Calibrated Coarsening in Human-AI Interaction, Theory and Experiments

Draft: Please do not circulate

Bnaya Dreyfuss
Harvard University

Ruru Hoong *
Harvard University & HBS

Job Market Paper

March 20, 2025

Abstract

Artificial intelligence (AI) signals are increasingly being deployed as human decision-making aids across many critical applications. However, human cognitive biases can prevent even informative AI predictions from improving outcomes. We propose coarsening at optimal thresholds—partitioning the signal space into a fewer number of cells—as a way to improve decision-making outcomes while (i) keeping humans in the loop, (ii) modifying signals without deception, and (iii) adapting to various forms of cognitive biases or decision-making contexts. Within a Bayesian persuasion framework with privately-informed receivers, we study the optimal structure of signals. We empirically show in a randomised experiment with loan underwriters that the provision of AI signals optimally coarsened at the right thresholds improves overall decision-making outcomes.

*Contact: ruruhoong@g.harvard.edu. Ruru Hoong is especially indebted to Katie Coffman, Shane Greenstein, Jesse Shapiro and David Yang for their support. We are especially grateful for funding from Harvard Business School and Harvard Economics department, as well as the 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.

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 with AI predictions themselves, but with the critical role *humans* play in decision-making. While AI provides valuable input, its effectiveness depends on how humans interpret and act on its signals. AI typically assists rather than replaces human decision-making, as human 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. Therefore, the quality of AI-assisted decisions depends therefore not only on the quality of AI predictions but also on the human's own beliefs, and, crucially, on the humans' ability to effectively integrate both into a final decision.

However, there is a large body of work showing that humans often struggle to process and integrate information, suffering from various cognitive biases that distort belief updating (Tversky and Kahneman, 1974; Benjamin, 2019). Consequently, providing even highly informative AI signals may fail to improve decision-making if humans misinterpret or incorrectly aggregate this information. While interventions such as training might help, the wide variability, complexity, and context-dependence of these biases limit the systematic implementation and effectiveness of these solutions. Since removing humans from the decision-making process is often not an option, a critical challenge emerges: how can we improve human-AI collaboration outcomes across diverse biases and contexts, whilst preserving humans' ultimate decision rights?

In this paper, we propose *calibrated coarsening*—i.e., partitioning information into fewer cells at optimised thresholds—as a broad approach to improving decision-making outcomes that fulfills

[1] Algorithmic risk assessments in sentencing have shown no gains in public safety, incarceration rates, or arrestee behavior (Stevenson and Doleac, 2019; Garrett and Monahan, 2020; Imai et al., 2023). Similarly, AI aids in clinical decision-making have not improved outcomes (Brocklehurst et al., 2017; Nunes et al., 2017; Gaube et al., 2021), nor have AI probability scores enhanced diagnostic performance in radiology (Agarwal et al., 2023).

[2] The European Council's recently approved AI Act—the first-ever comprehensive legal framework on AI— requires human oversight over algorithmic decision making (European Commission, 2024). Beyond the EU, examples also abound. In clinical decision support, liability still rests with the clinician (Greenes, 2011; Mezrich, 2022; Cestonaro et al., 2023). In hiring, potential litigation concerns surrounding disparate impact (where a facially neutral procedure unintentionally disadvantages a protected class, prohibited by Title VII of the Civil Rights Act) have led employers to ensure human involvement when integrating AI into hiring processes (Kassir et al., 2023).

the following desiderata: (i) it keeps humans in the loop, (ii) it transparently modifies signals without deception, and (iii) it adapts to various cognitive biases or decision-making contexts.

Coarsening is already commonplace in many real-world applications.[3], yet it is often implemented heuristically rather than systematically. Our key contribution is to provide a framework for the optimal implementation of this widely used practice to improve decision-making outcomes. Specifically, we show that *calibrated* coarsening—that is, coarsening that accounts for and adapts to users' cognitive biases—can better align the integration of human judgment with AI-generated information, ultimately enhancing decision quality. Crucially, this approach does not depend is context-agnostic and does not rely on knowledge of the particular form of human bias; algorithmic predictions in any setting can be coarsened through a few simple steps, irrespective of the form of human bias at play. This adaptability is particularly important as biases vary not only across fields, but also among individuals within the same profession. By adjusting thresholds to pool information in a way that adjusts for decision-makers' cognitive tendencies, calibrated coarsening flexibly accommodates various biases while maintaining full transparency, a critical factor in fostering trust and adoption of AI-assisted decision-making (Mahmud et al., 2022).

We present a theoretical framework for leveraging this technique to improve performance, while also explaining why coarsening may naturally emerge in decision-making settings. To demonstrate its real-world applicability, we then empirically evaluate it in an experiment with loan officers, assessing how signals coarsened at our derived optimal thresholds affect loan approval outcomes. We find that the average diagnostic accuracy increases over two benchmarks: human-alone (no provision of AI signal), and perfect revelation (provision of the continuous AI signal, without coarsening).

To provide some intuition on how coarsening can improve AI-assisted decision making, consider the following simple numerical example. A loan officer seeks to approve loans that will be repaid, and reject those that will not. Suppose only 20% ($p = 0.2$) of loans are actually repaid. The officer has access to an AI-generated signal, s, which follows a normal distribution with a mean of 1 and standard deviation of 1 if the loan is repaid, and a mean of 0 with the same standard deviation

[3]For example, pretrial risk assessment scores are not continuous algorithmic probabilities but instead range discretely from 1-6 (see Figure 2, Angelova et al. (2023)). Similarly, the leading FDA-approved AI assistant for dental health no longer offers probabilistic assessments of periodontal disease – as they did in their initial interface – but rather uses red and green boxes to signify merely two levels of disease risk (see Figures 1a and 1b in Appendix).

otherwise. A Bayesian loan officer seeking to maximize accuracy would approve the loan if:

$$s > \frac{1}{2} - log(\frac{p}{1-p}) \approx 1.89,$$[4]

leading to a 1.4 percentage points gain in accuracy (relative to solely relying on the prior).

Suppose another officer suffers from *automation bias*, meaning they over-rely on the AI's signal. Specifically, they overweight the AI signal by a factor of $\alpha = 2.5$, following the signal weighting framework of Grether (1992).[5] This implies they would approve the loan whenever:

$$s > \frac{1}{2} - \frac{1}{\alpha} log(\frac{p}{1-p}) \approx 1.05,$$

meaning they approve loans more frequently than they should, relative to the Bayesian benchmark. Consequently, the officer's accuracy with the AI signal becomes 2.2 percentage points *lower* than the 80% they would achieve without the signal.

Now, consider a coarsened AI system that reports only whether the signal is above or below $s = 1.89$. It is straightforward to show that the officer then follows a simple rule: approve if the AI reports "Above 1.89" and deny otherwise. This effectively restores the biased officer's decision rule to the Bayesian benchmark, as they adopt the same effective approval threshold. By pooling together signals from the "disagreement" range $[1.05, 1.89]$ with those below 1.05, the loan officer is persuaded to deny the loan even when the raw AI signal (without pooling) would have led them to approve. This example illustrates how cognitive biases can make AI assistance ineffective and even harmful, and how calibrated coarsening—by pooling signals together at the right thresholds—can help correct for this.

To formalize this intuition behind why coarsening can effectively improve human-AI decision-making outcomes, we adopt an information design approach, specifically leveraging the linear programming framework in Kolotilin (2018) in order to directly study optimal signaling structures and messaging strategies without explicitly characterizing all possible posterior belief distributions.

In our model, we consider a binary state and two agents: a Bayesian Sender (e.g., an AI designer or manager) with access to an AI-generated signal, and a potentially biased human Receiver

[4] To provide some intution, this decision rule consists of two key terms: the first, $\frac{1}{2}$, reflects the 50% cutoff for classification, since the officer is maximising accuracy. The second log-odds ratio term, $-\log \frac{p}{1-p}$, adjusts for the prior probability of repayment p: a higher prior lowers the requirement, making approval easier, and vice-versa.

[5] In Grether's framework, decision-makers assign weights to signals when updating beliefs that can deviate from Bayesian updating. A factor $\alpha > 1$ indicates overweighting, while $\alpha < 1$ reflects underweighting. Here, $\alpha = 2.5$ means the loan officer gives the AI's signal more weight than that of an optimal Bayesian agent.

(e.g., loan officer), who observes their own private signal. The Sender commits to a mapping from AI signals into messages, while the Receiver combines this message with their private information to take an action intended to match the true state. The Sender's objective is to choose the messaging strategy that maximizes their own true expected utility (reflecting the unbiased posterior), while the Receiver acts to maximize their perceived expected utility (reflecting their biased posterior) to match the state.

We first demonstrate theoretically that when the Sender can condition the messaging strategy on the Receiver's signal, the optimal messaging strategy is a binary coarsening, which always improves the Receiver's expected utility compared to fully revealing the AI signal. When the Sender cannot condition the strategy on Receiver's signal, the optimal signaling structure becomes significantly more complex. Through simulations reflecting common biases found in human-AI collaboration (e.g., correlation neglect, automation neglect), we find that—in the absence of human private information—the best-performing binary threshold typically coincides with directly recommending the action suggested by the AI signal alone.

To demonstrate this empirically, we then conduct a randomised evaluation in the context of consumer loan approvals, showing that providing signals coarsened at the decision threshold (henceforth "binary optimal threshold") indeed improves decision-making outcomes. In our first experiment, we recruit $\tilde{1}50$ loan underwriters to make loan application decisions. We present a binary prediction policy problem: approve the loan to applicants expected to make timely payments, and deny the loan otherwise. Participants are incentivised for accuracy, by balancing the cost between false positives and negatives: a bonus of \$0.10 is awarded for each loan correctly denied and for each loan correctly approved. Participants see both a subset of a real loan application—with variables determined by those deemed most important to loan underwriters we interviewed as well as from the results of our pilots—as well as a signal from an AI model trained on a loan repayment dataset with similar loans from Home Credit. We vary the type of AI signal that is shown $\{$binary, probability, none$\}$, with sub-treatments varying the thresholds at which the binary signals are shown, inclusive of the optimal threshold implied by our simulations. In the binary condition, instead of receiving the AI probability score, participants see "Yes, this applicant's AI score is above [threshold]" if the score is above the specified threshold, and "No, this applicant's AI score is below [threshold]". We focus mainly on loan approval accuracy as our main result.

We find that providing loan specialists with optimally coarsened AI signals significantly enhances decision-making performance. Loan underwriters' approval accuracy (our metric for deci-

sion quality, since participants are incentivised for accuracy) improves dramatically under the *Binary Optimal* treatment—outperforming both the continuous AI probability score and unassisted decision-making. On average, participants achieve 4 percentage points higher accuracy compared to those receiving the AI score as a probability, an improvement that closes nearly 40% of the gap between human performance and the Bayesian benchmark. Even more strikingly, underwriters provided with optimally coarsened signals make 6 percentage points more accurate decisions than those without any AI assistance. In contrast, other binary thresholds yield no significant improvement over unassisted decisions, highlighting the precision and effectiveness of our optimised coarsening strategy. These results demonstrate the immense potential in leveraging AI and informational design to enhance human judgment.

We emphasise several caveats to our study. First, though we suggest that calibrated coarsening can help alleviate several behavioural biases, we do not claim to isolate any one particular bias as the driving mechanism behind the human-Bayesian gap. Such a claim would inevitably be context-dependent, and we view the approach's agnosticism to specific behavioural biases as a strength. Studies show that behavioural biases can vary depending on the specific context of the decision: they are not uniformly present across all decision-making scenarios (National Academies of Sciences et al., 2023) – the sorts of biases a loan officer might face in making loan approvals could be very different from the biases encountered by radiologists diagnosing chest cancer. Our insight applies regardless of the exact reason *why* over or under-reaction takes place: as long as there is a mistake in updating, coarsening has the potential to improve human-AI decision making. Though coarsening does not *always* help, it proves to be effective across a number of contexts and behavioural biases, as also proves to be the case in our empirical experiments.

Second, 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 the information presentation of the problem, (ii) replacing the individual with a model, and (iii) training people to use the appropriate information processing strategies. Coarsening arguably falls within the first category. The second of these debiasing approaches is not feasible as this 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 – that is, assigning different cases to either an AI or to an unassisted human (Raghu et al., 2019; Mozannar and Sontag, 2020; Athey et al., 2020; Bansal et al., 2021; Agarwal et al.,

2023). However, this solution is infeasible in scenarios where humans *must* make the final decision, or where there is an institutional preference for human oversight and autonomy. Furthermore, by necessitating a choice between an AI or human agent, it falls short in leveraging the collective information offered by both humans and AI algorithms.

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 more recent papers have designed interventions that have been shown to counter some behavioural biases like confirmation bias (Morewedge et al., 2015; Sellier et al., 2019), but these interventions are typically specific to the context and type of bias they address, can be costly, and tend to have limited long-term effects. Moreover, training has not been shown to be effective in correcting biases we deem most relevant to our scenarios, such as the over-weighting signals or correlation neglect (Enke and Zimmermann, 2019). Coarsening, on the other hand, provides a versatile solution applicable across various contexts and regardless of the specific underlying bias involved. Addressing the issue of biased updating towards or away from the AI by training humans more directly (e.g., by telling people to update less in certain areas of the distribution) would also likely have limited long-term impacts even if successful–as the AI trains on more data over time, the places in which people may be over-updating may change dynamically. In contrast, coarsening provides an easily adaptable, low-cost solution that can be integrated into the system, eliminating the need to continually train and 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). Even slight improvements in decision accuracy are expected to lead to significant economic returns (Cockburn et al., 2018; Agrawal et al., 2019). Notably, Agrawal et al. (2022) examine the fundamental components of decision-making—prediction and judgment[6]—and discuss how AI advancements in prediction may force a decoupling of (AI) prediction from (human) judgment, thereby transforming managerial and organisational structures. We suggest that a more complete characterisation of this shift necessarily has to take into account human cognitive biases,

[6]Per Agrawal et al. (2022), "judgment" here means human preferences, i.e., how humans evaluate the utility of different outcomes.

as well as the tools at the disposal of organisations to address them–of which coarsening is one of them. Biases can impact both prediction and judgment: on the former, humans do not just use AI predictions blindly. They may adjust their decisions based on private knowledge, or deviate from rational behaviour with or without such knowledge. It is therefore important to take into account *how* humans use AI predictions. Moreover, even humans that are well-calibrated on predictions may fail to choose the *actions* that reflect those beliefs and preferences, leading to sub-optimal decision-making outcomes. AI's purported transformative potential may therefore be more challenging and costly to realise than first appears; our paper offers a practical solution to these barriers in a way that is agnostic to the biases that drive them.

