\bar{\pi}}(r)$, and chooses $a = 1$. Therefore, Receiver adheres the signal. The proof for the other case is symmetric. □

Proposition 3

Proof. Recall that the FOC is given by

$$\begin{aligned} V'(t) = & r'_0(t) \int_{-\infty}^t v(s, r_0(t)) g(r_0(t) | s) f(s) ds \\ & + r'_1(t) \int_t^{\infty} v(s, r_1(t)) g(r_1(t) | s) f(s) ds \\ & + \int_{r_1(t)}^{r_0(t)} v(t, r) g(r | t) dr f(t) = 0 \end{aligned}$$

which we can write as $r'_0(t)A_0(t) + r'_1(t)A_1(t) + f(t)B(t) = 0$.

Notice that $|v(s, r)|$ is bounded by $\max\{c_{FP}, c_{TP}\} \equiv \bar{v}$. Therefore, at $t = \hat{t}$ we have:

$$|A_0(\hat{t})| \leq \bar{v} \int_{-\infty}^{\hat{t}} g(r_0(\hat{t}) | s) f(s) ds \leq \bar{v} g(r_0(\hat{t})),$$

which, by Condition 1 in the proposition, implies:

$$r'_0(t)A_0(t) \leq \bar{v}\epsilon \quad \text{and} \quad r'_1(t)A_1(t) \leq \bar{v}\epsilon.$$

For the third term, by Condition 2 and the fact that $E_r[v(\hat{t}, r)] = 0$ at the AI-only boundary, we

have:

$$\begin{aligned}
|B(\hat{t})| &= \left| \int_{r_1(\hat{t})}^{r_0(\hat{t})} v(\hat{t}, r) g(r | \hat{t}) dr \right| \\
&= |E_r[v(\hat{t}, r) | r \in [r_1(\hat{t}), r_0(\hat{t})]]| \cdot \int_{r_1(\hat{t})}^{r_0(\hat{t})} g(r | \hat{t}) dr \\
&\leq \epsilon
\end{aligned}$$

Therefore,

$$|V'(\hat{t})| \leq 2\bar{v}\epsilon + f(\hat{t})\epsilon \leq C_1\epsilon$$

where $C_1 = 2\bar{v} + \sup_s f(s)$.

By the Mean Value Theorem, there exists some \tilde{t} between \hat{t} and t^* such that:

$$|V'(\hat{t}) - V'(t^*)| = |V''(\tilde{t})| \cdot |t^* - \hat{t}|$$

Since $V'(t^*) = 0$, it follows that:

$$|t^* - \hat{t}| = \frac{|V'(\hat{t})|}{|V''(\tilde{t})|} \leq \frac{C_1}{m}\epsilon$$

where $m > 0$ satisfies $|V''(t)| \geq m$ (the existence of such m is guaranteed by uniform concavity).

Defining $C \equiv \frac{C_1}{m}$ yields the desired result.

□

8.3 Experiment Design

8.3.1 Experiment I Instructions

Figure 12: Instructions to participants to highlight prior and AI details

Instructions

[Instructions](#) [More details on AI](#)

For the rest of the questions in the survey, we will show you profiles of people who are each applying for a loan. Imagine you are a loan approval officer who has to decide whether or not to approve/deny the loan.

For each of these people, we will show you:

1. A **loan application** like the ones you saw in the practice rounds
2. Additional information from an **artificial intelligence (AI) assistant** on how likely it thinks the applicant will repay their loan on time.

Sometimes the AI information will be shown first, and other times the loan application data will be shown first. Based on this information, you need to decide whether or not to approve the loan for them.

All of the AI assistants come from the same AI model that is trained on a dataset of over 300,000 loan applicants. The data available to the AI includes all the details available to you -- like those in the loan applications you saw in the practice rounds -- and more. More information on the underlying AI model can be found in "**More details on AI**".

Over the course of this survey, we will show you **51 applications**. These applications will be randomly drawn from a sample where **66.66% of applicants** will repay their loan on time. The other 33.33% will not repay their loan on time.

Of the 51 randomly selected loan applicants you will see, on average, how many of them will not repay their loan on time?

Please give your answer in the form of a number.

