Preference for Explanations: Case of Explainable AI

Alex Chan

Harvard University

(Please click here for the most current version)

October 16, 2025*

Abstract

I study how individuals use, avoid, or undervalue explanations of artificial intelligence in two experiments involving real US$10,000-loans. In the first, 45% of loan-approving participants sought explanations of an AI's credit-risk predictions. When explanations revealed that borrowers received worse risk scores partly due to race and gender, participants were 42% more likely to override profit-maximizing recommendations. Yet lender-aligned participants were 19% more likely to avoid explanations—even as they were more likely to review the AI's risk predictions—consistent with strategic information avoidance. This avoidance disappeared when explanations were framed as purely financial or when demographics were absent, and third parties sanctioned deliberate avoidance at personal cost. A second experiment shows systematic undervaluation: willingness-to-pay for explanations fell 26% when participants held private information that made explanations more valuable, reflecting failures of contingent reasoning. A simple tutorial reversed this effect. These results show that market demand understates the social value of AI explainability.†

Keywords: XAI, Interpretability, explanation, algorithms, artificial intelligence, information avoidance, moral wiggle room, algorithmic fairness, consumer finance

*Email address: achan@hbs.edu. This study was approved by the Institutional Review Board at Harvard University (IRB-25-0040 and IRB-25-1120). The study was pre-registered on the AEA RCT registry, ID AEARCTR-0015581 and AEARCTR-0016777. Acknowledgements to the HBS Dean's Office, Harvard Office of the Provost, and Harvard Office of the General Counsel for the thorough review process. Valuable research assistance was provided by Yetong Xu.

†*JEL* Codes: B4, C1, C91, C92, D81, G41, D12, D14, G21

1 Introduction

Explanations serve both individual and social good, though they are too often avoided, strategically or mistakenly.

AI explainability (the ability to explain "how" a prediction was made)[1] has surged in importance as a research topic. Over the past decade, there were over 5,000 publications[2] on the topic (Jacovi (2023)) and the number of papers containing "explainable AI", "XAI", "explainability", "interpretable AI", or "interpretability" has more than tripled since 2010 (Graziani et al. (2023)). And regulators are increasingly mandating AI explainability: in the U.S., the Equal Credit Opportunity Act requires creditors to provide specific reasons for adverse loan decisions (Consumer Financial Protection Bureau (2011)); and in the E.U., the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act establish rights to explanation for significant algorithmic decisions (European Parliament and Council of the European Union (2016); European Parliament and Council of the European Union (2023)).

Given the "unreasonable effectiveness" of AI algorithms (Ludwig et al. (2024)), demand for algorithmic predictions are ever on the rise, even though many AI systems are "Black-Box" and predictions cannot be presented in readily explainable terms. AI explainability can be used to confirm important desiderata (e.g., algorithmic fairness; see Doshi-Velez and Kim (2017), Kleinberg and Mullainathan (2019)) of AI systems already deployed in important decisions like hiring (Chalfin et al. (2016)), credit approval (Fuster et al. (2022), medical testing (Mullainathan and Obermeyer (2019), judicial decisions (Kleinberg et al. (2018)), and more. A staggering 85% of businesses believe that transparency in AI—how models work—is a critical factor in winning over consumers (IBM Corporation and Morning Consult (2022)), and economists have long been aware that information is a valuable asset that decision-makers seek (Stigler (1961)).

Much of the existing literature on AI explainability focuses on developing methods to generate and evaluate explanations, often presuming that individuals naturally demand them. If this assumption holds, consumers and firms should willingly trade off between explainability and other attributes, such as predictive accuracy, and regulation would

[1] Sometimes also called interpretability. Some scholars make a distinction between explainability and interpretability but I will use the words interchangeably here. See, e.g., Molnar (2020).

[2] For comparison, the National Bureau of Economic Research publishes roughly 1,200 working papers each year across all fields of economics.

only be necessary where competitive pressures fail to elicit adequate disclosure. Yet it is also possible that individuals may avoid explanations—for example, to enable self-serving or anti-social behavior they would otherwise forgo—or that behavioral biases lead them to undervalue explanations. If so, the scope for regulation is broader than the standard efficiency rationale would imply. Whether these latter concerns arise in common economic settings involving algorithmic decision-making remains an empirical question.

When self-interest and fairness come into conflict, the demand for AI explainability can diverge from the demand for AI predictions. A bank may wish to minimize credit risk, but disparities across demographic groups can disproportionately restrict access to loans (Bakker et al. (2025)). Self-image concerns may also lead individuals to avoid information instrumental to their choices if such avoidance allows them to "excuse selfishness" (Exley and Kessler (2023), Exley (2016), Dana et al. (2007)) or to distort beliefs in self-serving ways (Saccardo and Serra-Garcia (2023)). Credit markets illustrate both the promise and the limits of explainability. Banks often rely on vendor-provided credit scoring models, where features, weights, and preprocessing steps are not fully disclosed either to lenders or to borrowers. Current regulation emphasizes outcome fairness—such as nondiscriminatory denial rates—but generally shields lenders from liability so long as they use industry-standard scores. Responsibility for uncovering and addressing potential bias is therefore diffuse, and lenders' proactive interest in opening the black box must serve as the counterweight to sustain fairness.