We also add 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, 2022; Fogliato et al., 2022; Lakkaraju and Farronato, 2022; Imai et al., 2023; Kim et al., 2024; Snyder et al., 2024), as well as identifying who benefits most from their provision (Caplin et al., 2024). Many also study behavioural aspects of human-AI interaction that may impede performance, including algorithmic aversion (Dietvorst et al., 2015, 2018), algorithmic appreciation or over-adherence (Logg et al., 2019; Banker and Khetani, 2019; Bai et al., 2021; Buçinca et al., 2021; Fügener et al., 2021), rational inattention (Boyacı et al., 2024), prospect theory (Ye et al., 2022), and correlation neglect (Agarwal et al., 2023). There is an emerging literature that proposes methods to improve the outcomes of collaborative human-AI design, including the provision of explanations and information about when the 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). Several propose using adaptive techniques to accommodate human private 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). Most closely, McLaughlin and Spiess (2022) model joint human-AI decision-making when agents' payoffs depend on the algorithmic recommendation itself, and show that withholding information when the algorithm is least certain can improve decisions. We show, both theoretically and empirically, that coarsening—beyond witholding information in uncertain areas—can be a much broader and powerful tool to systematically counter a wide range of decision-making biases.

Our work also relates to a broader literature in organisational behaviour, psychology, sociology, operations 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). Recent work has also examined the broader operational implications of AI, particularly in contexts where fairness and discrimination play critical roles (Kelley et al., 2022). Such considerations are especially relevant when understanding how human judgment interacts with AI predictions, as both can contribute to or counteract biases in decision-making processes. We contribute to this literature by introducing an approach that alters how individuals infer knowledge from AI tools – through threshold adjustments, we influence their willingness to follow AI recommendations. Our findings demonstrate that this approach has potential to improve how humans utilise knowledge from AI tools, ultimately leading to better decision-making outcomes.

One strand of this literature also focuses more specifically how individuals seek and 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 screening, prognosis, and treatment (Staradub et al., 2002; Taylor and Potts, 2008). AI or algorithmic inputs can also be viewed as a form of second opinion. Our paper speaks to when—and how—this second opinion can improve important decision-making outcomes, focusing specifically on algorithmic signals because they allow us to manipulate signals precisely and non-deceptively through threshold choice. Whereas adjusting the sensitivity or specificity of human second opinions (e.g., a radiologist colleague's input) may be challenging, algorithmic models allow for such adjustments in any setting.

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). This work has been largely theoretical; to our knowledge, we are the first empirical application that implements this type of informational design to fix behavioural biases.

The paper proceeds as follows. Section 2 presents a model of decision-making and some motivational examples of how coarsening can address different types of behavioural biases. Section 3 presents the experimental design for our experiment with loan specialists, Section 4 the empirical strategy, and 5 the data and results. Section 6 concludes.

2 Model

2.1 Basic framework

In our framework, we focus on classification problems where a *Receiver* (e.g., loan officer) 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 loan 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 signal the loan officer receives about the applicant's quality, based on their review of the loan application. The *Sender* (e.g., a bank manager with direct access to the output of an AI system or the designer/engineer of the AI model) has type $s \in S \subseteq \mathbb{R}$, representing the underlying AI-generated signal of applicant quality available to the Sender. The signals s and r are jointly distributed conditional on the state, reflecting the potential dependence of these signals on the binary state ω. Specifically, conditional on ω, s is distributed according to $F_\omega(s)$; moreover, conditional distribution on s and the state, r is distributed according to $G_\omega(r|s)$, where both marginals are continuous in s and continuously differentiable in r, and have strictly positive densities $f_\omega(s)$ and $g_\omega(r|s)$. For notational convenience, we will often denote unconditional distributions (those not conditioned on the state ω, e.g., $g(r|s) = \int_\Omega g_\omega(r|s)P(\omega)\, d\omega$) by omitting the subscript ω.

The Sender's and Receiver's utilities from inaction, $a = 0$, are normalised to zero. To reflect our binary classification setting, the Sender's utility depends on both the true binary state and the action taken by the Receiver. Specifically, we can express the Sender's relative utility[7] from the Receiver's action as a function of the state as:

$$V(\omega) = -c_{FP}\mathbb{I}(\omega = 0) + c_{TP}\mathbb{I}(\omega = 1)$$

and analogously the Receiver's utility as

$$U(\omega) = -c_{FP'}\mathbb{I}(\omega = 0) + c_{TP'}\mathbb{I}(\omega = 1)$$

For convenience, we can also express agents' utility in terms of their subjective expected values conditional on types. We define $\pi(r, s)$ and $\tilde{\pi}(r, s)$, the Sender's and Receiver's posterior beliefs

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

over $Pr(\omega|s,r)$, respectively. The Sender is Bayesian: that is, they update their (potentially subjective) prior over the state to their posterior using Bayes' rule. We can think of the Sender as the Bayesian benchmark—the Sender is a Bayesian updater, as *Bayesian updating is the best that one can perform given the information available*. We remain agnostic to how the Receiver updates their posterior beliefs – in general, they can deviate from the Sender, whether due to differing subjective prior over the prevalence of the state or deviations in their updating function.

The Sender's and Receiver's utilities from action $a = 1$ can therefore be expressed as $v(r,s)$ and $u(r,s)$, respectively, where v and u are continuous in s and continuously differentiable in r:

$$v(r,s) \equiv \mathbb{E}_{\pi(r,s)}\left(\hat{v}(\omega)\right)$$

$$u(r,s) \equiv \mathbb{E}_{\tilde{\pi}(r,s)}\left(\hat{u}(\omega)\right)$$

Although the Sender cannot take actions directly, they influence the Receiver by choosing an information disclosure mechanism $\Phi : S \to \mathcal{M}$, where $\Phi(s) = m \in \mathcal{M}$ is the message observed by the Receiver as a function of the Sender's type. This mechanism is designed ensuring the marginal distribution of s under Φ is equal to the prior marginal distribution of s.

The timing of the communication game proceeds as follows: First, the Sender publicly selects a mapping Φ. Next, the state ω is realized, and the signals r and s are drawn according to the joint distribution H_ω. Third, 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 make only the following restrictions on how the Receiver updates from messages:

Assumption 1. *For any mapping Φ and any $\Phi(s) = m \in \mathcal{M}$, we can write $\tilde{\pi}(r,m)$ as:*

$$\tilde{\pi}(r,m) = \int_{s \in S} \tilde{\pi}(r,s)\tilde{f}(s \mid r,m)ds$$

where for a given a messaging strategy Φ, $F(s \mid r,m)$ is the conditional distribution of s given m and r, and $\tilde{f}(s \mid r,m)$ satisfies the following conditions:

1. *$\tilde{f}$ is a proper PDF*
2. *$\tilde{f}$ has the same support as the true (i.e., Bayesian) conditional PDF:*

$$\tilde{f}(s \mid m,r) > 0 \iff f(s \mid m,r) > 0$$

3. *$\tilde{f}$ is continuous w.r.t. to the length of the inverse of m for any s and r. Formally, given Φ s and r, let $L(m)$ be the total length (or more generally the measure) of $m^{-1} \equiv \{\tilde{s} : \Phi(\tilde{s}) = m\}$, then for any two messages m, m' such that $m'^{-1} \subset m^{-1}$, for every $\epsilon > 0$ there exists $\delta > 0$ such that $L(m) - L(m') < \delta$ implies $|\tilde{f}(s \mid r, m) - \tilde{f}(s \mid r, m')| < \epsilon$.*

Essentially, Assumption 1 requires Receiver's posterior to respect the law of total probability: for each message m, the posterior must be a weighted average of the signals contained in its inverse.

In general, the Receiver's utility $u(r, s)$ can deviate from the Sender's $v(r, s)$ in two main ways: differences in fundamental underlying preferences (e.g., differing costs of true/false positives, $c_{TP} \neq c_{TP'}$ or $c_{FP} \neq c_{FP'}$) as well as differences in posterior beliefs (e.g., $\pi(r, s) \neq \tilde{\pi}(r, s)$). For the first part of this section, we abstract away from what drives the deviations in $u(r, s)$ and $v(r, s)$—in particular, unlike in Bayesian Persuasion we do not work with the set of possible posterior beliefs—and instead work directly with utilities expressed as functions of the Sender's and Receiver's types, as this allows us to take a linear programming approach per Kolotilin (2018).

Let $P_\Phi(s|m)$ denote the distribution of s given m under Φ. The distribution of s given m and r is then

$$P_\Phi(s|m, r) = \frac{\int_{\underline{s}}^{s} g(r|\tilde{s}) dP_\Phi(\tilde{s}|m)}{\int_S g(r|\tilde{s}) dP_\Phi(\tilde{s}|m)}.$$

The Receiver's expected utility from $a = 1$ given m and r is

$$\int_S u(r, s) dP_\Phi(s|m, r) = \frac{\int_S u(r, s) g_\omega(r|s) dP_\Phi(s|m)}{\int_S g(r|s) dP_\Phi(s|m)}.$$

Therefore, the Receiver strictly prefers to act if $\int_S \tilde{u}(r, s) dP_\Phi(s|m) > 0$ and strictly prefers not to act if $\int_S \tilde{u}(r, s) dP_\Phi(s|m) < 0$, where $\tilde{u}(r, s) = u(r, s)g(r|s)$.

We now impose a single crossing assumption that allows us to restrict attention to mechanisms Φ such that each message induces the Receiver to act, if and only if the Receiver's private information r is above some threshold type m. [8]

[8] There is work employing the notion of "potential decisions" where a monotonicity assumption (candidates with more favourable s are more likely to be approved with $a = 1$) allows the expression of the objective function in terms of the Receiver's compliance to the algorithmic assessment, and the decisions they make when deviating from the recommendation.(McLaughlin and Spiess, 2024) Our framework allows us to consider a wider range of mechanisms, not limited to those with an implied "algorithmic recommendation" — for each message m of a mechanism Φ, we can partition Receivers into two groups: those with r below an action threshold, and those above who do act. However, our assumption is stricter: in fact, it is stricter than the condition that for all $s_1, s_2 \in S$, functions $\tilde{u}(r, s_1)$ and $\tilde{u}(r, s_2)$ of $r \in R$ satisfy signed-ratio monotonicity (see Remark 1 of Kolotilin (2018)) and Theorem 1 of Quah and Strulovici (2012)).

Assumption 2. *(Single crossing assumption from Kolotilin (2018))*

For each distribution Q on S, there exists $r_Q \in R$ such that

$$\int_S \tilde{u}(r,s)\, dQ(s) \gtrless 0 \quad \textit{if} \quad r \gtrless r_Q.$$

Additionally, there exists a strictly decreasing function r^* that satisfies $u(r^*(s), s) = 0$ for all $s \in S$.

Now, consider a mechanism Φ where each message m induces a distribution Q of s. By Assumption 2, after observing m, the Receiver of type r_Q is indifferent between the two actions, while a Receiver of type $r > r_Q$ strictly prefers to act.

From here on, if a Receiver of type r is indifferent between two actions, we assume that he acts. Without loss of generality, we can focus on mechanisms Φ where each message m of Φ induces the Receiver to act if and only if $r \geq m$. This means the set of feasible messages is the image $R^* \equiv r^*(S)$ of S under r^*, and the Sender's expected utility from message $m \in R^*$ is

$$V(m,s) \equiv \int_m^{\bar{r}} v(r,s) g(r \mid s)\, dr.$$

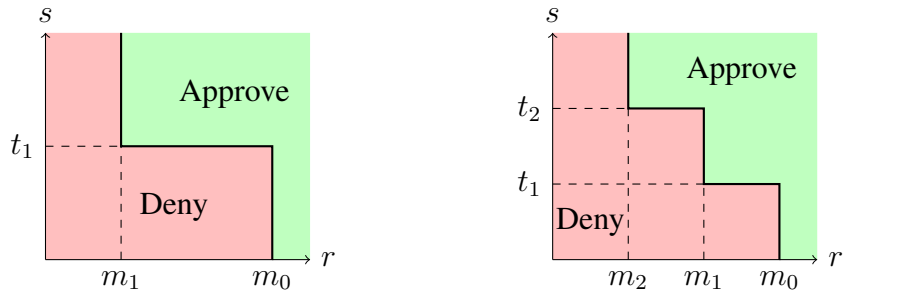
We define the following messaging strategies:

- A *perfect revelation mechanism*, denoted by Φ_{full}, is one that generates a different message for all $s \in S$
- A *full censorship mechanism*, denoted by $\Phi_\emptyset$, is one that generates the same message $\Phi(s) = m$ for all for all $s \in S$
- A *K-coarsening at a set of thresholds* T, denoted by Φ_T is a mechanism that generates K messages $\mathcal{M} = \{m_0, ..., m_{K-1}\}$, where $K = |T| + 1$. The mechanism maps each $s \in S$ to a message m_k according to the rule:

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

where $t_k < t_{k+1}$ for all k, $t_0 = \underline{s}$, and $t_K = \bar{s}$. We denote $T = \{t_1, ..., t_{K-1}\}$ as the set of thresholds chosen by the Sender to coarsen the signal space. Note of course that a K-coarsening generates $K - 1$ thresholds.

A quick note on the above mechanisms. A *perfect revelation mechanism* is a special case of a *K-coarsening mechanism* where $K = |\mathcal{M}| = |S|$. In this case, each threshold t_k corresponds to the k-th value of $s \in S$, i.e., $t_k = s_k$. Moreover, a *full censorship mechanism* is a *1-coarsening*. To illustrate *K-coarsening*, let us consider a 2-coarsening at thresholds T which generates two messages, m_0 for all $s \in [\underline{s}, t_1)$, and another message m_q for all $\in [t_1, \bar{s}]$. We can think of this as a binary coarsening that pools together all signals below the threshold t_1, and then pools together all signals above the threshold. As per Assumption 2, after observing m_k, a Receiver of type $r \geq m_k$ acts. The sets T and $\mathcal{M}$ are sufficient to characterise the actions of the Receivers (given their underlying type) under the messages in mechanism Φ_T, as we can see illustrated for 2 and 3-coarsening per Figures 1a and 1b below.



(a) Binary (2-coarsening) (b) Trinary (3-coarsening)

Figure 1: Two examples of *K-coarsening* mechanisms

2.2 Optimal mechanisms

2.2.1 Optimal mechanism when $u(r, s) = v(r, s)$

When the Receiver's utility is aligned with the Sender's utility $u(r, s) = v(r, s)$—that is, the Receiver acts as a Bayesian updater with identical priors and the same costs for true and false positives as the Sender—the Sender's optimal mechanism becomes straightforward: the optimal strategy is to fully reveal their type, effectively disclosing all the information they possess.