[Next](#)

Figure 13: Detailed instructions give “More details on AI”



Figure 14: Instructions for each treatment condition highlight AI performance (*Binary Medium (50)* condition shown here as example)

[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:

Yes

This applicant's AI score is above 50

No

This applicant's AI score is below 50

Information on the AI-generated assistance

- When the score for an applicant is above **50** out of 100, Bailey the AI will tell you **Yes**; otherwise, it will tell you **No**
- For cases where the applicant **will repay the loan on time**:
 - The AI assistant tells you **Yes** 97% of the time
 - The AI assistant tells you **No** 3% of the time
- For cases where the applicant **will not repay the loan on time**:
 - The AI assistant tells you **Yes** 47% of the time
 - The AI assistant tells you **No** 53% of the time

[Next](#)

Figure 15: Detailed instructions for each treatment condition (*Binary Medium (50)* condition shown here as example)

[Instructions for next 2 tasks](#) [More details](#)

Information on the loan applicants

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

Information on the AI-generated assistance

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

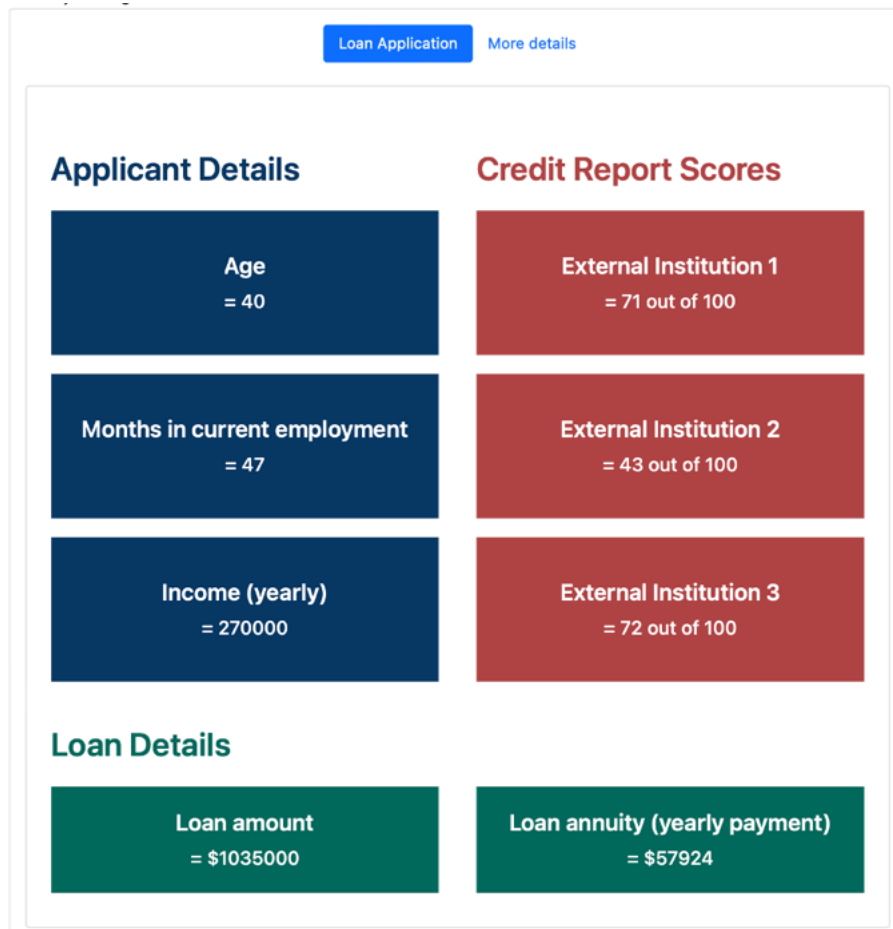


Information on performance in practice rounds

- During the practice rounds, you decided on 1 applications, where 66.66% applicants would go on to repay their loans. Here is how you performed, as well as how Bailey the AI would've performed if it had assisted you:
 - Of the 6 applicants who went on to **repay the loan on time**:
 - You correctly approved the loan for 1 out of 6 of them
 - Bailey the AI would've told you **Yes** for 6 out of 6 of them
 - Of the 5 applicants who would go on to **not repay the loan on time**:
 - You correctly denied the loan for 0 out of 5 of them
 - Bailey the AI would've told you **No** for 2 out of 5 of them

[Next](#)

Figure 16: 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 17: Example of first “posterior” elicitation

Applicant 1

Given what you currently know, 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.)