Even when fairness is not the central issue, explanations may still be undervalued. They often serve as complements to private information, helping users interpret predictions more effectively. For example, if a model's prediction hinges on a feature that a private signal renders uninformative, knowing this linkage is essential for correct inference. Yet individuals may treat explanations as redundant rather than complementary when they already hold private signals. This reasoning gap, related to failures in contingent reasoning (Shafir and Tversky (1992), Martínez-Marquina et al. (2019), Niederle et al. (2013), Luttmer et al. (2025), Esponda and Vespa (2014)), helps explain why willingness-to-pay for explanations remains low, as people fail to fully appreciate their usefulness.[3]

This paper is the first to provide direct empirical evidence on when and why explana-

[3]There is a related literature that looks at how to structure AI outputs to improve human decision-making; for example, see Hoong and Dreyfuss (2025) on calibrated coarsening in loan approvals.

tions are under-demanded in high-stakes AI-supported decisions. I study loan allocation decisions involving real $10,000-loans in collaboration with a private U.S. lender. Two mechanisms emerge. First, I show strategic avoidance of explanations: lender-aligned participants were roughly one fifth more likely than neutrals to avoid viewing explanations that could reveal discriminatory implications of their profit-maximizing choices, even while actively seeking predictions. Second, I show systematic undervaluation of explanations: in a prediction task, willingness-to-pay fell by over 25% when participants held private information that made explanations more useful, but rose by nearly 24% after a brief tutorial clarified their complementarity with private signals. These findings extend the literature beyond the consumption or persuasion value of explanations (e.g., Graeber et al. (2024)) by demonstrating their distinctive role in helping users integrate AI predictions with their private information—and by being the first to show that individuals often fail to recognize this value ex ante.

From a policy perspective, both strategic avoidance and systematic undervaluation matter for regulation. Decision-makers may shun explainable AI to preserve moral wiggle room, leading to inefficiently low demand in settings such as consumer finance, where incentives align with ignoring fairness concerns. At the same time, individuals may undervalue explanations ex ante because they fail to anticipate their complementarity with predictions—even though once shown explanations ex post they can use them to improve decisions. Correcting this failure of contingent reasoning justifies policies that nudge, subsidize, or otherwise reduce the cost of explanation. Importantly, the technical literature on explainability typically emphasizes the accuracy–explainability trade-off: full transparency is infeasible, so explanations are designed to maximize understanding within cognitive limits (e.g., Lage et al. (2019)). But if demand is also shaped by avoidance and undervaluation, a purely technical assessment will misestimate the true social value of explainability, underscoring the need to reconsider the welfare implications of explainability regulation.

In the first experiment, I examine how financial incentives shape engagement with AI-generated information in loan decisions in collaboration with a private US-based lender. In a survey experiment, participants decide whether to approve real US$10,000-loans for two borrowers: a binary decision to accept the AI's recommendation and only approve the low-risk borrower or override the AI's recommendation and approve both, with the

AI assessing default risk. Participants are made aware that the borrowers differ by race (White/Black) and gender (Male/Female). Participants are randomized into a *neutral treatment* (fixed bonus) or a *lender-aligned treatment* (bonus tied to repayment) and choose whether to view the AI's risk predictions and/or its explanation. To test for self-image concerns and the tension between self-interest and fairness, I vary the salience of the possibility of the inclusion of race and gender as explanatory variables in AI explainability. *Lender-aligned* participants are more likely to avoid explanations but more likely to seek risk scores and AI recommendations. Difference in avoidance of explanation is amplified when potential for fairness auditing is more salient. To my knowledge, this is the first study to document simultaneous information-seeking and avoidance while attempting to isolate image concerns from alternative explanations like aversion of interpersonal trade-offs and confusion/inattention.[4]

Avoiding explainability is particularly intriguing when the explanation serves an instrumental purpose beyond simply providing moral wiggle room. Does understanding how an AI assigns default risk impact decision-making? More specifically, does it change how individuals allocate loans? Many participants seem to believe so—between $45\% - 56\%$ chose to view the explanation across both the *neutral* and *lender-aligned* treatments. Moreover, among those who either opted to see an explanation or were randomized into receiving one, those given a detailed breakdown of key predictive factors (including race/gender)—leveraging the widely used Shapley-Additive-Explanation-Values (SHAP) method—were different in how likely they override the AI from those who received only a generic AI description (akin to marketing claims about AI capabilities). The latter were more likely to just accept the AI's recommendation and approve only the loan to the lower-risk White male borrower, compared to others. This is not because due to shifts in any difference between posteriors about prediction accuracy. This suggests that explainability is not just a passive feature of AI models—it has real consequences for economic decision-making.

If explanations are meant to improve loan repayment, do participants perceive them as somehow useful ex ante? To test this, we varied how the option to receive an explanation was presented. In one framing, participants were told the explanation would highlight only the role of financial information in the model's prediction; in another, they were also

[4]For more on the rich literature on information avoidance, Golman et al. (2017) offer a great review.

told it would shed light on the role of race and gender (as well as financial information). The results are striking: when race and gender were also salient, *lender-aligned* participants avoided the explanation at higher rates. But when the explanation was framed only around financial variables, *lender-aligned* participants similarly sought out the information. This pattern suggests that explainability is selectively valued—participants embrace it when it seems likely to align with their incentives (and only avoid it when it could expose uncomfortable truths).