Proposition 1. *Let $u(r, s) = v(r, s)$ for all $r \in R$ and $s \in S$. Then the Sender's optimal messaging strategy is the perfect revelation mechanism Φ_{full}.*

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 m induces the Receiver r to act if and only if $r \geq m$. Under the *perfect revelation mechanism*, Φ_{full} generates the message $r^*(s)$ for each $s \in S$. Hence the expected utility of the Sender under Φ_{full} is:

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

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:

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

By Assumption 2, 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_{full} 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, $\mathbb{E}_{\Phi_{\text{full}}}[u] - \mathbb{E}_{\Phi}[u] > 0$ for any Φ that differs from Φ_{full}, making Φ_{full} the optimal mechanism that maximises the Sender's expected utility.

□

2.2.2 Optimal coarsening with known r

We now extend to the case where $u(r, s) \neq v(r, s)$, but where the Sender knows r. Consequently, for the remainder of this subsection, we treat r as a known constant.

If the Sender fully reveals s, the Receiver's strategy simplifies to:

$$\tilde{a}(r, s) = \mathbb{I}\left(s \geq \tilde{b}(r) \right),$$

where $\tilde{b}(r)$ is the Receiver's *decision boundary*, or the lowest s needed for the Receiver to approve the loan given their type r.

It is important to note that if the Sender could take an action directly, they would generally choose a different decision boundary since $u(r, s) \neq v(r, s)$. We denote the Sender-optimal strat-

egy by $a'(r,s)$, with the associated decision boundary $b'(r)$. The following proposition shows that when the Sender's and Receiver's decision boundaries diverge, the Sender can always do better than fully revealing s.

Let $\tilde{a}(r,m;\Phi)$ denote the action taken by a type-r Receiver conditional on a messaging strategy Φ and a message m. Finally, we say that the Sender can *fully implement* a' if there exists a strategy Φ such that $\tilde{a}(r,\Phi(s);\Phi) = a'(r,s)$ for all $s \in S$.

Proposition 2. *Let $r \in R$. If Sender's and Receiver's decision boundaries differ, i.e., $\tilde{b}(r) \neq b'(r)$, then:*

- *The Sender's optimal messaging strategy is a binary coarsening Φ_{t^*} with threshold $t^* \in (\min\{\tilde{b}(r), b'(r)\}, \max\{\tilde{b}(r), b'(r)\})$.*
- *Φ_{t^*} strictly increases the Sender's expected utility relative to perfect revelation.*
- *If $b'(r) > \tilde{b}(r)$ and $\tilde{a}(r,\Phi_\emptyset(s)) = 0$, or if $b'(r) < \tilde{b}(r)$ and $\tilde{a}(r,\Phi_\emptyset(s)) = 1$, the Sender can fully implement a'.*

Proof. Fix r. Starting with the first part, assume without loss of generality that the Sender's decision boundary is higher $b'(r) > \tilde{b}(r)$. Let

$$t^* = min\{max\{t : \tilde{a}(r,m_0;\Phi_t) = 0\}, b'(r)\}$$

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

To see why, notice that the Receiver will follow $\Phi_{\tilde{b}(r)}$, since by assumption 1, given Φ_t $\tilde{\pi}(r,m_0)$ is a weighted average of posteriors, all lower than $\tilde{\pi}(r,\tilde{b}(r))$, and $\tilde{\pi}(r,m_1)$ is a weighted average of posteriors all higher than $\tilde{\pi}(r,\tilde{b}(r))$. Clearly, for any $t' > \tilde{b}(r)$, Receiver follows $a(r,\Phi_{t'}(s)) = 1$. Moreover, by the last part of assumption 1, we can choose $\Phi_{t'}$ with t' close enough to $\tilde{b}(r)$ where $\tilde{\pi}(r,m_0) < \tilde{\pi}(r,\tilde{b}(r))$, so Receiver will also follow $\Phi_{t'}(s) = m_0$. Therefore, $t^* > \tilde{b}(r)$.

Notice that this immediately implies that Φ_{t^*} improves Sender's utility over perfect revelation: their utility over the segment $s \in [t^*, b'(r)]$ is strictly higher relative to perfect 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^*}. If Φ_{t^*} fully implements a', then we trivially arrive at a contradiction.

Assume not, then we must have $\tilde{a}(r, \Phi(s)) = 0$ for all $s_A \in (t^*, t']$ where $t' \leq t^s$. However, this immediately leads to a contradiction: if $\tilde{a}(r, \Phi(s)) = 0$ for some $s' > t^*$, there must exist a binary coarsening Φ' where $\tilde{a}(r, \Phi'(s)) = 0$, since this binary coarsening induces a distribution over s that is weakly stochastically dominated by Φ. This contradicts the definition of t^*. The proof for the case $b'(r) < \tilde{b}(r)$ is symmetric.

Moving on to the last part, first notice that a' is fully implemented iff Receiver follows $\Phi_{b'(r)}$. Assume $b'(r) > \tilde{b}(r)$, and $\tilde{a}(r, \Phi_\emptyset(s)) = 0$. Clearly, this implies $\tilde{a}(r, m_0; \Phi_{b'(r)}) = 0$, since this message generates a distribution over s that is stochastically dominated by the empty message. Similarly, since $b'(r) > \tilde{b}(r)$, the distribution over s given $\Phi_{b'(r)}(s) = m_1$ stochastically dominates the distribution given the message $\Phi_{\tilde{b}(r)}(s) = m_1$, and since $\tilde{a}(r, 1; \Phi_{\tilde{b}(r)}) = 1$, we must also have $\tilde{a}(r, 1; \Phi_{b'(r)}) = 1$. Therefore, Receiver follows the signal. The proof for the other case is symmetric. □

Contingent on knowing r, the Sender can always construct a binary coarsening that improves outcomes over perfect revelation – and this holds for a large class of deviations from Bayes.

2.3 Application: Binary coarsening with automation bias

We now turn to specific, well-documented behavioral biases to contextualize and simplify our problem. To develop intuition for how the extent of these different biases can affect decision-making—and therefore the optimal coarsening—we introduce several assumptions about the decision-making process (e.g., breaking down the optimal mechanism into choosing a joint distribution over posterior beliefs and actions to maximise utility; as well as *how* the Receiver updates their posteriors) while ensuring that these do not alter our previous results. We begin with automation bias (or automation neglect), the tendency to overweight 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).

We now explicitly model how the Receiver updates their beliefs in response to signals and optimizes over the induced posterior distributions. The Receiver's strategy, $a : R \times M \to \{0, 1\}$, maximizes their subjective expected utility:

$$a(r, m) = \mathbb{I}\left(\tilde{\pi}(r, m) \geq \frac{c_{FP}}{c_{TP} + c_{FP}}\right)$$

Note that $\tilde{\pi}(r, \tilde{b}) = \frac{c_{FP}}{c_{TP}+c_{FP}}$. It will also prove helpful at this point to re-write the utilities from taking an action as functions of the log likelihoods:

$$v(s, r) = log\left(\frac{\pi(s, r)}{1 - \pi(s, r)}\right) - log(c)$$

$$u(s, r) = log\left(\frac{\tilde{\pi}(s, r)}{1 - \tilde{\pi}(s, r)}\right) - log(c)$$

where $c \equiv \frac{c_{FP}}{c_{TP}}$.

2.3.1 When r is known to the Sender

As before, we fix $r = r'$. Assume the Receiver is automation biased, i.e., places incorrect weight on AI signals when calculating the posterior:

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

Since we fix r, $\frac{g_1(r'|s)}{g_0(r'|s)}$ becomes a known function of s. Therefore we can write:

$$v(s) = \log\left(\frac{p}{1 - p}\right) + \log\left(\frac{f_1(s)}{f_0(s)}\right) + \log h(s) - \log(c)$$

and

$$u(s) = \log\left(\frac{p}{1 - p}\right) + \alpha \log\left(\frac{f_1(s)}{f_0(s)}\right) + \log h(s) - \log(c)$$

where $\log h(s) = \frac{g_1(r'|s)}{g_0(r'|s)}$.

Assume that conditional on s, r is uninformative, i.e., the human has no private information. Conditional on r being known, this assumption is non-substantive and we make it to simplify the presentation. Note that since $g_0(r|s) = g_1(r|s)$ for all r in the case of uninformative r, the $\log h(s)$ term drops out of the utility functions. In what follows, we also assume without loss of generality that $c\frac{1-p}{p} < 1$, i.e., the "default" action (absent any information on s) is $a = 0$ (the opposite case is symmetric).

We now show that for α high enough, the Sender can choose a binary coarsening that fully

implements their preferences. Let $s \mid r, \omega \sim N(\omega, \sigma_s)$.[9] Since s is normally distributed, we have:

$$\log\left(\frac{f_1(s)}{f_0(s)}\right) = \frac{2s-1}{2\sigma_s^2}$$

Therefore, under full revelation the Receiver chooses $a = 1$ if:

$$s \geq \frac{1}{2} + \frac{\sigma_s^2}{\alpha}\log\left(c\frac{1-p}{p}\right)$$

whereas the Sender would choose $a = 1$ if:

$$s \geq \frac{1}{2} + \sigma_s^2\log\left(c\frac{1-p}{p}\right) \equiv s^*$$

Intuitively, when $\alpha \to 0$, the Receiver places increasingly lower weight on s, and therefore chooses $a = 0$ even for high values of s. In contrast, as $\alpha \to \infty$, the Receiver completely relies on the signal, choosing $a = 1$ for values of s closer to 1 (the mean under $\omega = 1$), and $a = 0$ for values closer to 0.

Observe that since the decision boundary is determined by the ratio $\frac{\sigma_s^2}{\alpha}$, automation bias is isomorphic to misperception of the variance s: over-weighting the signal in the posterior calculation is decision-equivalent to treating it as less noisy than it actually is.

Applying Proposition 2 to our case, we can show the following result:

Result 1. *Under automation bias and jointly normal signals:*

1. *For any $\alpha > 0$, optimal binary coarsening strictly improves the Sender's utility over full revelation*

2. *There exists $\overline{\alpha} < 1$ such that if $\alpha \geq \overline{\alpha}$, the optimal binary coarsening is $t^* = s^*$, and t^* fully implements the Sender's preferences*

In other words, under *automation bias* ($\alpha > 1$), binary coarsening can always fully implement the Sender's preferences. Under *automation neglect*, binary coarsening can always achieve higher utility than full revelation, but can fully implement the Sender's preferences only when α is sufficiently close to 1.

[9] Note that under a Gaussian distribution, setting $\mu = \omega$ is without loss of generality, as normal distributions can be shifted and rescaled to get mean 1 and 0 under each state, respectively.

For intuition, observe that under binary coarsening, the Receiver chooses $a = 1$ if and only if:

$$\alpha E(s \mid m) \geq \frac{1}{2} + \sigma_s^2 \log\left(c\frac{1-p}{p}\right)$$

When the binary threshold $t = s^*$, the mean of s is a weighted average of values above s^*, and therefore by the definition of s^*, the inequality holds strictly when $\alpha = 1$. Clearly, it also holds for $\alpha > 1$. Moreover, there exists $\overline{\alpha} < 1$ such that it holds with equality.

In words, since non-biased ($\alpha = 1$) agents adhere to a Sender-optimal binary threshold, automation-biased ($\alpha > 1$) agents place higher weight on the signal, and are therefore even *more* willing to adhere to it. In contrast, sufficiently automation-neglectful ($\alpha < 1$) agents are persuaded to choose $a = 1$ only for relatively high values of mean s, and therefore the Sender must choose a threshold high enough to induce them to act.

Figure 2 illustrates this result graphically. For simplicity, we set $c_{FP} = c_{TP} = c'_{FP} = c'_{TP}$ in this example. The dotted red line is the Receiver's decision boundary, i.e., for any given value of α, the values of s over which they would have chosen $a = 1$ (if they were to fully observe the Sender's type s). The horizontal black dotted line is the Sender-optimal boundary, i.e., the value of s above which the Sender would want the Receiver to act. Note that where the red dashed curve crosses this black line (exactly at $\alpha = 1$) is the case of no bias: if the Receiver is unbiased, the Sender's ideal threshold matches the Receiver's actual threshold under full revelation. When $\alpha < 1$ (automation neglect), the Receiver's threshold for approval is higher than the Sender's optimal, meaning they require a stronger signal to act, since they place less weight on s. When $\alpha > 1$ (automation bias), the threshold gets lower because the Receiver over-reacts to s, approving more readily for any given s.

The blue area denotes the set of bias parameters and thresholds (α, t) pairs such that the Receiver would "adhere" to a binary-coarsened signal at threshold t. Since we are in a scenario where the default inclination is to deny the loan absent any information (the prior probability that $\omega = 1$ is low, i.e., $p = 0.35$), the threshold needs to be sufficiently high to induce an action. This region extends below the red dashed line, showing that the Sender can choose a binary threshold lower than the Receiver's full-information cutoff and still induce the Receiver to act. In particular, the optimal threshold in this case is given by the upper envelope of the solid blue line and the dashed black line.

Consequently, by choosing the proper binary coarsening, the Sender can always strictly im-

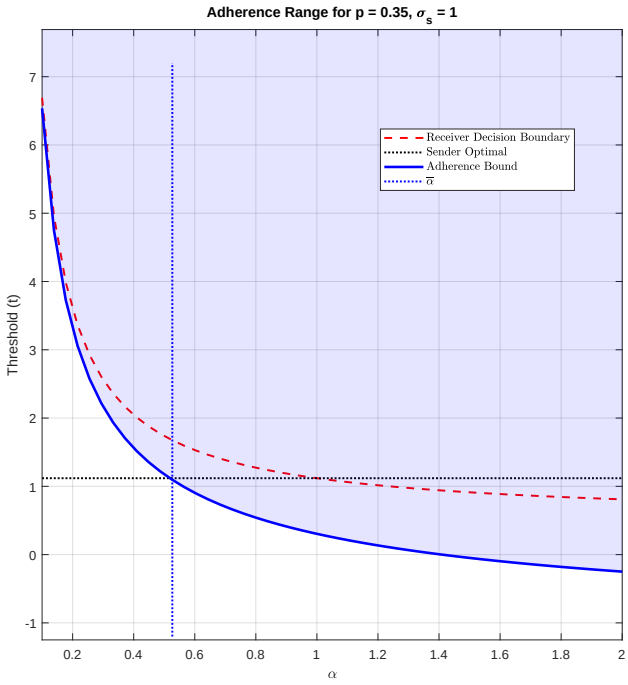


Figure 2: Automation neglect with no private information: graphical illustration

prove their own utility over what would happen under full revelation by inducing the Receiver to act even for s values such that under full revelation, they would not have acted.