0

100

Next

Figure 18: Example of second signal (in this case, AI with *Binary Medium (50)* condition shown here as example)

Applicant 1

You have to decide whether or not to grant **Applicant 1** a loan.

You are given access to a score generated by Bailey the AI assistant that will tell you **Yes** if it thought the person was likely to repay the loan on time and **No** if it thought the person was unlikely to repay the loan on time.


More information on the composition of applicants, and the accuracy of the human-generated risk score, can be found by clicking "More details".

AI-generated score

More details

Yes

This applicant's
AI score
is above 50



Next

Figure 19: Approve/deny screen displays both signals again for easy decision-making

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.)

0

100

Do you want to approve this applicant for a loan?

☐ Approve

☐ Deny

Applicant Details

Age
= 40

Months in current employment
= 47

Income (yearly)
= 270000

Credit Report Scores

External Institution 1
= 71 out of 100

External Institution 2
= 43 out of 100

External Institution 3
= 72 out of 100


Loan Details

Loan amount
= \$1035000

Loan annuity (yearly payment)
= \$57924

Yes

This applicant's
AI score
is above 50



Next


8.3.2 Experiment II Additional Instructions

Figure 20: Instructions for *Trinary High (50, 70)* condition in Experiment II

[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:



This applicant's score is between **0-50**

This applicant's score is between **50-70**

This applicant's score is between **70-100**

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 21: Instructions for *Binary Medium (50)* condition in Experiment II

[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

8.3.3 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, automation bias, 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.

8.3.4 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 :

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

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.)**

Automation Bias We adapt this question from the base-rate neglect cognitive task in [Enke et al. \(2023\)](#). Automation bias (AB) is calculated as the ratio of the response to the automation bias question (see below) to the response to the base-rate neglect question (see above):

$$AB = \frac{\text{automationBias}}{\text{baseRateNeglect}}$$

An AB value of 1 indicates no automation bias. If $AB > 1$, it suggests automation bias. If $AB < 1$, it indicates automation neglect.

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 color of this randomly-drawn chip below, then you need to figure out (in percent) which bag was selected.

You are told that one **red chip** has randomly been drawn from the secretly selected bag. **What do you think is the likelihood (percentage chance) that the selected bag is the one with more red chips?** (Round to the nearest integer.)

Aggregation bias We adapt a variant of the leading example of [Piccione and Rubinstein \(2024\)](#) 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 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?**

8.4 Empirical Evidence

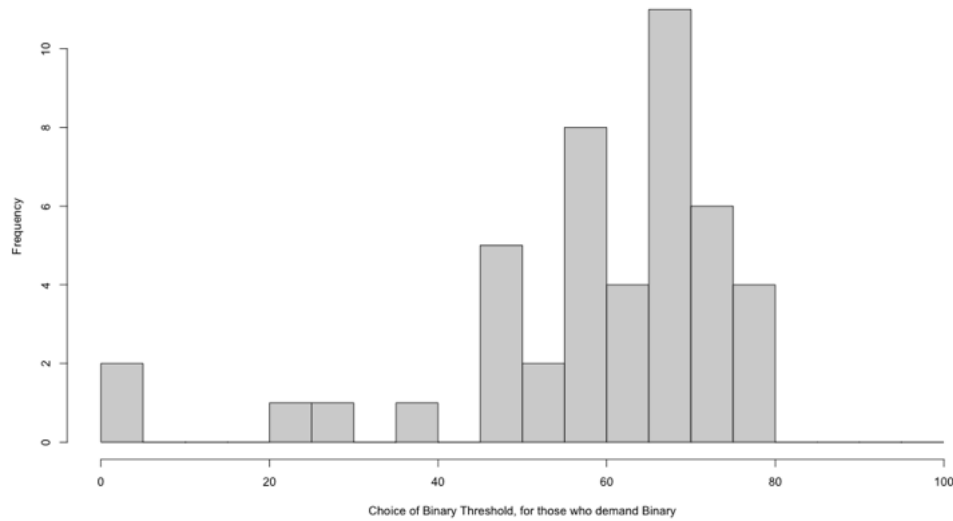
8.4.1 Demand

Table 4: 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 22: 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.