I conducted a *demographic-free* borrower variant of the experiment, where participants were never made aware of the borrowers' race or gender. Instead, these demographic differences were replaced with arbitrary distinctions (e.g., odd/even-numbered phone numbers). In this setting, concerns about AI bias might still be present but were minimized, and lender-aligned participants' tendency to avoid explanations when demographic factors were salient diminished—the gap between *neutral* and *lender-aligned* participants shrank to insignificance. This strongly suggests that information avoidance is not about explainability per se, but rather about what might be revealed. When explainability threatens to expose factors that could create moral or ethical dilemmas, decision-makers turn away; when it reinforces their financial interests, they remain eager to engage.

To convincingly attribute information avoidance to image concerns, it is crucial to hold constant the decision environment and underlying decision problem while varying only the salience or framing of moral implications. For instance, participants might strategically avoid explanations if these explanations could reveal biases linked to race or gender. Consistent with this interpretation, I find (within the *lender-aligned* treatment) that participants' tendency to avoid explanations disappears when (1) demographic characteristics (race or gender) of the borrowers are not disclosed, or (2) options to see the explanation are explicitly framed in terms of purely financial information. This pattern strongly suggests that information avoidance is driven not by differences in the decision task itself, but rather by participants' self-image concerns regarding fairness and ethics. A potential alternative explanation—that *lender-aligned* participants might simply be less interested in information overall—is ruled out by the finding that these participants exhibit higher interest in information about the risk prediction itself. This, strategic, phenomenon of seeking information while avoiding information about that information is scantly documented in economics before this paper.

Information avoidance as a means of creating moral wiggle room or plausible deniability is established in economics. Prior research demonstrates that third parties frequently sanction individuals who violate social norms and are willing to incur personal costs to enforce fairness (e.g., Chen and Li (2009)). And there is evidence that third parties similarly sanction strategic avoidance of information (Krupka and Weber (2013), Bartling et al. (2014), Grossman and Van Der Weele (2017)). Recent work by Saccardo and Serra-Garcia (2023) shows individuals are sophisticated about their own tendency to exploit moral wiggle room, but do those who exploit moral wiggle actively penalize others who also attempt to avoid information to create plausible deniability?

This paper provides evidence that third parties do sanction such strategic ignorance. In my experimental setup, participants were asked whether they wished, at a cost to themselves, to penalize another randomly chosen, *lender-aligned* participant who had accepted the AI's recommendation to only approve the loan to the lower-risk White male borrower. I find that participants imposed significantly (51.0%) higher penalties on *lender-aligned* participants who had an opportunity—but chose not—to learn whether race and gender influenced the AI-assigned default risk, compared to those who never had such an opportunity. Strikingly, even *lender-aligned* participants who themselves opted for ignorance imposed these higher penalties (39.8% higher). These findings suggest that individuals, acting as third parties, not only recognize but actively penalize deliberate attempts by others to seek moral wiggle room through strategic information avoidance. They are not going to let someone wiggle out of something they are trying to wiggle out of themselves!

Experiment 2 turns to a simple prediction task - participants make an incentivized, AI-supported prediction regarding the repayment outcome of a real US$10,000-loan. Each participant sees a loan application labeled as high default risk by an AI system trained on real credit data, informed that the AI has high accuracy, and offered a chance to purchase a one-line explanation naming the main feature behind the prediction. Without private information, the explanation offers some consumption or reassurance value (or instrumental value for future predictions); with private information, however, it can become more instrumentally valuable. Specifically, a private note informs participants that the applicant's two-year employment gap is benign (due to a professional certificate program), even though such gaps typically signaled termination in the training data. In

this setting, the explanation reveals whether the gap actually drove the high-risk label, allowing participants to combine their private note with the model's reasoning to form an accurate final prediction.

Despite this complementarity, participants' willingness-to-pay (WTP) for an explanation declines by 25.6% when they hold private information compared to when they do not—contrary to theoretical predictions. This is surprising because winning the explanation guarantees a more accurate signal: if the gap drove the label, participants should correctly infer repayment; if not, the AI's high-risk label should be trusted. The under-demand for explanations thus cannot be rationalized as optimal behavior. To probe the mechanism, we implement a second treatment arm in which participants are guided through a two-case tutorial clarifying how explanations interact with private signals. In this condition, WTP rises by 23.7% with private information, consistent with theory, suggesting that failures of contingent reasoning—rather than indifference or random bidding—drive the observed undervaluation.

We further test whether low WTP reflects an inability to use explanations effectively once acquired. Comparing participants who increase their WTP under private signals to those who do not, we find no significant differences in prediction accuracy, indicating that undervaluation is not due to an incapacity to benefit from explanations. Finally, in a treatment where explanations are revealed only after predictions are locked, participants still bid substantial amounts, indicating a sizable value (not directly instrumental to the current prediction)—about 64.6% of the valuation on the explanation. Together, these results suggest that explanations are systematically under-demanded because people fail to reason through contingencies ex ante even though they improve individual decisions ex post.