Finally, the vertical dotted blue line at $\overline{\alpha}$ shows the lower bound on α for which the Sender can fully implement their preferences. This happens when the black dotted line—the Sender's optimal boundary—is contained in the blue area, so the Receiver will strictly adhere to the Sender's optimal boundary. When $\alpha < \overline{\alpha}$, the Receiver's under-weighting of the signal forces the Sender to pick a threshold higher than s^* if she wants the Receiver to act (but still improves Sender's utility over full revelation).

2.3.2 When r is unknown to the Sender: an illustrative example

What happens, then, if the Sender does not know r? In some contexts, it might be impractical to elicit r for every case, or to tailor the threshold to every case. Moreover, it could be challenging to handle strategic r reports by the Receiver.

We start with an illustrative example where the Receiver has automation neglect and no private information (r is uninformative conditional on s), and where $c_{TP} = c_{FP} = c'_{TP} = c'_{FP}$. Figure 3 then shows the decision boundaries b and b' described in the model.

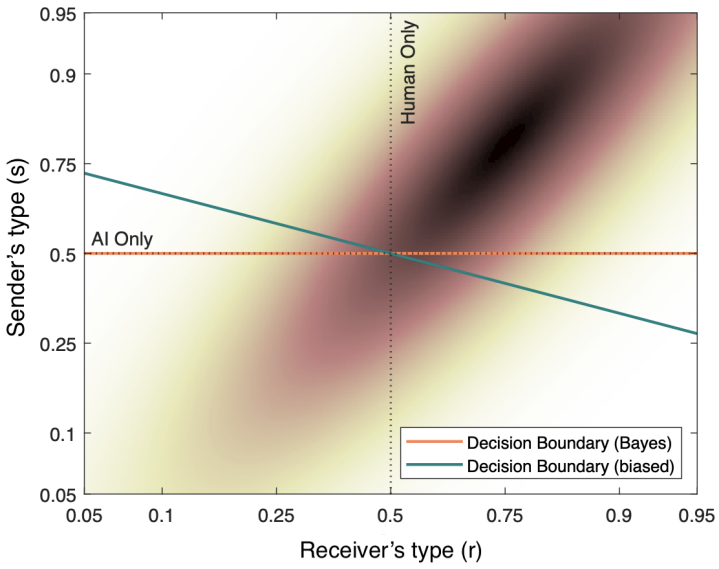
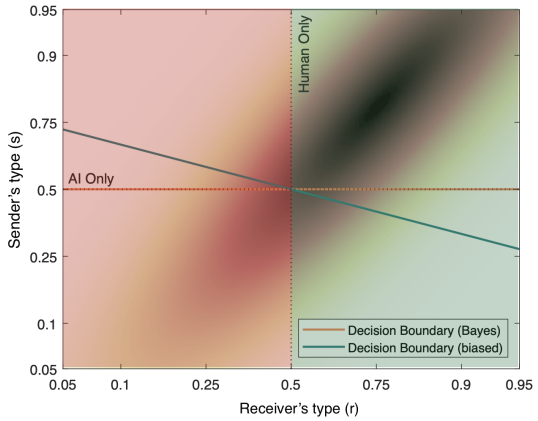


Figure 3: Decision boundaries, under automation neglect

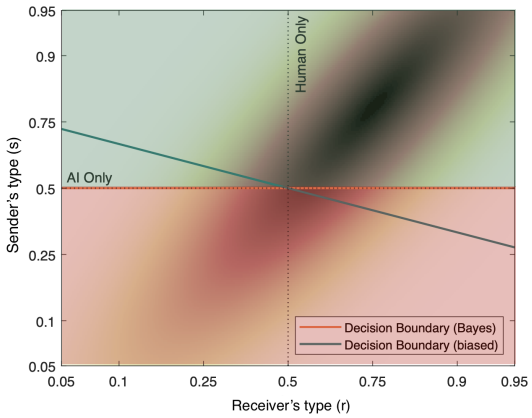
The biased Receiver's boundary denoted by the blue line and the Sender-optimal Bayesian boundary by the orange line. The shading denotes the joint distribution of signals, which here we simulate as highly correlated, as is often the case in the settings we care about (e.g., there is substantial overlap in the type of information AI and the human experts both have access to). Notice that because we are in a case with no private information, the Bayesian orange line is horizontal: the best that a Bayesian can do is to approve when the AI signal implied posterior is above 0.5, and deny when below. The biased Receiver's boundary is steeper because they face automation neglect: they are more responsive to their *own* signal than they should be.

A nice feature of this graphical illustration is that we can then easily visualise our desired benchmarks by the grids and boundaries. For example, Figure 4a shows what a human with no AI assistance at all would do – they would approve everything to the right of $r = 0.5$, and deny everything else. On the other hand, Figure 4b illustrates the Bayesian benchmark—in other words, the best any decision-maker can do given the information at hand—they would approve everything above the orange line, and deny anything below. The scope for improvement of outcomes (over full revelation), then, must be the area between the Bayesian and the biased decision boundaries denoted in Figure 4c. In the blue area to the left of the graph, a biased Receiver would deny whilst a Bayesian would approve the loan, and vice-versa for the blue area to the right.

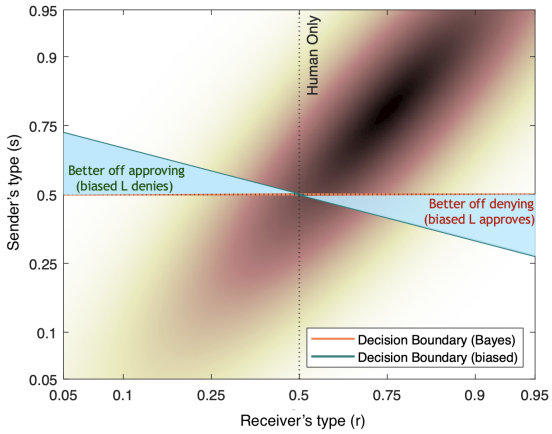
To understand the potential treatment effect of coarsening the AI signal, we then define one more object – adherence bounds. In the context of a binary coarsening, the adherence bounds are



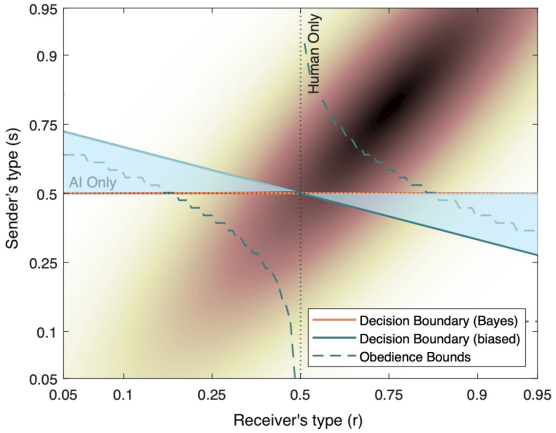
(a) Human-only Benchmark



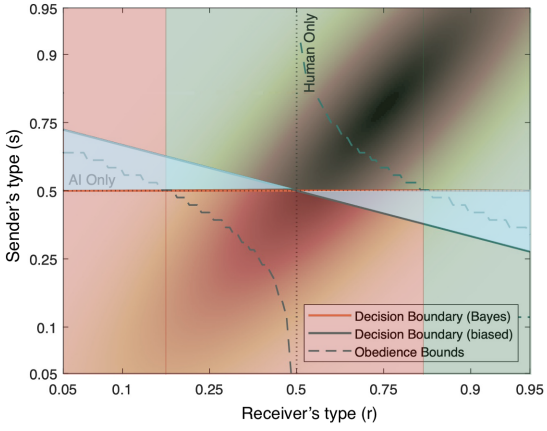
(b) Bayesian Benchmark



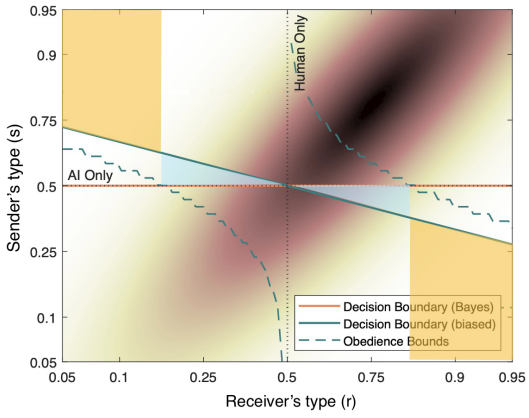
(c) Scope for Outcome Improvement



(d) Adherence Bounds



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



(f) Net Treatment Effect of Coarsening at Decision Threshold

Figure 4: Decision boundaries, automation neglect example

the two dotted curves in 4d denoting the lowest and highest binary AI signal thresholds such that a biased Receiver with a given r will *adhere* to the binary signal. When the Sender coarsens the AI signal at a particular threshold (take, for example, where $s = 0.5$), the biased Receiver follows the signal between the adherence bounds, and otherwise follows their own private signal. This is illustrated in 4e: between the adherence bounds, the Receiver approves when the AI signal is above the coarsening threshold, and denies otherwise. At low enough r (where the lower adherence bound intersects 0.5), the Receiver always denies. At high enough r (where the higher adherence bound intersects 0.5), they always approve. The net treatment effect of that binary coarsening (over full revelation) then therefore depends on the size of these area(s), weighted by the joint distribution. The blue areas in Figure 4f denotes the gains from coarsening, whereas the yellow are denotes the loss from coarsening.

As it happens, when the human has automation neglect ($\alpha < 1$) and no private information, the optimal binary threshold is indeed at $s = 0.5$, and is robust over a range of values and a range of biases like correlation neglect. This is what we test empirically in our loan experiment, finding that coarsening at the optimal binary threshold improves decision-making outcomes.

3 Experimental design

Our first randomised evaluation tests the deployment of coarsened signals in consumer loan approvals, in a context where loan underwriters have no private information. This is an important case to address: in a world dominated by big data and multi-modal models, many decision-making contexts now involve scenarios where AI has access to the same—or even more—underlying data than humans, but humans still have to make the final decision. We recruit 149 loan underwriters to make hypothetical loan approval decisions, showing that providing them with signals coarsened at the optimal threshold leads to better decision-making outcomes than providing full information.

3.1 Experimental context: loan approvals

We choose to focus on loan approvals as they are an important use-case of AI, as well as an industry with a long history of leveraging algorithmic predictions. Financial institutions—and the decision-makers within them—face a binary prediction policy problem: approve the loan to applicants expected to make timely payments, and deny the loan otherwise. So long as applicants

will repay the loan, institutions want to approve the loan, not only because interest and fees are often a primary source of revenue, but also for reasons of financial inclusion.

In our experiment, we utilise a dataset of real loan applications from the multi-national financial institution Home Credit, which includes a wide range of relevant variables including: current and previous loan applications and repayment history, credit bureau records (previous loans, credits, cash), demographic information (e.g., gender, age, car/house ownership and residential information, family, education, and employment), credit card information, normalised credit scores from external sources, amongst others. One nice advantage of this particular context is that ground truth exists: the dataset includes real labels for whether or not a person eventually repaid their loans or time or not.[10] 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—whether or not the loan is actually repaid.

3.2 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 officer, underwriter, or processor (henceforth, "loan specialists"). Recruitment primarily took place through temporary job postings on LinkedIn and Indeed, supplemented by targeted outreach to members of the National Association of Mortgage Underwriters (NAMU) and the National Association of Mortgage Processors (NAMP), as well as contract hiring on Upwork. A subset of the loan specialists were recruited through custom screening on Prolific, with stringent screening processes implemented to filter out scammers.[11] Participants recruited over Prolific must additionally meet the following requirements to be included in the study: they need 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 branch workplace, *or* demonstrate familiarity with loan officer, underwriter, or processor responsibilities in their decision descriptions

[10] We thank Kirill Odontsov at Home Credit for providing us with the data and the labels.

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

(such as referencing income-to-debt ratio, income stability, capacity, or collateral). Besides the first condition, these filtering criteria were *not* made known to them so as to reduce gaming.

To ensure the robustness of our findings, we conduct a check in the Appendix that excludes participants recruited through Prolific. This allows us to assess whether the inclusion of Prolific participants—who were subject to more stringent screening but recruited through a general-purpose survey platform—affects the results. While Prolific offers flexibility and cost-advantages in recruitment, it differs fundamentally from targeted industry outreach. In fact, in our pilots, attempts to recruit Prolific participants with Accounting and Business background (majors common amongst loan specialists) proved ineffective, as their familiarity with industry-specific practices and decision-making deviated significantly from that of loan specialists recruited elsewhere. Only after implementing rigorous screening criteria were we able to identify Prolific participants resembling loan specialists; this underscores the importance of prioritising targeted recruitment to ensure participants actually have domain expertise, even if it is at greater cost and effort. As shown in Appendix 7.3.1, our main results remain robust even when Prolific participants are excluded.

Once through with screening, qualified participants are directed to complete an online survey hosted on the Otree platform via Heroku. The survey takes an average of 60 minutes, and participants are told that they will have an opportunity to participate in a follow-up survey in the coming months, as part of a potential separate follow-up experiment.