8.4.2 Heterogeneous treatment effects

For the regression tables in the following subsection, Column (1) shows coefficients from a regression of decision accuracy on treatment indicators and case fixed effects for the below median group. Column (2) shows coefficients from a regression of decision accuracy on treatment indicators and case fixed effects 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 \times 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 7) and the “Overall Accuracy” columns of these true/false positive and negative summary tables (e.g., Table 8) arise because the former incorporate case fixed effects, while the latter report simple means.

Gender

Table 5: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	Male (1)	Female (2)	Difference (2) - (1) (3)
Human-Only	−0.095*** (0.022)	−0.026 (0.019)	0.072** (0.036)
Probability	−0.048** (0.022)	−0.014 (0.019)	0.036 (0.022)
Binary High (70)	−0.078*** (0.022)	−0.040** (0.019)	0.037 (0.022)
Binary Low (30)	−0.054** (0.022)	−0.029 (0.019)	0.024 (0.022)
Constant	0.747*** (0.052)	0.663*** (0.044)	
Observations	3,100	4,350	7,450
R ²	0.304	0.275	0.283
Adjusted R ²	0.291	0.266	0.277

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 6: 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
<i>Female</i>						
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
<i>Male</i>						
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 7: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	Age (Below Median)	Age (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	−0.056*** (0.020)	−0.058*** (0.021)	−0.003 (0.035)
Probability	−0.031 (0.020)	−0.031 (0.021)	0.0002 (0.020)
Binary High (70)	−0.064*** (0.020)	−0.049** (0.021)	0.011 (0.020)
Binary Low (30)	−0.020 (0.020)	−0.062*** (0.021)	−0.040** (0.020)
Constant	0.716*** (0.047)	0.679*** (0.048)	
Observations	3,950	3,500	7,450
R ²	0.279	0.300	0.283
Adjusted R ²	0.269	0.289	0.277

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 8: 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
<i>Age (Above Median)</i>						
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
<i>Age (Below Median)</i>						
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 9: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	Exp (Below Median)	Exp (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.072*** (0.020)	-0.037* (0.021)	0.031 (0.035)
Probability	-0.040** (0.020)	-0.021 (0.021)	0.017 (0.019)
Binary High (70)	-0.089*** (0.020)	-0.019 (0.021)	0.069*** (0.019)
Binary Low (30)	-0.058*** (0.020)	-0.021 (0.021)	0.041** (0.019)
Constant	0.680*** (0.046)	0.722*** (0.049)	
Observations	4,200	3,250	7,450
R ²	0.277	0.318	0.284
Adjusted R ²	0.268	0.307	0.279

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 10: 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
<i>Exp (Above Median)</i>						
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
<i>Exp (Below Median)</i>						
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 11: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	Perceived (Below Median)	Perceived (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	−0.074*** (0.020)	−0.042** (0.021)	0.035 (0.035)
Probability	−0.038* (0.020)	−0.020 (0.021)	0.020 (0.020)
Binary High (70)	−0.053*** (0.020)	−0.062*** (0.021)	−0.006 (0.020)
Binary Low (30)	−0.044** (0.020)	−0.036* (0.021)	0.011 (0.020)
Constant	0.698*** (0.047)	0.700*** (0.049)	
Observations	3,800	3,650	7,450
R ²	0.293	0.282	0.283
Adjusted R ²	0.283	0.271	0.277

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 12: 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
<i>Perceived (Above Median)</i>						
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
<i>Perceived (Below Median)</i>						
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 13: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	Prac Acc (Below Median)	Prac Acc (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	−0.062*** (0.018)	−0.045* (0.025)	0.015 (0.034)
Probability	−0.034* (0.018)	−0.020 (0.025)	0.013 (0.017)
Binary High (70)	−0.063*** (0.018)	−0.048* (0.025)	0.013 (0.017)
Binary Low (30)	−0.054*** (0.018)	−0.008 (0.025)	0.047*** (0.017)
Constant	0.708*** (0.041)	0.680*** (0.059)	
Observations	5,100	2,350	7,450
R ²	0.275	0.313	0.283
Adjusted R ²	0.267	0.297	0.277