This paper is the first to show that explanations, despite their instrumental and social value, are systematically under-demanded—either because individuals strategically avoid them to preserve moral wiggle room or because they fail to reason through their complementarities with private information. Both behaviors distort market demand for AI explainability, with implications for welfare and regulation. If individuals undervalue explanations ex ante, policies that make them salient, subsidize their provision, or lower access costs may improve efficiency. If, instead, avoidance is strategic, interventions such as "right to explanation" laws may be necessary to counteract incentives to hide behind

black-box models. In either case, understanding when and why explanations are shunned is critical for calibrating regulation to balance the costs of explainability against its private and social benefits.

The paper is organized as follows. Section 2 provides contextual knowledge on XAI. Section 4 describes the experiments. In Section 5 I describe my empirical specification and empirical findings. Section 6 concludes.

2 Explainable AI

Explainable AI (XAI) emerged as a response to the increasing complexity of machine learning (ML) models and their "black box" nature.[5] Early AI systems and simpler models (e.g., linear regression) were inherently more transparent but lacked predictive power. As deep learning and ensemble methods began to dominate predictive tasks, it became evident that stakeholders needed not only accurate predictions but also clear insight into how the models arrived at these outputs. Over time, regulatory concerns and high-stakes domains (like healthcare or finance) accelerated the demand for explainability, resulting in the flourishing of XAI research.

In practice, explainable AI methods can be separated into intrinsically interpretable models (e.g., small decision trees, sparse linear regression) and post hoc explanations for complex models. Even linear regression can become too hard to parse once higher-dimensional interactions are introduced. For black-box models, analysts focus on dissecting model internals (partial dependencies, feature importance) or approximating the model through surrogate means (like training a simpler model on the same inputs and outputs). Local methods (e.g., LIME) clarify individual decisions by examining small perturbations near a point, while global methods (e.g., partial dependence plots) offer a broader perspective. One widely used local-and-global approach is Shapley-Additive-Explanation-Values explanations, sometimes referred to as SHAP (Štrumbelj and Kononenko (2014), Lundberg and Lee (2017)), treat each feature in a model as a "player" in a cooperative game (Roth (1988)) whose overall "payout" is the model's prediction. They calculate each feature's contribution by averaging that feature's marginal effect across all possible subsets of features, thus fairly distributing the total prediction among them. In prac-

[5] See Doshi-Velez and Kim (2017), Linardatos et al. (2020), and Molnar (2020) for excellent overviews.

tice, this yields a local feature-attribution perspective for each instance, while aggregate statistics of these attributions also reveal global patterns. It is SHAP that we will turn to when presenting participants in this experiment with an explanation of why the high-default-risk borrower was deemed as such.

To illustrate, imagine a credit scoring model that predicts a numerical risk score for a loan applicant. A Shapley explanation for one applicant might say: "Starting from the average risk score of 5%, having a high debt-to-income ratio added +3% risk, multiple recent credit inquiries added +2%, whereas a strong credit history subtracted −4%, and a high income subtracted −1%, yielding a final predicted risk of $5\%+3\%+2\%-4\%-1\% = 5\%$". Each feature's Shapley value quantifies its contribution in percentage points to this particular applicant's risk. An important point is that these contributions are context-dependent – they depend on the presence of other features and the model's interactions. That is why the Shapley value calculation considers all combinations of features: to properly account for interaction effects (if, say, "income" matters only in conjunction with "debt-to-income ratio," the Shapley framework will allocate the appropriate credit to that interaction). In summary, Shapley values provide a principled way to explain why a complex model gave the prediction it did for a given instance, grounded in an analogy to a fair payoff distribution in a cooperative game.

However, perfect explainability is typically unachievable. The more complex a model is, the more its full logic resists translation into a concise explanation. Evaluating explanations remains ad hoc: no universal measure exists, and humans can only comprehend so many interacting features before it becomes overwhelming (Halford et al. (2005)). Nonetheless, partial explanations or highlights of key features can offer valuable insights, even if they do not fully capture the underlying model's intricacy.

Ultimately, explanations serve multiple purposes—debugging, verifying reliability, building trust, auditing fairness, and ensuring ethical operation—broadly categorized into performance and ethical considerations. From a performance perspective, explainability enhances AI models by enabling developers and users to diagnose errors, detect biases, and improve reliability. For instance, explanations can reveal whether a credit-scoring model is overly sensitive to spurious inputs or whether a medical diagnostic AI is relying on irrelevant background features, allowing practitioners to refine these models. In this way, explainability acts as a tool for quality control: it can catch modeling errors, dataset

biases, or logical glitches that pure accuracy metrics might miss. Moreover, explainability facilitates trust and collaboration, ensuring that human users—such as doctors, financial analysts, or policymakers—are more likely to rely on AI recommendations when they understand the reasoning behind them.

Beyond performance, the second broad impetus for XAI comes from principles of transparency, fairness, and accountability. Explainability helps audit whether models disproportionately affect certain groups—such as by relying on proxies for race or gender—and ensures compliance with fairness.