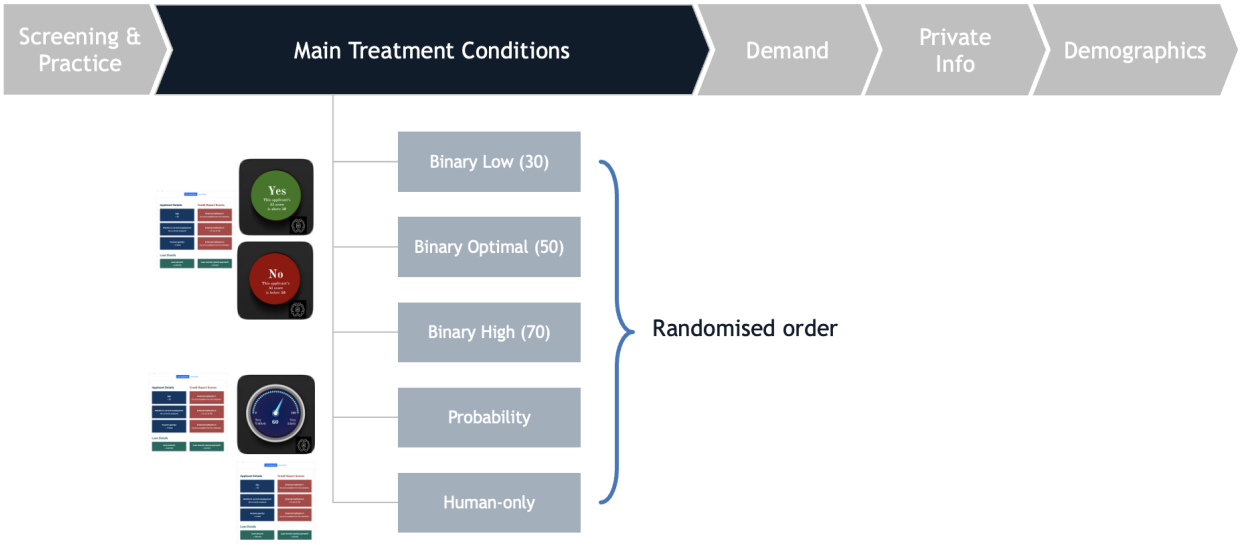


Figure 5: Experimental survey flow

3.3 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 randomized order. Depending on their assigned treatment group, participants may also receive a signal from an AI model trained on a loan repayment dataset from Home Credit. After reviewing this information, they are directed to a landing page where they respond to two questions:

- "What do you think is the % chance that the applicant will make their future payments on time? [Click on the blue bar to choose a number between 0 and 100.]"
- "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:

- **The prior:** The share of applicants in the population expected to repay their loans.
- **Their performance in the practice rounds:** How many loans they approved/denied were repaid or defaulted.
- **Details of the underlying AI model:** Information on the AI training dataset, score interpretation, performance metrics, and signal density based on repayment outcomes.
- **Details of the assisting AI:**
  - The threshold above which the AI says Yes, and below which they say No (Binary only)
  - The percentage of time the AI assistant says Yes/No, separately for when the applicant will repay the loan on time and when they will not m (Binary only)
  - The average score the AI assistant gives, separately for when the applicant will repay the loan on time and when they will not (Probability only)

Participants can access most of this information throughout the experiment via a "More details" tab on the right-hand side.

Incentives are structured to balance the cost between false positives and negatives: a bonus of $0.10 is awarded for each loan correctly denied (for default applicants) and each loan correctly approved (for on-time payers). 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).

3.3.1 Loan application details

Participants are shown the following variables as part of a loan application (see Figure **??** in the Appendix for an example). These key variables were identified through interviews with loan underwriters 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 underwriters use to assess and make decisions on applications.

- **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:** Normalised credit score from external data source #1 (e.g., FICO, VantageScore), ranging from 0-100, with 100 being the most credit-worthy.
- **External Institution 2:** Normalised credit score from external data source #2 (e.g., FICO, VantageScore), ranging from 0-100, with 100 being the most credit-worthy.

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

3.3.2 AI model details

We use one of the top-performing AI models from a public competition on the Home Credit dataset, trained on 300,000 loan applicants and their repayment outcomes. The data available to the AI includes details of the loan, applicant demographics (age, gender, education, family, car and house ownership), employment details, previous Credit Bureau enquiries and loans, normalized credit scores (if available), and whether or not they repaid their loans on time. 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.

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. The score shown as the AI probability score to participants represents the implied posterior probability that an individual will repay the loan. This probability is determined using Bayesian updating, which incorporates both prior information (the initial likelihood that any borrower will repay the loan) and new evidence from the AI model.

3.4 Treatment conditions

In our main experimental variation, we manipulate the type of AI signal participants receive (if any). The treatment conditions are structured as follows: {Human Only, Probability, Binary Optimal, Binary High, and Binary Low}.

In the Human Only condition, participants make decisions based solely on the loan application information, without any AI assistance. In the Probability condition, they 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, instead of receiving a probability score from the AI, participants are given a simple binary outcome: "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. Across participants, blocks appear in randomised order. We also cross-randomise this 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 before 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 the posterior is elicited both before and after the second signal is shown and the loan application is presented first. The randomisation of these conditions allows us to later conduct robustness checks for potential anchoring effects, ensuring that the order in which participants see the AI signal or the posterior elicitation does not influence their loan decisions or final posteriors, as shown in Appendix 7.3.3.

To increase power, loan specialists are shown all 5 decision blocks in randomised order. The specific loan applications are also randomised in order, and so in general will differ across the different treatment conditions for different people.

3.5 Additional Survey Questions

Demand for AI – We 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. We then 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 randomly implement one row of the multiple price list for the final two decisions.

Perceived private information – At the start 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.

Biases – Participants answer a series of cognitive questions documented in Appendix 7.1 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).

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 underwriters 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.

4 Empirical strategy

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

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

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 probability treatment as the reference base level, or control). 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_ts – in particular we test the hypothesis that:

- How the *Binary Optimal* arm performs relative to the *Probability* treatment, i.e., test the null that $\beta_{\text{binary-opt}} = 0$
- How the *Binary Optimal* arm performs relative to the *Human Only* treatment, i.e., test the null that $\beta_{\text{binary-opt}} = \beta_{\text{human-only}}$

The above tests were pre-registered in our pre-analysis plan (AEARCTR-0013716). We also documented in the plan that we expected that *Binary Optimal* will perform better than the *Bi-*

nary High/Low treatment conditions as well, although we remained agnostic to whether *Binary High/Low* will perform better than the *Human* or *Probability* treatment.

5 Empirical evidence: calibrated coarsening improves decision-making outcomes

5.1 Data

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%

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.

5.2 Main results

Our main finding provides empirical evidence that calibrated coarsening improves decision-making, as illustrated in Figure 6 and Table 2. Figure 6 clearly demonstrates that *Binary Optimal (50)* treat-

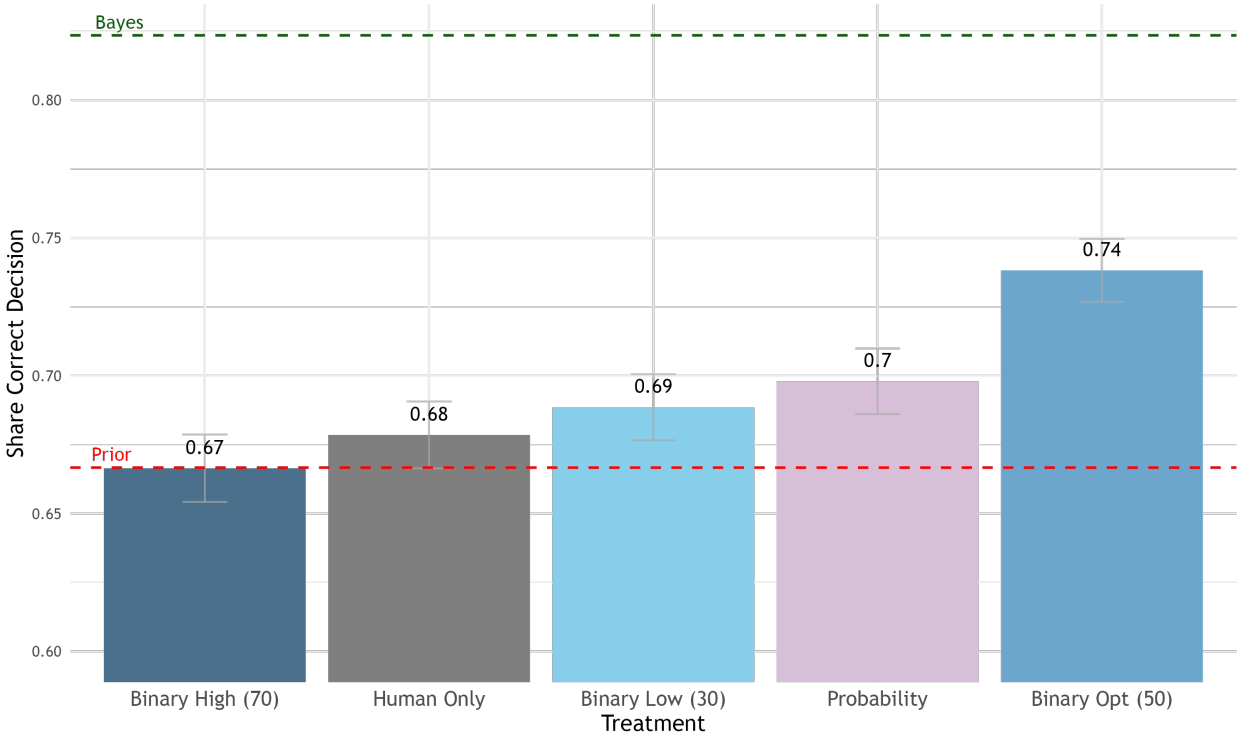


Figure 6: Decision-making outcomes under different treatment conditions

ment performed best amongst all the treatments, achieving an accuracy of 74%. Figure 6 shows that the *Binary Optimal (50)* treatment outperformed all other treatments, achieving an accuracy of 74%. This exceeds the prior of 66%—the best achievable outcome without information, obtained by approving every case (marked by the red line)—and moves closer to the Bayesian benchmark of 82%, which represents the accuracy a fully Bayesian agent would achieve with complete access to AI probability signals.

Table 2 complements this by presenting regression results from Equation 1, using the *Binary Optimal (50)* treatment as the reference level to quantify differences across treatments. 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). These results highlight the effectiveness of calibrated coarsening—closing 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.

Table 2: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Human-Only	-0.060^{***}	-0.057^{***}
	(0.017)	(0.014)
Probability	-0.040^{**}	-0.029^{**}
	(0.017)	(0.014)
Binary High (70)	-0.072^{***}	-0.058^{***}
	(0.017)	(0.014)
Binary Low (30)	-0.050^{***}	-0.040^{***}
	(0.017)	(0.014)
Constant	0.738^{***}	0.699^{***}
	(0.012)	(0.034)
Observations	7,450	7,450
R^2	0.003	0.282
Adjusted R^2	0.002	0.277
Note:	*p<0.1; **p<0.05; ***p<0.01	

Column (1) does not include fixed effects, whilst Column (2) and includes application case-level fixed effects, omitted from display due to the large number of cases.

A quick note on the other comparisons: while providing participants with the full AI score in the *Probability* condition leads to some improvement over *Human Only*, the difference is not statistically significant. Moreover, the alternative binary treatments, *Binary High (70)* and *Binary Low (30)*, yield 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 optimal threshold, as in the *Binary Optimal (50)* condition, decision-making performance improves substantially, bringing participants significantly closer to the Bayesian benchmark.

Heterogeneous treatment effects, displayed only with case fixed effects, are presented in Appendix 7.2.1. Gender emerges as a key moderator, with men benefiting significantly more from AI assistance. The *Binary Optimal (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 Optimal (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 AI assistance disproportionately benefits those with less prior domain knowledge and with more room for improvement, at least in this context, a finding supported in the literature (Dell'Acqua et al., 2023).

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 Optimal (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.

Baseline bias moderators also shape treatment effects. For **automation bias**, baseline accuracy in the *Human Only* condition is similar across groups, but treatment effects are higher for those who face less automation bias (that is, more automation neglect). This is consistent with our illustrative example of automation neglect: when the loan specialist underweights the AI signal, coarsening can help pool signals together and induce the loan specialist to listen more to the AI signal. 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 show much difference in either baseline accuracy or treatment effects.

It is important to note that all of these bias measures were elicited in abstraction from the loan context, which may explain why some, like base-rate neglect and balls-urns, show meaningful moderation of treatment effects while others, like automation bias, aggregation bias, and correlation-neglect, do not. The extent to which these biases translate into real-world decision-making likely depends on how closely the elicitation task aligns with the domain of interest.

5.3 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 are participants following the AI recommendations—and what factors are driving these outcomes.

5.3.1 Algorithmic adherence

We begin by taking a deeper dive into how participants react to AI signals. Figure 7 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. However, approval rates vary significantly across treatments, demonstrating that loan specialists are sensitive to the AI's threshold settings.

Logically, 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" recommendation weaker, leading to fewer denials. Figure 7 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 threshold conditions.

For the *Probability* condition (rightmost cluster of bars), loan approval rates are further broken out by the implied binary recommendation. For example, under the *Low (30)* threshold in the *Probability* condition, the figure displays the share of loans approved when the AI model's risk score exceeded 30. The purple bar represents the average loan approval rate across the entire *Probability* condition.

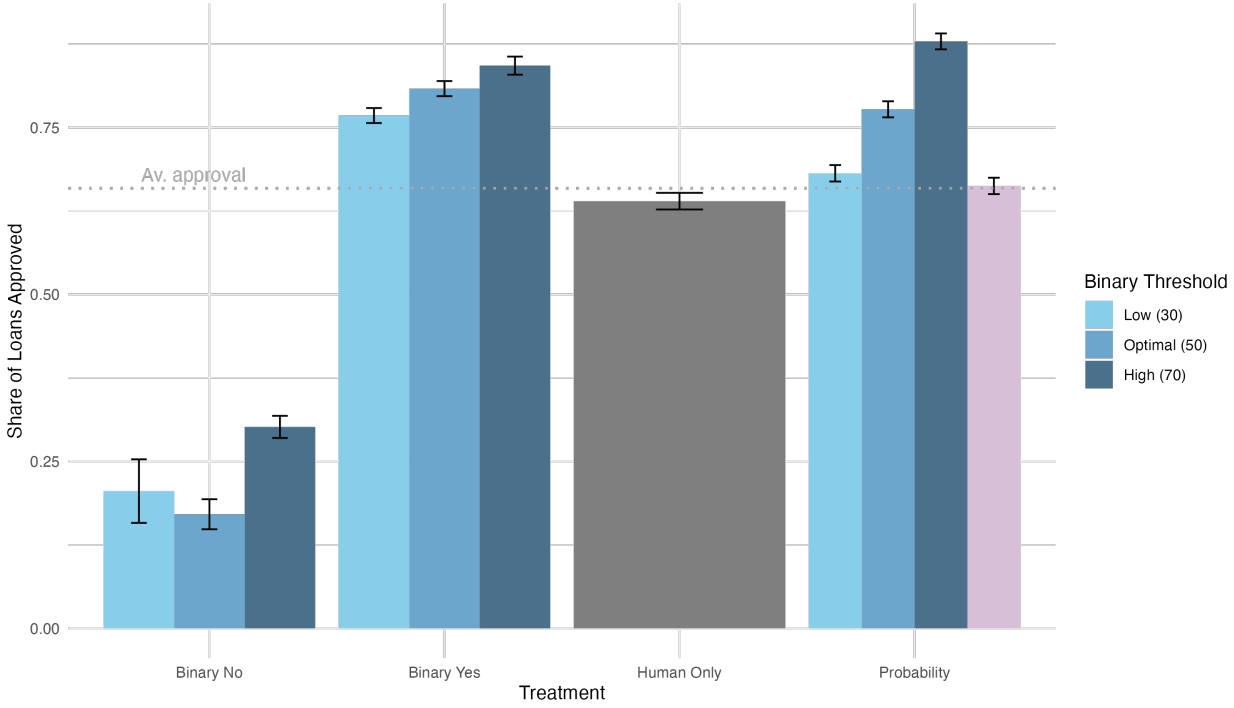


Figure 7: Loan Approval Rates, by Treatment and (implied) AI Recommendation

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 private information—rather than blindly following the AI's suggestions.