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 14: 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
<i>Prac Acc (Below Median)</i>						
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
<i>Prac Acc (Above Median)</i>						
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 15: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		Difference (2) - (1)
	Rubinstein (Below Median)	Rubinstein (Above Median)	
	(1)	(2)	(3)
Human-Only	−0.060*** (0.018)	−0.051** (0.026)	0.011 (0.035)
Probability	−0.038** (0.018)	−0.011 (0.025)	0.028 (0.018)
Binary High (70)	−0.072*** (0.018)	−0.029 (0.025)	0.044** (0.017)
Binary Low (30)	−0.057*** (0.018)	−0.006 (0.025)	0.053*** (0.017)
Constant	0.702*** (0.041)	0.696*** (0.059)	
Observations	5,050	2,400	7,450
R ²	0.279	0.302	0.283
Adjusted R ²	0.271	0.286	0.277

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 16: Share of True/False Positives and Negatives by Treatment and Rubinstein

Treatment	TP	TN	FP	FN	Overall.Accuracy	Total.Count
<i>Rubinstein (Below Median)</i>						
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
<i>Rubinstein (Above Median)</i>						
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 17: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	BU (Below Median)	BU (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	-0.064*** (0.019)	-0.027 (0.028)	0.032 (0.037)
Probability	-0.027 (0.019)	-0.002 (0.028)	0.023 (0.019)
Binary High (70)	-0.074*** (0.019)	-0.031 (0.028)	0.040** (0.019)
Binary Low (30)	-0.038** (0.019)	-0.026 (0.028)	0.009 (0.019)
Constant	0.691*** (0.044)	0.782*** (0.066)	
Observations	4,300	1,900	6,200
R ²	0.288	0.310	0.289
Adjusted R ²	0.279	0.289	0.282

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 18: 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
<i>BU (Below Median)</i>						
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
<i>BU (Above Median)</i>						
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

Automation Bias (AB)

Table 19: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	AB (Below Median)	AB (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	−0.067*** (0.017)	−0.022 (0.030)	0.045 (0.034)
Probability	−0.044*** (0.017)	0.027 (0.030)	0.066*** (0.017)
Binary High (70)	−0.072*** (0.017)	−0.021 (0.030)	0.058*** (0.017)
Binary Low (30)	−0.040** (0.017)	−0.037 (0.030)	0.008 (0.017)
Constant	0.706*** (0.038)	0.668*** (0.070)	
Observations	5,650	1,800	7,450
R ²	0.290	0.289	0.283
Adjusted R ²	0.284	0.267	0.277
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01			

Table 20: 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
<i>AB (Above Median)</i>						
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
<i>AB (Below Median)</i>						
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 21: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	BRN (1)	BRN (2)	Difference (2) - (1) (3)
Human-Only	−0.052*** (0.018)	−0.081*** (0.030)	−0.030 (0.036)
Probability	−0.044** (0.018)	−0.001 (0.030)	0.042** (0.018)
Binary High (70)	−0.071*** (0.018)	−0.038 (0.030)	0.036** (0.018)
Binary Low (30)	−0.028 (0.018)	−0.064** (0.030)	−0.031* (0.018)
Constant	0.706*** (0.042)	0.674*** (0.069)	
Observations	4,800	1,850	6,650
R ²	0.291	0.274	0.281
Adjusted R ²	0.282	0.252	0.275
<i>Note:</i> *p<0.1; **p<0.05; ***p<0.01			

Table 22: 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
<i>BRN (Below Median)</i>						
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
<i>BRN (Above Median)</i>						
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 23: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>		
	Share of Decisions Correct		
	CN (Below Median)	CN (Above Median)	Difference (2) - (1)
	(1)	(2)	(3)
Human-Only	−0.068*** (0.018)	−0.060** (0.027)	0.006 (0.035)
Probability	−0.026 (0.018)	−0.052* (0.027)	−0.029 (0.018)
Binary High (70)	−0.064*** (0.018)	−0.051* (0.027)	0.011 (0.018)
Binary Low (30)	−0.045** (0.018)	−0.046* (0.027)	−0.003 (0.018)
Constant	0.697*** (0.042)	0.714*** (0.063)	
Observations	4,750	2,100	6,850
R ²	0.293	0.296	0.290
Adjusted R ²	0.285	0.278	0.284