In principle, the benefits of explainability should appeal to all users. However, in some contexts, decision-makers may deliberately avoid explanations that reveal insights conflicting with their incentives, effectively "hiding behind the black box" to evade acknowledging biases or flaws. In doing so, they may forgo the benefits of explainability, such as improved performance or the ability to audit for errors, in favor of maintaining a self-serving position.

3 Conceptual Framework

To interpret the patterns in both experiments—and to clarify when demand for explanations diverges from demand for predictions—I develop a simple conceptual framework. The framework isolates two forces suggested by the data: (i) strategic avoidance, where explanations eliminate plausible deniability by revealing unfairness and thus impose a psychological cost on profit-maximizing choices; and (ii) systematic undervaluation, where users fail to anticipate how explanations raise the effective accuracy of decisions by complementing private information. A decision-maker chooses between a profit-maximizing and a fair action, can request an explanation that both reveals the fairness state and sharpens beliefs, and derives a direct consumption value from doing so. The model delivers testable predictions that map directly to the experimental treatments (framing of explanations, presence/absence of demographics, aided reasoning), organizing the empirical results and clarifying the welfare stakes of XAI policy.

3.1 Actions and Payoffs

A Decision-Maker (DM) chooses an action $a \in \{a_p, a_f\}$, where a_p is a profit-maximizing action and a_f is a fair alternative. The material payoff $\pi(a, w)$ depends on the state of the world $w \in \{0, 1\}$.

- $\pi(a_p, w = 1) = \Pi_{high}$ and $\pi(a_p, w = 0) = 0$.
- $\pi(a_f, w) = \Pi_{low}$, for any w.

We assume $\Pi_{high} > \Pi_{low} > 0$. The AI provides a prediction $p = P(w = 1|x)$.

3.2 Fairness and Explanations

The action a_p can be either fair or unfair, $F \in \{\text{Fair}, \text{Unfair}\}$. The DM has a prior belief $P(F = \text{Unfair}) = q$. The DM can choose to see an explanation ($e = 1$) or not ($e = 0$).

- **The Explanation ($e = 1$)** perfectly reveals the fairness state F. It also provides information that allows the DM to update their belief about the state of the world from p to a more accurate p_{eff}.

3.3 Accuracy Improvement via Explanations

We can formalize how an explanation helps the DM update their belief from p to p_{eff}. Imagine the feature space is discretized into a set of features $\{x_1, x_2, ..., x_F\}$. The AI's prediction $p = P(w = 1|x)$ is based on observing one of these features. The DM possesses a private binary signal, $b \in \{0, 1\}$, representing contextual knowledge the AI lacks. This signal is relevant only for a specific feature, say x_k. For example, the DM knows that while $P(w = 1|x = x_k)$ is high, the conditional probability $P(w = 1|x = x_k, b = 0)$ is very low. For all other features x_f where $f \neq k$, the signal is uninformative, i.e., $P(w = 1|x = x_f, b = 0) = P(w = 1|x = x_f)$. Without an explanation, the DM only knows the prediction p and cannot apply their private signal. With an explanation, the AI reveals which feature x_f was the primary driver of the prediction.

- If the explanation indicates the feature was x_k and the DM holds the signal $b = 0$, they update their belief to $p_{eff} = P(w = 1|x = x_k, b = 0)$, which is much lower than the original p.

- If the explanation points to any other feature x_f (where $f \neq k$), or if the DM's signal is $b = 1$, their belief remains $p_{eff} = p$.

This mechanism illustrates how explanations provide instrumental value by allowing users to integrate their own knowledge and form a more accurate assessment.

3.4 Utility Function

The Decision-Maker (DM) derives utility from material payoffs, may incur a psychological cost when *knowingly* choosing an unfair action, and may obtain direct consumption value from seeing an explanation. Let $e \in \{0, 1\}$ indicate whether the DM sees an explanation. The utility is

$$U(a, w, F, e) \;=\; \pi(a, w) \;-\; C(a, F, e) \;+\; \nu\, e,$$

where:

- $\pi(a, w)$ is the material payoff from taking action $a \in \{a_p, a_f\}$ in state $w \in \{0, 1\}$.
- $\nu \geq 0$ captures the direct consumption value of receiving an explanation (e.g., curiosity, sense of control, or knowledge that may be instrumentally useful for future decisions).
- $C(a, F, e)$ is a psychological cost $\psi > 0$ incurred *only* if the DM knowingly takes an unfair action: $C(a_p, \text{Unfair}, e = 1) = \psi$ and $C(a, F, e) = 0$ otherwise.

This formulation captures "moral wiggle room" or plausible deniability: when $e = 0$, the DM remains uncertain about the fairness state F and can engage in motivated reasoning (Exley and Kessler (2023); Saccardo and Serra-Garcia (2023); Dana et al. (2007)), such as down-weighting the prior $q \equiv P(F = \text{Unfair})$ or adjusting risk preferences in a self-serving manner (Exley (2016)). For tractability, I do not model these mechanisms explicitly. Instead, I assume that when F is ambiguous (i.e., $e = 0$), the psychological cost is not incurred even though the DM knows there is probability q that a_p is unfair. Under $e = 1$, the explanation removes ambiguity about F, eliminating plausible deniability and triggering the cost ψ if and only if $F = \text{Unfair}$ and the DM chooses a_p.[6]

[6]Building on Bénabou and Tirole (2002), Rabin (1995), and related work in behavioral economics, the DM faces both pecuniary incentives and moral concerns.