5.3.2 Estimating pooled decision boundaries

If participants are not simply following the AI signal, how are they actually making decisions? To investigate this, we recreate the decision boundary graphs detailed in the simulations from the

model section, but this time using our experimental data.

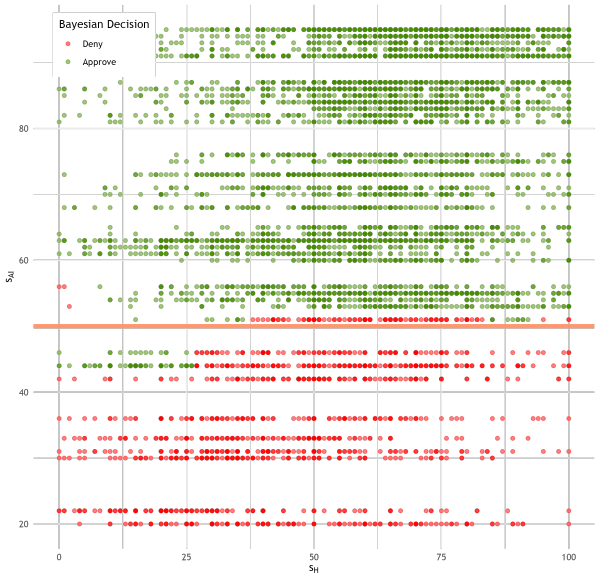
First, we verify the Bayesian decision boundary, represented by the orange line in Figure 8a. For all cases where we have elicited the human signal over the loan application (r),[12] we compute the Bayesian posterior using the joint distribution of r and s conditional on ω. We then apply the thresholding 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 the two outcomes.

As expected in this setting—where humans have no private information—the Bayesian decision boundary is a flat line around $s = 50$. This reflects the fact that a Bayesian decision-maker would perfectly follow the AI signal, approving any case where the AI score exceeds 50 and denying otherwise. This serves as a benchmark to evaluate the extent to which participant decisions deviate from this ideal behavior.

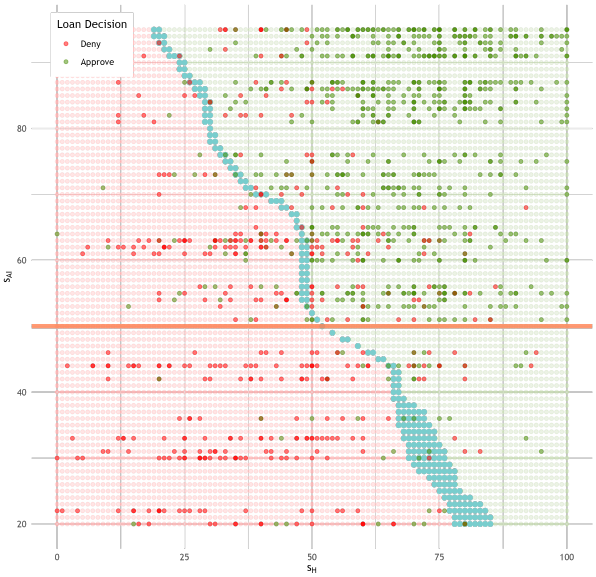
Next, we estimate the decision boundary of the loan specialists in our experiment. To do this, we pool together all decisions under the *Probability* condition (where participants were provided with the full AI risk score) in which we also elicited the human signal. This allows us to examine how loan 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 $\Pr(a^r = 1 \mid r, s) = 0.5$ (within a specified tolerance), yielding the estimated decision boundary, shown as the blue line in Figure 8b.

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

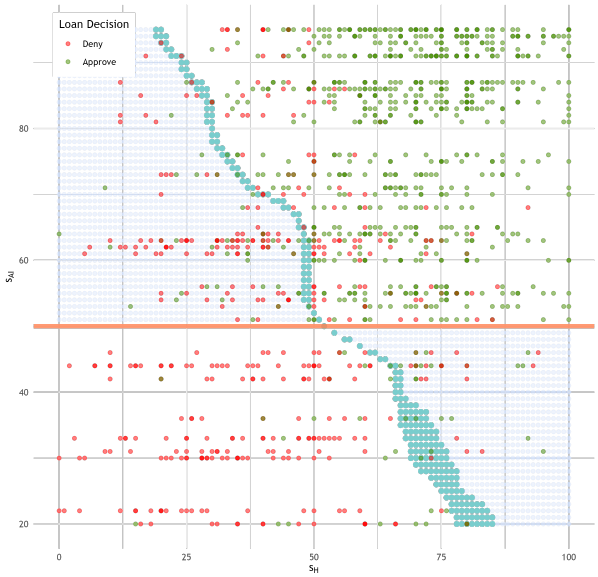
[12] Since 90% of cases present the loan application first (before the AI signal), and we independently randomize whether the posterior is elicited after the first signal in 90% of cases, this results in our observing r for 81% of cases.



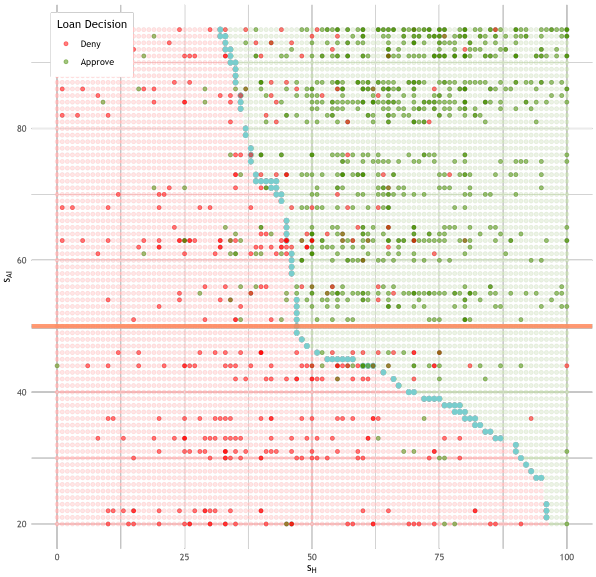
(a) Bayesian Decision Boundary
(All cases with elicited r, pooled)



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



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



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

Figure 8: Pooled Decision boundaries, Visualised Using Experiment Data

closer to Bayesian-optimal behavior.

Examining decisions under the *Binary Optimal (50)* treatment reveals that coarsening the AI signal can indeed improve decision-making outcomes. Figure 8d illustrates how thresholding the AI signal at the decision cutoff ($s = 50$) influences loan specialists' choices. In general, this coarsening helps participants adhere more closely to the AI signal within the range of $r = 40$ to $r = 90$. Outside this range, however, participants continue to rely on their own human signal, which—when strong enough—leads them to override the (sometimes implied) AI recommendation. By pooling continuous signals into actionable recommendations at key thresholds, coarsening helps align human behavior more closely with optimal decision-making benchmarks. While there remains room for improvement, particularly outside the 'adherence' range, these results underscore the potential of coarsening as a powerful tool to enhance decision-making.

5.3.3 Estimating individual decision boundaries

Given the potential heterogeneity in our experimental population, we estimate individual decision boundaries to understand how each loan specialist integrates AI-probability signals when making loan decisions. The challenge lies in the limited number of observations per specialist—on average, eight in the *Probability* condition for which we have also elicited the human signal over the loan application (r)—making traditional estimation methods unreliable. To address this, we adopt a Bayesian 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. In brief, the method specifies a prior distribution for the model parameters and an explicit model of the likelihood function, and applies Bayes' rule to derive the posterior distribution of the parameters, reflecting updated beliefs after seeing the data.

Model specification

We model loan approval as a latent probit process. Each loan application n is associated with a loan specialist i, where $i(n)$ denotes the loan specialist making the decision n. For each application, we observe two signals: the Sender's type s (the "AI" signal), and the Receiver's type r (the "human" signal). The observed binary outcome $a_n \in \{0, 1\}$ indicates whether the loan specialist decided to approve loan n ($a_n = 1$) or not ($a_n = 0$). We assume that the binary decision arises from an

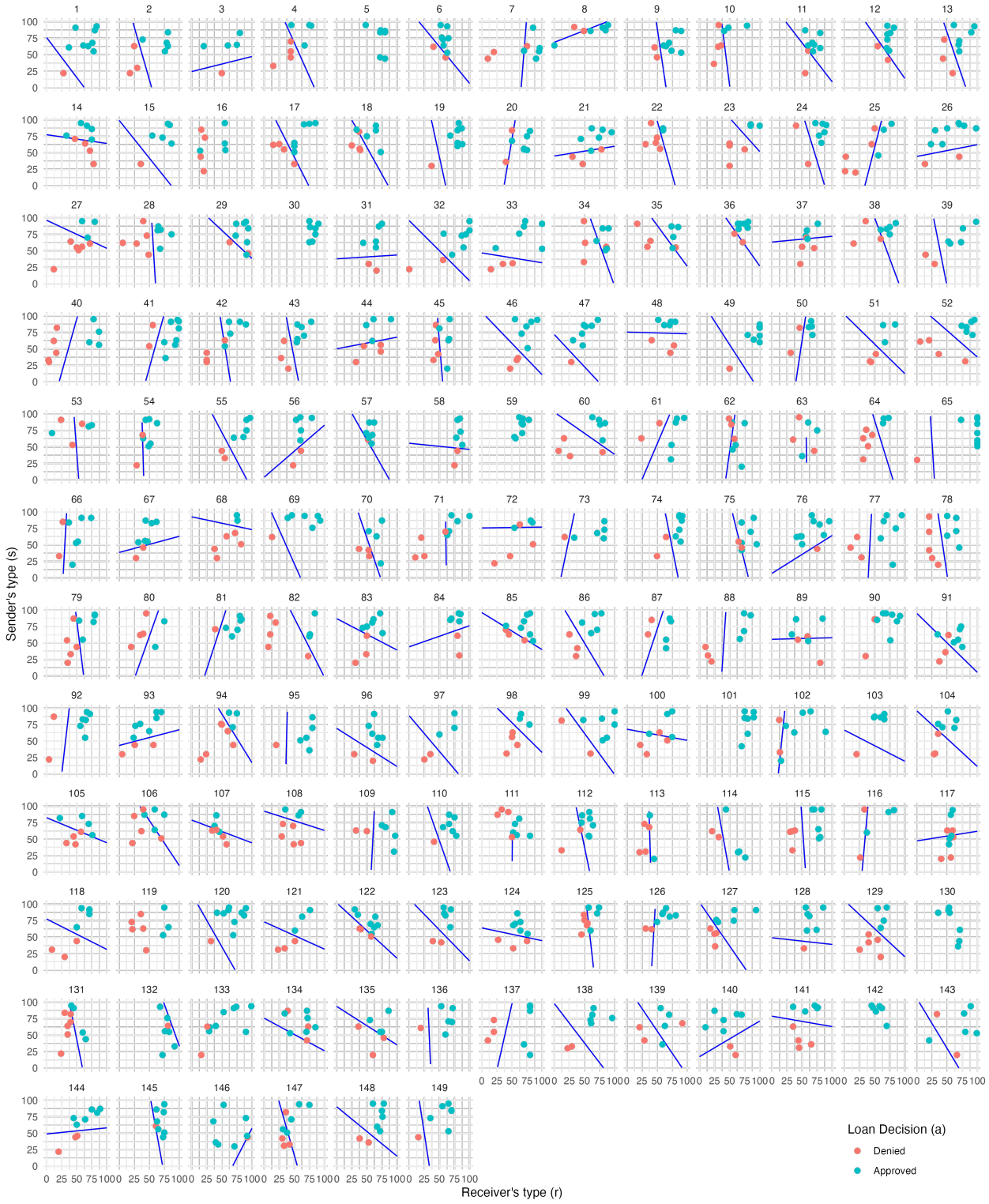


Figure 9: Loan Specialists' Individual Decision Boundaries

unobserved continuous variable y_n^*, given by:

$$y_n^* = \alpha_{i(n)} + \beta_{r,i(n)} \cdot r_n + \beta_{s,i(n)} \cdot s_n + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}(0,1)$$

where $\alpha_{i(n)}$ is the random intercept for individual i, while $\beta_{r,i(n)}, \beta_{s,i(n)}$ are the random slopes capturing how that individual weights signals r_n and s_n, respectively. We do not observe y_n^* directly but instead observe the binary loan decision outcome $a_n = \mathbb{1}[y_n^* \geq 0]$.

Since we assume that the error term follows a standard normal distribution, it follows that:

$$Pr(a_n = 1) = Pr(y_n^* \geq 0) = \Phi(\alpha_{i(n)} + \beta_{r,i(n)} \cdot r_n + \beta_{s,i(n)} \cdot s_n)$$

where $\Phi(\cdot)$ is the standard normal CDF.

To accommodate individual-level heterogeneity without overfitting, we employ a hierarchical approach with partial pooling, allowing individual-level parameters $(\alpha_i, \beta_{r,i}, \beta_{s,i})$ to vary while being informed by population-level distributions where data is scarce. Specifically, each set of individual-level effects is drawn from a normal distribution whose mean depends on a vector of participant-specific covariates Z_i (e.g., demographics, baseline biases). We let:

$$\alpha_i \sim \mathcal{N}(Z_i \cdot \gamma_\alpha, \sigma_\alpha^2)$$

$$\beta_{r,i} \sim \mathcal{N}(Z_i \cdot \gamma_r, \sigma_r^2)$$

$$\beta_{s,i} \sim \mathcal{N}(Z_i \cdot \gamma_s, \sigma_s^2)$$

where $\gamma_\alpha, \gamma_r, \gamma_s$ are the population-level coefficient vectors, and $\sigma_\alpha, \sigma_r, \sigma_s$ are the standard deviations representing how individuals vary from those means. This structure allows individuals with similar characteristics to have similar coefficients while still permitting individual deviations.