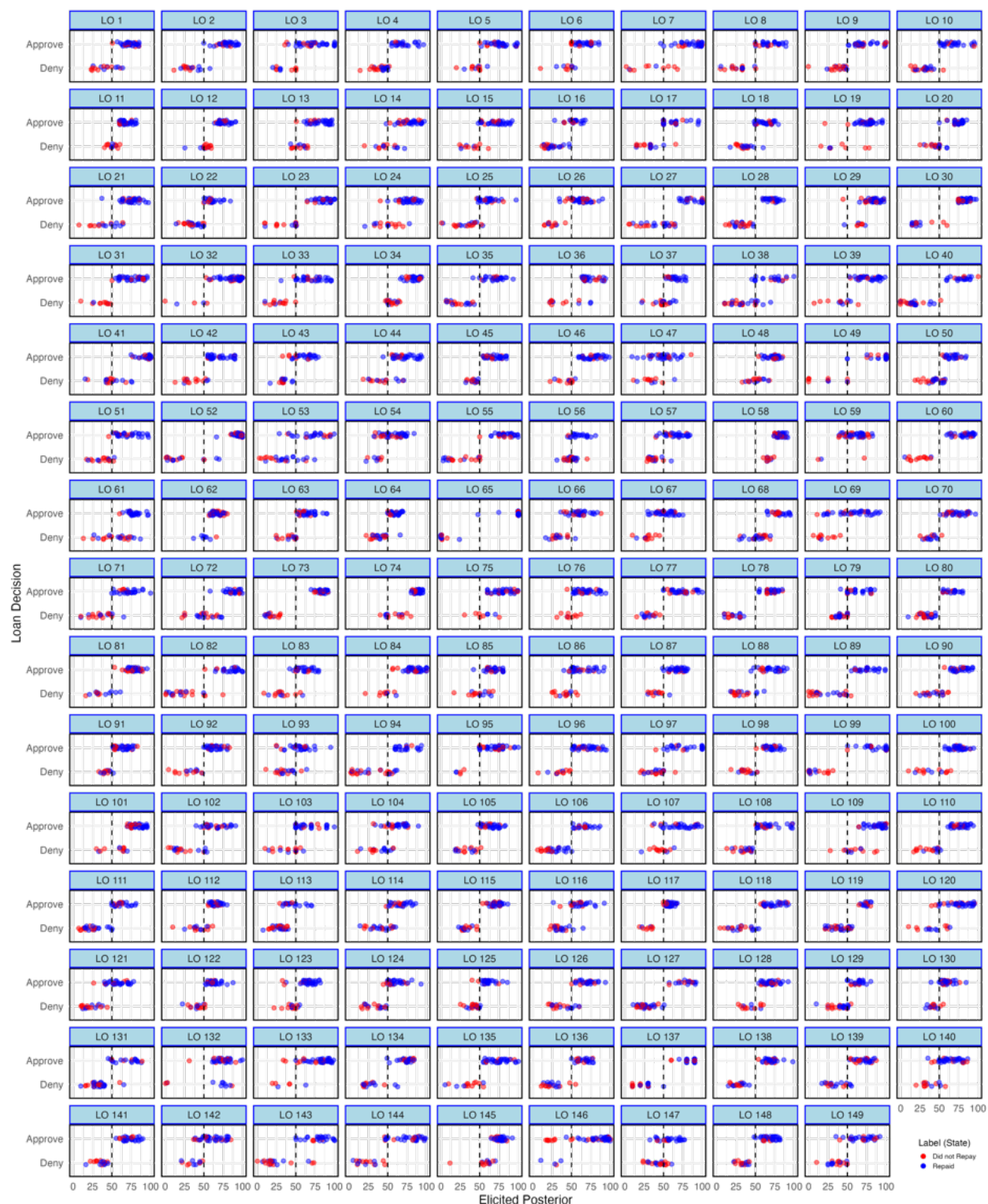
Note: *p<0.1; **p<0.05; ***p<0.01

Table 24: 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
<i>CN (Below Median)</i>						
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
<i>CN (Above Median)</i>						
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

8.4.3 Effective action thresholds

Figure 23: 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.

8.4.4 Where loan specialists are performing well

Table 25-27 present the shares of true and 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 the treatments influence decision-making accuracy across different levels of signal strength.

Table 25: 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 25, regardless of treatment, loan specialists are much better at approving correctly³¹ than denying correctly,³² despite the fact that they are incentivised equally for true positives and true negatives. This aligns with anecdotal evidence of real-life loan specialists' behavior: 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 26 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). 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

³¹True positive rate = $TP/(TP+FP)$, or the sensitivity

³²True negative rate = $TN/(TN+FN)$, or the specificity

Table 26: 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 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 27. 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.³³ 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.