3.5 The Decision Problem

We compare the DM's expected utility with and without the explanation. Let $\Pi(p) = p \cdot \Pi_{high}$ be the expected material payoff of action a_p given belief p.

3.5.1 Case 1: No Explanation ($e = 0$)

The DM does not know the fairness state. Due to the ambiguity, the DM has plausible deniability and can engage in motivated reasoning to justify choosing the profitable action. Therefore, the psychological cost ψ is not incurred, even though the DM knows there is a probability q of the action being unfair. The decision is based purely on expected material payoff:

$$E[U|e = 0] = \max(\Pi(p), \Pi_{low})$$

3.5.2 Case 2: With Explanation ($e = 1$)

The explanation removes all ambiguity about the fairness state F, thereby eliminating plausible deniability. The DM sees the explanation, updates their belief to p_{eff}, and learns F.

- With probability $(1 - q)$, F is 'Fair'. The DM's utility is $\max(\Pi(p_{eff}), \Pi_{low})$.
- With probability q, F is 'Unfair'. The DM's utility is $\max(\Pi(p_{eff}) - \psi, \Pi_{low})$.

The total expected utility is the sum of these outcomes, plus the consumption value:

$$E[U|e = 1] = \nu + (1 - q)\max(\Pi(p_{eff}), \Pi_{low}) + q\max(\Pi(p_{eff}) - \psi, \Pi_{low})$$

The DM will request the explanation if its value, $\mathcal{V}(e) = E[U|e = 1] - E[U|e = 0]$, is positive.

3.6 Model Implications: Why Explanations are Under-Demanded

3.6.1 Result 1: Willful Ignorance

A DM may avoid an explanation to dodge a difficult choice. As established, not seeing an explanation provides moral wiggle room, allowing the DM to pursue the profitable

action without psychological cost. Seeing an explanation removes this deniability. **Assumptions for this scenario:**

1. For simplicity, assume the explanation offers no accuracy gain ($p_{eff} = p$) and no consumption value ($\nu = 0$).
2. The DM initially prefers the profitable action: $\Pi(p) > \Pi_{low}$.
3. The fairness cost is high enough to make the DM switch their choice if unfairness is revealed: $\Pi(p) - \psi < \Pi_{low}$.

Now, let's calculate the expected utilities under these assumptions.

- **Without explanation**, the DM leverages plausible deniability and chooses a_p:

$$E[U|e = 0] = \Pi(p)$$

- **With explanation**, the DM's choice depends on what is revealed:
  - If 'Fair' (prob $1 - q$), they choose a_p for a payoff of $\Pi(p)$.
  - If 'Unfair' (prob q), they are forced to switch to a_f for a payoff of Π_{low}.

  The expected utility is therefore:

$$E[U|e = 1] = (1 - q)\Pi(p) + q\Pi_{low}$$

The DM will avoid the explanation if $E[U|e = 0] > E[U|e = 1]$, which means:

$$\Pi(p) > (1 - q)\Pi(p) + q\Pi_{low} \implies q\Pi(p) > q\Pi_{low} \implies \Pi(p) > \Pi_{low}$$

This condition is true by our initial assumption. The DM anticipates that the explanation is likely to reveal bad news that would force them into a lower-payoff action. By remaining ignorant, they can choose the profit-maximizing action without incurring the psychological cost, thereby securing a higher expected utility. This is strategic, willful ignorance. **Corollary:** The incentive for willful ignorance increases with the stakes. The value of avoiding the explanation in this scenario is $\mathcal{V}_{avoid} = E[U|e = 0] - E[U|e = 1] = q(\Pi(p) - \Pi_{low}) = q(p \cdot \Pi_{high} - \Pi_{low})$. This value is strictly increasing in Π_{high}. Therefore,

as the potential payoff from the profit-maximizing action grows, the DM's incentive to avoid information that could compromise their ability to choose it becomes stronger.

3.6.2 Result 2: Welfare Loss from Low Cognitive Ability

The benefit from an explanation's accuracy improvements may not be fully realized if the DM struggles to interpret it. Let $\gamma \in [0, 1]$ be the DM's cognitive ability to use the explanation. The effective belief update is a function of γ, $p_{eff}(\gamma)$, making the expected utility with explanation, $E[U|e = 1, \gamma]$, an increasing function of γ.

- **True value of explanation:** $\mathcal{V}(e|\gamma = 1) = E[U|e = 1, \gamma = 1] - E[U|e = 0]$.

However, a DM with low ability γ_{low} perceives a lower value:

- **Perceived value of explanation:** $\mathcal{V}(e|\gamma_{low}) = E[U|e = 1, \gamma_{low}] - E[U|e = 0]$.