Model estimation

We use weakly informative priors, ensuring that estimates are primarily data-driven. The model is fit in Stan using Hamiltonian Monte Carlo (HMC), which jointly samples from the posterior distribution of all parameters.

We then use these draws to infer the individual decision boundaries. For each individual i, the

latent decision boundary is defined by the set of points (r, s) at which $y_i^* = 0$:

$$\alpha_i + \beta_{r,i} \cdot r + \beta_{s,i} \cdot s = 0 \implies s = -\frac{\alpha_i + \beta_{r,i} \cdot r}{\beta_{s,i}}$$

provided $\beta_{s,i} \neq 0$. Figure 9 plots these boundaries across individuals, illustrating substantial variation in how different specialists integrate the two signals. The substantial variation across individuals suggests the potential for adaptive, individual-specific interventions, which we are actively exploring.

5.3.4 Effective action thresholds

One important aspect we want to verify is whether loan specialists are behaving as if they have an effective action threshold induced by the incentives. In our setup, because false positives and false negatives are rewarded equally, loan specialists should generally act as though their posterior threshold for approval is 50. This means they should approve a loan if their posterior is above 50 and deny it otherwise. As shown in Figure 10, we observe that this is indeed the case for most loan specialists. The approve and deny decisions are generally well-separated at the vertical line passing through 50, with some degree of noise. This suggest that for most people, the deviation from the Bayesian benchmark is happening through mistakes in belief updating (to the posterior), and less so from the translation from posterior beliefs to actions.

5.3.5 Where loan specialists are performing well

A related question, then, is where people are doing well at approving as well as denying. Table 3-4 present the shares of true and false positives and negatives across treatments. These results are broken down by treatment arm (the main treatment arms: *Binary Optimal*, *Human Only*, and *Probability*) and further categorised by buckets of AI risk scores and human-elicited signals (where available). The tables highlight the areas where participants perform well in both approving and denying decisions, offering insight into how the treatments influence decision-making accuracy across different levels of signal strength.

As seen in Table 3, regardless of treatment, loan specialists are much better at approving correctly[13] than denying correctly,[14] despite the fact that they are incentivised equally for true posi-

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

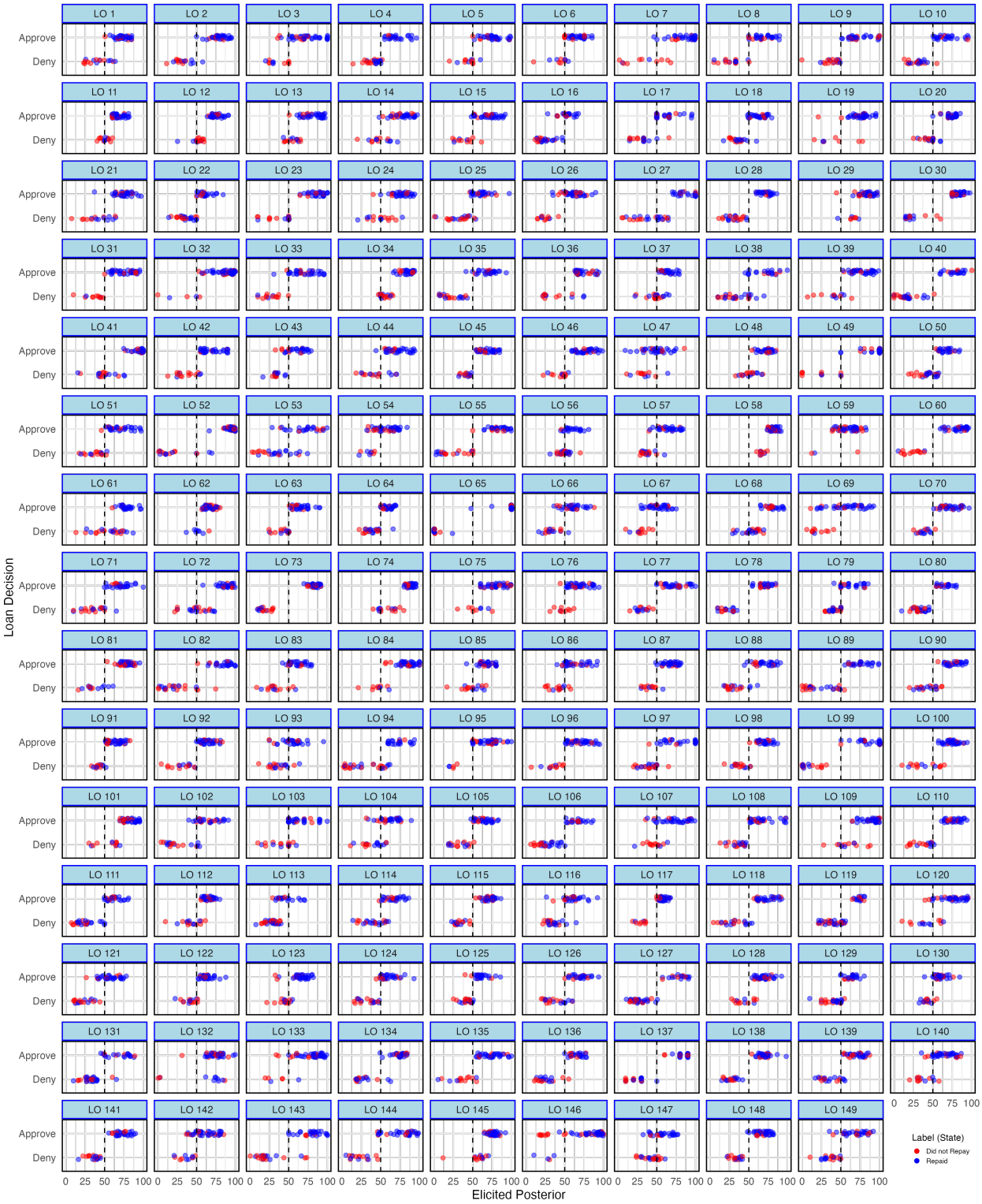


Figure 10: Loan Specialists' Effective Action Thresholds

Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)
Binary Opt (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

Table 3: Share of True/False Positives and Negatives by Treatment

tives 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 the share of 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.

Heterogeneous performance by baseline moderators are additionally presented in Appendix 7.2.1. Any discrepancies between the coefficients in the regression tables (e.g., Table 6) and the "Overall Accuracy" columns of these true/false positive and negative summary tables (e.g., Table 7) arise because the former incorporate case fixed effects, while the latter report simple means.

r	Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)
0-33	Binary Opt (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 Opt (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 Opt (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

Table 4: Share of True/False Positives and Negatives by s_H Bucket and Treatment

Breaking this down further by the human signal inferred from observing the loan application (for cases where the "human only" posterior was elicited), Table 4 shows that loan specialists

perform worst when their own signal is low (0-33), i.e., in the implied "Deny" region, and the AI treatments do not appear to have positive treatment effects on accuracy. Specialists perform best when their own signal is high (67-100). The *Binary Optimal* treatment offers the greatest improvement (6.6pp over *Probability* and 8.0pp over *Human Only*) when the human signal falls in the uncertain range (34-66), suggesting that this treatment is particularly effective in aiding decision-making under uncertainty by reducing the share of cases falsely denied when participants are uncertain.

s	Treatment	TP (% of total)	TN (% of total)	FP (% of total)	FN (% of total)	Overall Accuracy (%)
0-33	Binary Opt (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 Opt (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 Opt (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

Table 5: Share of True/False Positives and Negatives by s_{AI} Bucket and Treatment

We can alternatively break this out by the underlying AI risk score of each loan case, as seen in Table 5. 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.[15] The *Binary Optimal* treatment shows the greatest improvement over Probability when the AI score falls in the ambiguous range (34-66). This implies that 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.

[15]Note, of course, they don't see this signal in the *Human Only* condition, and only a coarsened version of the signal in the *Binary Optimal* .

6 Conclusion and Discussion

To fully leverage the potential of AI to improve decision-making and transform the future of work, the optimised design of human-AI collaboration is paramount. In this paper, we show—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.

Crucially, our study focuses on real loan approval scenarios, leveraging ground truth data to demonstrate that our approach improves outcomes we care about in practice, such as loan approval accuracy and reduced default rates. A common challenge in such contexts is the selective labels problem, where the outcome of interest is available only for a subset of the population. Specifically, ground truth is typically observed only for individuals approved for loans, as they were deemed low-risk by loan officers in the original Home Credit setting. Our AI model is trained on this subset (with repayment as the outcome). However, this does not pose a significant issue in our experiment as participants evaluate a randomly sampled subset of cases from the Home Credit dataset,[16] ensuring the applicant distribution matches that of the AI training data. This setup mitigates traditional selective label concerns.

Instead, the primary consideration is of external validity: whether this applicant pool differs meaningfully from contexts in which calibrated coarsening might be deployed, and whether such differences could in turn interact with the treatments to skew results. For this issue to materially affect external validity, the potential difference between the population of the AI training data and actual pool of applicants that loan officers face would have to systematically alter how participants respond to a full probability AI risk-score, compared to a coarsened AI signal. Whilst this is possible, we do not believe this to be a large risk. Moreover, firms deploying calibrated coarsening in real-world contexts often have access to more comprehensive datasets, including data on applicants for whom the repayment outcome is unobserved. With these datasets, selective label issues could be mitigated more directly through methods such as reweighting or other causal inference techniques. By leveraging the full range of data available and tailoring the approach to their spe-

[16] Of course, these cases are excluded from the training of the AI model.

cific applicant pools, firms can ensure that calibrated coarsening remains a practical, robust, and effective decision-support tool in their unique operational contexts.

A related and broader question, then, is on 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 developing adaptive thresholds to account for heterogeneity in bias types. Additionally, we are exploring extensions to other domains, including hiring decisions and radiology diagnostics, to assess the broader applicability of calibrated coarsening in different decision-making contexts.

On the theoretical front, we are working to demonstrate that under mild conditions regarding deviation from Bayesian updating and the information structure, optimal coarsened messaging always benefits the Sender compared to either no signal or perfect signal revelation. This approach does not require knowledge of the Receiver's private signals or assume specific behavioral biases. Instead, it uses the joint distribution of historical Receiver decisions with the state and AI signal.

In conclusion, calibrated coarsening is a practical and effective approach to improving human-AI decision-making. Future research avenues include scenarios involving strategic considerations, objective function may diverge from that of the Sender (e.g., management, AI designer, or the AI itself) 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. 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.

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.

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.

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).

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.

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.

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 Cambridge, MA, USA.

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

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

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

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

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

Enke, B. and Zimmermann, F. (2019). Correlation neglect in belief formation. *Review of Economic Studies*, 86:313–332. © The Author(s) 2017. Published by Oxford University Press on behalf of The Review of Economic Studies Limited. Advance access publication 22 December 2017.

European Commission (2024). Eu artificial intelligence act: Article 14. https://artificialintelligenceact.eu/ai-act-explorer/.

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.

Fogliato, R., De-Arteaga, M., and Chouldechova, A. (2022). A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. *Available at SSRN*.

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.

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

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

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

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

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. (2022). The impact of algorithmic tools on child protection: Evidence from a randomized controlled trial. *Job market paper*.

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

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

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

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

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

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.

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. 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.

National Academies of Sciences, E., Medicine, et al. (2023). Behavioral economics: Policy impact and future directions.

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*.

Quah, J. K.-H. and Strulovici, B. (2012). Aggregating the single crossing property. *Econometrica*, 80(5):2333–2348.

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., 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.

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*.

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.

7 Appendix

7.1 Bias Elicitations

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

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

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

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

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

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

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.

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.

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

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

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

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

> Three people: Ann, Bob, and Charlie. Each of them is interested in estimating the weight of a water bucket in pounds.
>
> Ann and Bob both get to take a peek at the bucket. They are equally good at estimating. Each tends to get their weight estimates right on average, but sometimes they make random mistakes. Ann and Bob are equally likely to make mistakes in any given estimate they make.
>
> Ann and Bob both share their estimates with Charlie, who never sees the bucket. 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.
>
> We will pay you more points the closer your decision is to the statistically-correct estimate, given the information provided.
>
> - 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.
>
> **What is your best estimate of the weight of the bucket?** (Round to the nearest integer.)

Balls and Urns We utilize a question from the balls-and-urns belief updating task in Enke et al. (2023). In this task, subjects are asked to indicate the percentage chance that the selected bag

is the one that contains more red chips, given that the drawn chip is red. Often, subjects exhibit conservatism bias by providing posterior estimates strictly between 50% and 70%, even though the Bayesian answer is 70%. We define the belief updating parameter BU as the ratio of the subject's response to 70:

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

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

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

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.