Table 27: 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

³³Note, 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*.

8.5 Robustness checks

8.5.1 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 recruiting Prolific participants with Accounting and Business backgrounds (common among loan specialists) was ineffective due to differences in industry-specific knowledge and decision-making. Only after implementing rigorous screening did we identify Prolific participants resembling loan specialists, highlighting the need for targeted recruitment to ensure domain expertise, even at higher cost and effort.

Table 28: Decision-making outcomes under different treatment conditions

	<i>Dependent variable:</i>	
	Share of Decisions Correct	
	(1)	(2)
Human-Only	−0.061*** (0.021)	−0.056*** (0.018)
Probability	−0.040* (0.021)	−0.018 (0.018)
Binary High (70)	−0.073*** (0.021)	−0.043** (0.018)
Binary Low (30)	−0.043** (0.021)	−0.020 (0.018)
Constant	0.745*** (0.015)	0.645*** (0.041)
Observations	4,800	4,800
R ²	0.003	0.297
Adjusted R ²	0.002	0.289

Note: Regression estimates per Equation 1 of loan approval accuracy by treatment condition, restricting data to only LinkedIn/Indeed recruitment and with *Binary Medium (50)* as the omitted reference group. Column (1) reports estimates without fixed effects; Column (2) includes case-level fixed effects. Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01

Table 28 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*.

8.5.2 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 29, neither the order of the AI signal nor the elicitation of posterior beliefs significantly influences the outcomes.

Table 29: Robustness Checks

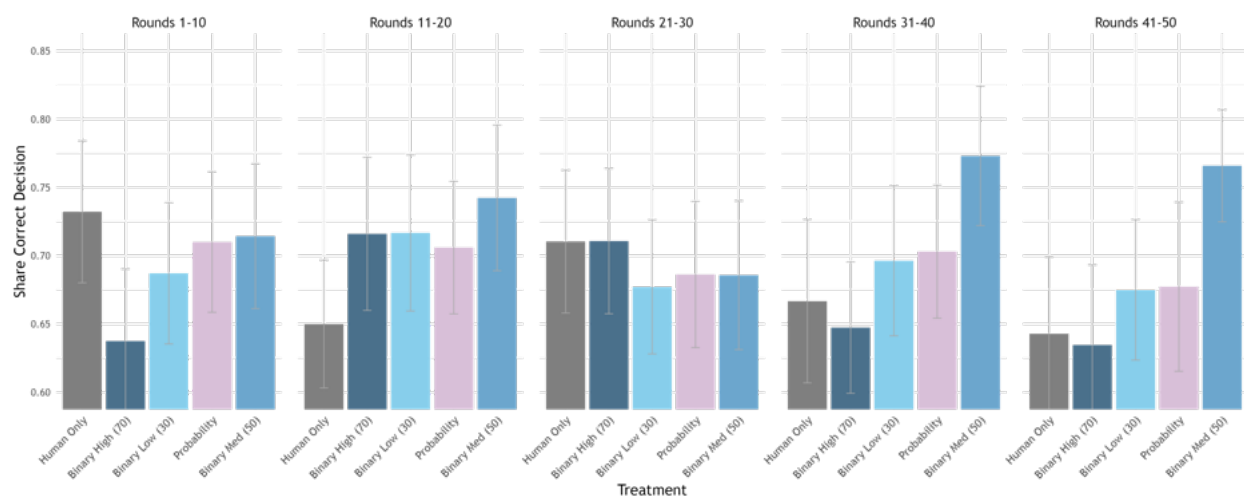
	<i>Dependent variable:</i>			
	Loan Decision	Posterior Belief	Loan Decision	Posterior Belief
	(1)	(2)	(3)	(4)
Elicitation Order			0.001 (0.018)	-0.424 (0.861)
First Posterior	-0.023 (0.019)	-1.081 (0.898)		
Constant	0.680*** (0.018)	60.454*** (0.855)	0.658*** (0.017)	59.855*** (0.815)
Observations	7,450	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 3, 4) and whether posterior beliefs are elicited after the second signal (Columns 1, 2). Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01

8.5.3 Restricting to each decision block

Figure 24 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 24: 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.