Consider a case where the explanation is objectively valuable, $\mathcal{V}(e|\gamma = 1) > 0$, but the DM's ability is so low that their perceived value is negative, $\mathcal{V}(e|\gamma_{low}) < 0$. The degree of undervaluation can be defined as the gap between the true value and the perceived value: $E[U|e = 1, \gamma = 1] - E[U|e = 1, \gamma]$. Since $E[U|e = 1, \gamma]$ is an increasing (and generally convex) function of γ, this gap widens as γ decreases. Therefore, individuals with lower cognitive ability to process an explanation are predicted to undervalue it more severely. In this situation, the DM will decline to see the explanation, forgoing a potential welfare gain. The resulting **welfare loss** is the difference between the optimal utility and the utility the DM actually achieves:

$$\text{Welfare Loss} = E[U|e = 1, \gamma = 1] - E[U|e = 0] = \mathcal{V}(e|\gamma = 1)$$

When a DM's inability to process an explanation prevents them from recognizing its value, they will under-demand it, leading to a welfare loss equal to the full potential value of that explanation.

4 Experimental Design

To evaluate the comparative statics derived in Section 3, I implement two complementary experimental designs. The first experiment examines whether participants strategically

avoid explanations when the underlying risk signal is known to be highly accurate, as in modern AI-based lending. By fixing beliefs about predictive quality, this design isolates the incremental effect of monetary incentives on the propensity to evade explanations, allowing us to test whether avoidance reflects a desire to preserve self-image rather than uncertainty about algorithmic performance. The second experiment turns to a pure prediction task, where participants evaluate a loan application with an outcome already realized. In this setting, fairness trade-offs are absent, and the focus shifts to whether behavioral factors—such as limits on contingent reasoning—lead participants to undervalue explanations even when they would improve accuracy. Together, these designs separately identify strategic avoidance and systematic undervaluation as distinct forces shaping demand for AI explainability.

4.1 Experiment 1: Allocation of Loans

4.1.1 Decision Task

The experimental framework centers on a decision about whether to allocate an interest-free, non-commercial US$10,000 loan to one or both unemployed individuals in need of funds for essential living expenses. The experiment is conducted in collaboration with a private lender based in the United States.[7] The decision task is embedded within a high-stakes AI application innew financial decision-making, specifically loan allocations. The lender has committed (in a notarized note) to following the input of the selected participant in this study in the allocation of the US$10,000 to each borrower.

Participants, recruited online, engage in a structured decision-making exercise where they assess borrower characteristics and might assess AI-generated default risk predictions. The lender uses an AI algorithm to determine the default risk and the algorithm had been perfectly accurate for the lender so far (participants complete a comprehension check on this fact). The AI model classifies one borrower as "low risk" (default probability $< 10\%$) and the other as "high risk" (default probability $> 90\%$). Participants then determine loan distribution, selecting between:

- **Override the AI's recommendations:** Approve the loan for both borrowers.
- **Accept the AI's recommendations:** Approve the loan only for one of the two

[7]The loans are non-commercial. Therefore, the AI used by the private lender is not regulated by laws that commonly regulate how banks or commercial lenders can deploy algorithmic decision-making.

borrowers (specifically the low-risk borrower if they were informed by the AI already).

The study systematically varies participant incentives, information access, and AI explanation provision to examine their impact on loan allocation decisions and AI reliance.

4.1.2 Key Treatments

Participants are given the context of the private interest-free US$10,000 loan at the beginning of the experiment and asked to check a "Statement of Understanding". Subsequently, participants review real, non-identifying details of two unemployed borrowers earning below the U.S. median income. One is White, the other Black; one is male, the other female.[8] Both are single, childless, similarly educated, and seeking the loan for immediate living expenses. They are then informed of the decision task described above and told about how they can earn a bonus, described below.

4.1.3 Incentive Structures

Participants are randomized into one of two bonus conditions:

- **Neutral Treatment:** A $0.50 bonus for completing the survey, independent of loan repayment outcomes.
- **Lender-Aligned Treatment:** A $1.00 bonus if 100% of approved loans are repaid, introducing an incentive to align recommendations with lender's financial interests.

Participants in the *Lender-Aligned Treatment* maximizes expected monetary returns for themselves (and the lender) if they accept the AI's recommendations and approve the loan only for the low-risk borrower and not for the high-risk borrower.

4.1.4 Information Availability

After being informed of the incentive structure, participants are randomly assigned to one of two groups regarding AI default risk visibility. One fourth of participants can opt to view AI-generated default risk assessments and see for whom the AI recommended approval or disapproval of the loan application (and if they decline, they won't be exposed

[8] In an alternative Demographic-Free treatment, race and gender are replaced by arbitrary markers (odd vs. even digits in phone numbers). This is outlined later on.

to subsequent AI explanations; but they will see an explanation if they chose to see the AI risk predictions) while the rest of participants automatically receive AI default risk predictions.

After a brief reminder of the prospective borrower characteristics and the bonus, participants are then asked if they would like to see an an explanation of how the AI Algorithm made the default risk predictions (except for those participants who opted to not view AI-generated default risk assessments entirely). The way that participants are asked varies across participants. In particular, they are randomized into five explanation exposure conditions:

- **Optional Explanation (Main Treatment):** Participants are offered the option:

  > "Review an explanation of the AI algorithm first, explicitly learning whether and how borrower characteristics influenced the AI's predictions before making your decision to accept or override the AI's approval recommendation."

  and that:

  > "It will show you what borrower characteristics (e.g., borrower's race [Black or White], borrower's gender [Man or Woman], annual income, debt-to-income ratio, home ownership, ...) led to the recommendation to approve or deny the loan."