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?**

7.2 Results

7.2.1 Heterogeneous treatment effects

Age

Table 6: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Age (Below Median)	Age (Above Median)
	(1)	(2)
Human-Only	-0.056^{***}	-0.058^{***}
	(0.020)	(0.021)
Probability	-0.031	-0.031
	(0.020)	(0.021)
Binary High (70)	-0.064^{***}	-0.049^{**}
	(0.020)	(0.021)
Binary Low (30)	-0.020	-0.062^{***}
	(0.020)	(0.021)
Constant	0.716^{***}	0.679^{***}
	(0.047)	(0.048)
Observations	3,950	3,500
R^2	0.279	0.300
Adjusted R^2	0.269	0.289
Note:		*p<0.1; **p<0.05; ***p<0.01

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% of total)	Overall Accuracy (%)
Age (Above Median)					
Probability	53.20	17.79	14.95	14.06	71.00
Human Only	49.14	19.14	16.29	15.43	68.29
Binary Opt (50)	59.47	18.41	11.15	10.97	77.88
Age (Below Median)					
Probability	48.82	20.06	14.85	16.27	68.88
Human Only	48.10	19.37	14.56	17.97	67.47
Binary Opt (50)	53.11	19.25	15.22	12.42	72.36

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

Gender

Table 8: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Male	Female
	(1)	(2)
Human-Only	-0.095^{***}	-0.026
	(0.022)	(0.019)
Probability	-0.048^{**}	-0.014
	(0.022)	(0.019)
Binary High (70)	-0.078^{***}	-0.040^{**}
	(0.022)	(0.019)
Binary Low (30)	-0.054^{**}	-0.029
	(0.022)	(0.019)
Constant	0.747^{***}	0.663^{***}
	(0.052)	(0.044)
Observations	3,100	4,350
R^2	0.304	0.275
Adjusted R^2	0.291	0.266
Note:	*p<0.1; **p<0.05; ***p<0.01	

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

Table 9: Share of True/False Positives and Negatives by Treatment and Gender

Years of Loan Specialist Experience

Table 10: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Experience (Below Median)	Experience (Above Median)
	(1)	(2)
Human-Only	-0.072^{***}	-0.037^{*}
	(0.020)	(0.021)
Probability	-0.040^{**}	-0.021
	(0.020)	(0.021)
Binary High (70)	-0.089^{***}	-0.019
	(0.020)	(0.021)
Binary Low (30)	-0.058^{***}	-0.021
	(0.020)	(0.021)
Constant	0.680^{***}	0.722^{***}
	(0.046)	(0.049)
Observations	4,200	3,250
R^2	0.277	0.318
Adjusted R^2	0.268	0.307
Note:		*p<0.1; **p<0.05; ***p<0.01

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

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

Perceived Private Information

Table 12: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Perceived (Below Median)	Perceived (Above Median)
	(1)	(2)
Human-Only	-0.074^{***}	-0.042^{**}
	(0.020)	(0.021)
Probability	-0.038^{*}	-0.020
	(0.020)	(0.021)
Binary High (70)	-0.053^{***}	-0.062^{***}
	(0.020)	(0.021)
Binary Low (30)	-0.044^{**}	-0.036^{*}
	(0.020)	(0.021)
Constant	0.698^{***}	0.700^{***}
	(0.047)	(0.049)
Observations	3,800	3,650
R^2	0.293	0.282
Adjusted R^2	0.283	0.271
Note:		*p<0.1; **p<0.05; ***p<0.01

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% of total)	Overall Accuracy (%)
Perceived (Above Median)					
Probability	51.45	18.40	14.65	15.50	69.85
Human Only	48.22	19.73	15.48	16.58	67.95
Binary Opt (50)	55.14	18.38	14.17	12.31	73.52
Perceived (Below Median)					
Probability	50.33	19.57	15.13	14.97	69.90
Human Only	48.95	18.82	15.26	16.97	67.76
Binary Opt (50)	56.98	19.32	12.50	11.20	76.30

Table 13: Share of True/False Positives and Negatives by Treatment and Perceived Private Info

Aggregation bias

Table 14: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Rubinstein (Below Median)	Rubinstein (Above Median)
	(1)	(2)
Human-Only	-0.060^{***}	-0.051^{**}
	(0.018)	(0.026)
Probability	-0.038^{**}	-0.011
	(0.018)	(0.025)
Binary High (70)	-0.072^{***}	-0.029
	(0.018)	(0.025)
Binary Low (30)	-0.057^{***}	-0.006
	(0.018)	(0.025)
Constant	0.702^{***}	0.696^{***}
	(0.041)	(0.059)
Observations	5,050	2,400
R^2	0.279	0.302
Adjusted R^2	0.271	0.286
Note:		*p<0.1; **p<0.05; ***p<0.01

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

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

Balls and Urns

Table 16: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Balls-Urns (Below Median)	Balls-Urns (Above Median)
	(1)	(2)
Human-Only	-0.064^{***}	-0.027
	(0.019)	(0.028)
Probability	-0.027	-0.002
	(0.019)	(0.028)
Binary High (70)	-0.074^{***}	-0.031
	(0.019)	(0.028)
Binary Low (30)	-0.038^{**}	-0.026
	(0.019)	(0.028)
Constant	0.691^{***}	0.782^{***}
	(0.044)	(0.066)
Observations	4,300	1,900
R^2	0.288	0.310
Adjusted R^2	0.279	0.289
Note:		*p<0.1; **p<0.05; ***p<0.01

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% of total)	Overall Accuracy (%)
Balls Urns (Below Median)					
Probability	53.17	17.97	13.99	14.87	71.13
Human Only	47.44	19.30	15.93	17.33	66.74
Binary Opt (50)	56.01	18.25	14.00	11.74	74.26
Balls Urns (Above Median)					
Probability	49.03	22.08	14.94	13.96	71.10
Human Only	50.79	19.21	13.16	16.84	70.00
Binary Opt (50)	57.81	19.27	10.63	12.29	77.08

Table 17: Share of True/False Positives and Negatives by Treatment and Balls Urns

Automation Bias

Table 18: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Automation Bias (Below Median)	Automation Bias (Above Median)
	(1)	(2)
Human-Only	-0.067^{***}	-0.022
	(0.017)	(0.030)
Probability	-0.044^{***}	0.027
	(0.017)	(0.030)
Binary High (70)	-0.072^{***}	-0.021
	(0.017)	(0.030)
Binary Low (30)	-0.040^{**}	-0.037
	(0.017)	(0.030)
Constant	0.706^{***}	0.668^{***}
	(0.038)	(0.070)
Observations	5,650	1,800
R^2	0.290	0.289
Adjusted R^2	0.284	0.267
Note:		*p<0.1; **p<0.05; ***p<0.01

Treatment	TP	TN	FP	FN	Overall.Accuracy
Automation Bias (Above Median)					
Probability	52.25	21.80	13.49	12.46	74.05
Human Only	50.83	16.67	17.50	15.00	67.50
Binary Opt (50)	55.56	18.18	14.14	12.12	73.74
Automation Bias (Below Median)					
Probability	50.44	18.10	15.34	16.11	68.54
Human Only	47.88	20.09	14.69	17.35	67.96
Binary Opt (50)	56.25	19.08	13.05	11.62	75.33

Table 19: Share of True/False Positives and Negatives by Treatment and Automation Bias

Base-rate Neglect

Table 20: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Base-Rate Neglect (Below Median)	Base-Rate Neglect (Above Median)
	(1)	(2)
Human-Only	-0.052^{***}	-0.081^{***}
	(0.018)	(0.030)
Probability	-0.044^{**}	-0.001
	(0.018)	(0.030)
Binary High (70)	-0.071^{***}	-0.038
	(0.018)	(0.030)
Binary Low (30)	-0.028	-0.064^{**}
	(0.018)	(0.030)
Constant	0.706^{***}	0.674^{***}
	(0.042)	(0.069)
Observations	4,800	1,850
R^2	0.291	0.274
Adjusted R^2	0.282	0.252
Note:		*p<0.1; **p<0.05; ***p<0.01

Treatment	TN (% of total)	TP (% of total)	FN (% of total)	FP (% of total)	Overall Accuracy (%)
BR Neglect (Below Median)					
Probability	50.13	17.79	15.84	16.23	67.92
Human Only	49.79	18.96	14.58	16.67	68.75
Binary Opt (50)	55.10	19.77	12.88	12.24	74.87
BR Neglect (Above Median)					
Probability	54.39	21.28	11.15	13.18	75.68
Human Only	45.14	18.92	18.38	17.57	64.05
Binary Opt (50)	56.19	18.39	13.04	12.37	74.58

Table 21: Share of True/False Positives and Negatives by Treatment and Base Rate Neglect

Correlation Neglect

Table 22: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	Correlation Neglect (Below Median)	Correlation Neglect (Above Median)
	(1)	(2)
Human-Only	-0.068^{***}	-0.060^{**}
	(0.018)	(0.027)
Probability	-0.026	-0.052^{*}
	(0.018)	(0.027)
Binary High (70)	-0.064^{***}	-0.051^{*}
	(0.018)	(0.027)
Binary Low (30)	-0.045^{**}	-0.046^{*}
	(0.018)	(0.027)
Constant	0.697^{***}	0.714^{***}
	(0.042)	(0.063)
Observations	4,750	2,100
R^2	0.293	0.296
Adjusted R^2	0.285	0.278
Note:		*p<0.1; **p<0.05; ***p<0.01

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

Table 23: Share of True/False Positives and Negatives by Treatment and Correlation Neglect

7.3 Robustness checks

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

Table 24 and Figure 11 present the results shown in Table 2 and Figure 6, respectively, restricting data to those recruited over LinkedIn (NAMU/NAMP inclusive) and Indeed, representing around 65% of our full sample. As seen below, the results still hold: *Binary Optimal* performs better than both *Probability* and *Human Only*. These differences are significant for Binary Optimal vs. *Human Only*, but due to the smaller sample size are not significant when compared to *Probability*. This consistency in results across recruitment sources increases our confidence that the findings hold for loan specialists targeted through our outreach efforts and are not driven by Prolific participants.

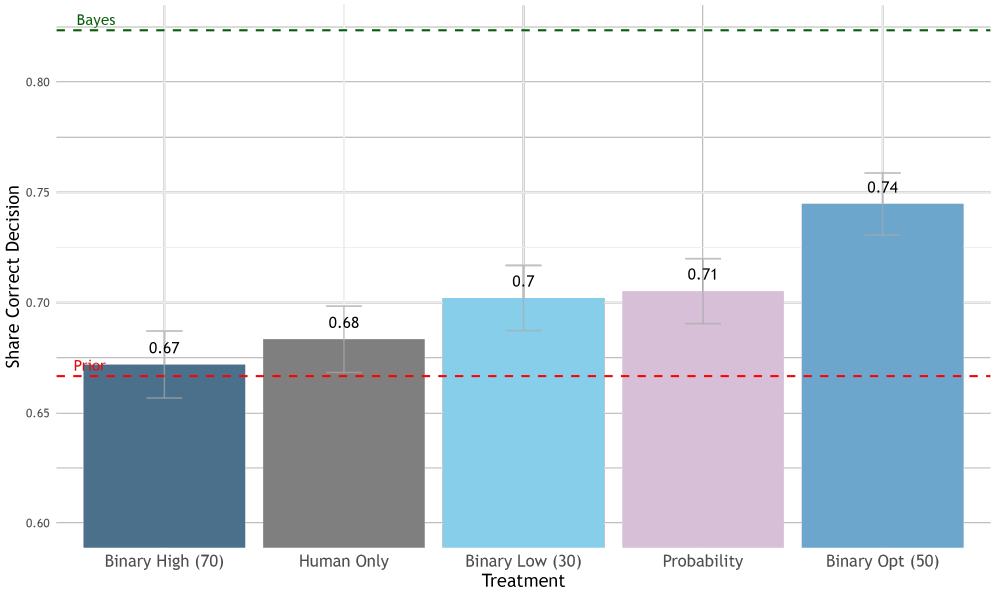


Figure 11: Decision-making outcomes under different treatment conditions

Table 24: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Human-Only	-0.061^{***}	-0.056^{***}
	(0.021)	(0.018)
Probability	-0.040^{*}	-0.018
	(0.021)	(0.018)
Binary High (70)	-0.073^{***}	-0.043^{**}
	(0.021)	(0.018)
Binary Low (30)	-0.043^{**}	-0.020
	(0.021)	(0.018)
Constant	0.745^{***}	0.645^{***}
	(0.015)	(0.041)
Observations	4,800	4,800
R^2	0.003	0.297
Adjusted R^2	0.002	0.289
Note:	*p<0.1; **p<0.05; ***p<0.01	

7.3.2 Restricting to first round only

Table 25 and Figure 12 present the results shown in Table 2 and Figure 6, respectively, restricting data to those obtained in the first decision block only for all participants.

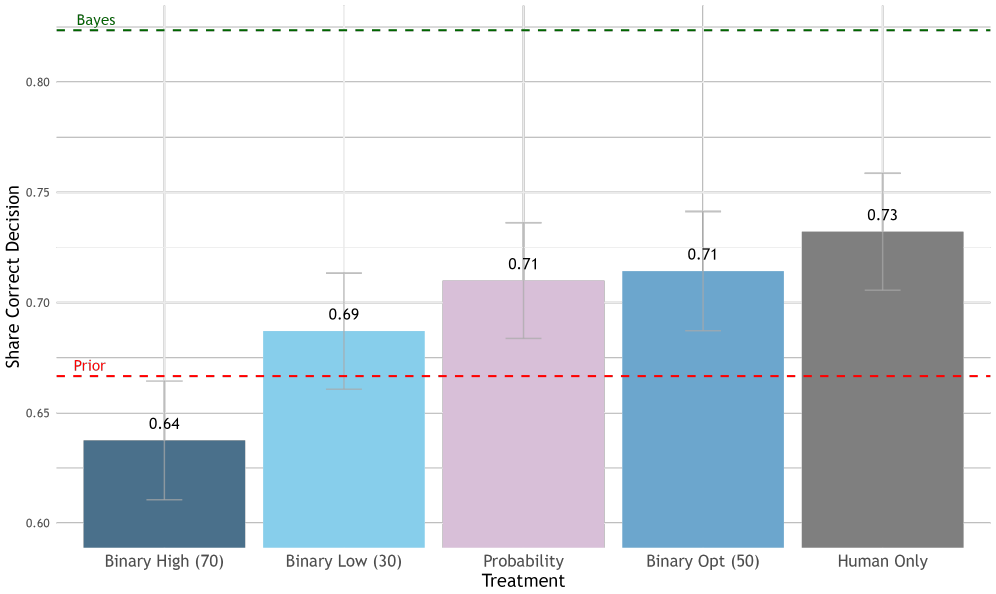


Figure 12: Decision-making outcomes under different treatment conditions

Table 25: Decision-making outcomes under different treatment conditions

	Dependent variable:	
	Share of Decisions Correct	
	(1)	(2)
Human-Only	0.018	0.021
	(0.039)	(0.034)
Probability	−0.004	0.040
	(0.038)	(0.033)
Binary High (70)	−0.077**	−0.037
	(0.038)	(0.032)
Binary Low (30)	−0.027	−0.008
	(0.038)	(0.033)
Constant	0.714***	0.672***
	(0.027)	(0.075)
Observations	1,490	1,490
R^2	0.005	0.301
Adjusted R^2	0.002	0.275
Note:	*p<0.1; **p<0.05; ***p<0.01	

Column (1) does not include fixed effects, whilst Column (2) and includes application case-level fixed effects, omitted from display due to the large number of cases.

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

Recall that we implemented case-level randomisation of:

- **Order of information presentation:** Whether the AI signal or loan application shown first {AI (10%), Loan (90%)}
- **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 to show that they are not influenced by the order in which participants see the AI signal, or whether or not a "posterior" was elicited in between the two signals. As shown in Table 26,

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

Table 26: Robustness Checks

	Dependent variable:			
	Loan Decision	Posterior Belief	Loan Decision	Posterior Belief
	(1)	(2)	(3)	(4)
Elicitation Order			0.001	−0.424
			(0.018)	(0.861)
First Posterior	−0.023	−1.081		
	(0.019)	(0.898)		
Constant	0.680***	60.454***	0.658***	59.855***
	(0.018)	(0.855)	(0.017)	(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:				*p<0.1; **p<0.05; ***p<0.01