- **Optional Explanation (Financial Focus):** Participants are asked a question similar to the arm above except that the possibility of the inclusion of financial information as explanatory variables in AI explainability is made salient:

  > "Review an explanation of the AI algorithm first, explicitly learning whether and how borrower financial information/predictors like debt-to-income ratio or annual income influenced the AI's predictions before making your decision to accept or override the AI's approval recommendation."

  and that:

> "It will show you what borrower financial information and predictors (e.g., annual income, debt-to-income ratio, home ownership, ...) led to the recommendation to approve or deny the loan."

- **Mandatory Explanation:** AI explanations are always shown.
- **No Explanation:** AI explanations are neither provided nor optional. This arm is given only one fourth as much randomization weight as the other three arms above.

It is important to note that the first two arms are similar except for the salience (Bordalo et al. (2013)) it places on features that might be available in the explanation. In particular, in the *Optional Explanation (Main Treatment)* treatment arm, the possibility of auditing algorithmic fairness is made salient to the participant.

4.1.5 Demographic-free Borrower Treatment

Of the 75% of participants automatically receive AI default risk predictions, a share of participants similar to the share sent to *Optional Explanation (Main Treatment)* or *Optional Explanation (Financial)* or *Mandatory Explanation* arms are randomized to NOT receive any information about race and gender of the borrowers from the very beginning of the experiment (throughout the whole experiment). For those in this *Demographic-free Borrower Treatment*, the fact that the borrowers are of different races and genders is replaced by the fact that one borrower's phone number is an odd number and the other's phone number is an even number. Intuitively the participants in the *Demographic-free Borrower Treatment* will face the same decisions and optimize over the same bonus scheme, except that they will not know that loan decisions might lead to different outcomes by race and gender (nor would they see an explanation of the AI about how race and gender might have been used for predicting default risks). Participants in the *Demographic-free Borrower Treatment* might have less elevated concerns about algorithmic fairness along race and gender lines.[9] Similar to the arms in subsection 4.1.4, I include a *No Explanation*, and, as in subsection 4.1.4, this arm is given only one fourth as much randomization weight.

[9]It remains possible that participants might find it not fair for people whose phone numbers are different to be treated differently.

4.1.6 AI Explanation Content

For those receiving AI explanations, either because they choose to see an explanation of how the AI Algorithm made the default risk predictions or if they were randomized into a mandatory explanation treatment, a truthful description of the neural network-based default risk prediction that is proprietary to the private lender will be provided. This explanation, if provided, will be provided before the particpant actually make the loan allocation decision. Specifically, an additional randomization determines the how much details the explanation provides:

- **Minimal Explanation:** High-level description of the AI as a neural network-based "black-box" system.
- **Detailed Explanation:** A breakdown of key feature contributions (listing order is randomized across participants), including:[10]

  > "The AI Algorithm (AI) is based on a "default risk prediction model" using machine learning (ML) is a system that leverages historical data and published algorithms to predict the likelihood of a borrower failing to repay a loan, essentially estimating the probability of a "default" on a loan.
  >
  > The specific machine learning technique deployed for this model is neural networks, a powerful form of artificial intelligence. While neural network is a type of "black box" algorithm, we give an intuition of how the AI prediction model estimates the chance that Borrower B defaults below. We list the Shapley-Additive-Explanation-Values which quantifies, on average, the magnitude (positive or negative) of each feature's contribution towards the predicted default risk of a borrower, relative to the average borrower. Some of the top features include:
  >
  > – Because Borrower B's race is non-White, AI decided that default risk is 11% higher
  >
  > – Because Borrower B's sex is woman, AI decided that default risk is 7% higher

[10] The participants in the *Demographic-free Borrower Treatment* will not see this version of explanation.

  - Because Borrower B's does not own a home, AI decided that default risk is 22% higher
  - Because Borrower B's U.S. zip code, AI decided that default risk is 8% higher
  - Because Borrower B's Debt-to-income ratio is over 50%, AI decided that default risk is 30% higher
  - Because Borrower B has no derogatory public records, AI decided that default risk is 19% lower
  - Because Borrower B's annual income is less than $\$40,000$, AI decided that default risk is 17% higher
  - Because Borrower B faces 0% interest rate, AI decided that default risk is 11% lower"

- **Partial Explanation:** The same breakdown as the detailed explanation, but excluding race, gender, and zip code as contributing factors.

I am interested to investigate whether the actual AI explanatory impacts the loan allocation decision. The *Detailed Explanation* arm shows clearly that a racial-minority and female borrower was assigned higher default risk partly due to their race and gender, something that can raise algorithmic fairness concerns.

4.1.7 Accuracy Belief Elicitation

To assess participants' beliefs about AI accuracy, especially whether it shifted after seeing the explanation and after making a loan allocation decision, I include an incentivized accuracy belief elicitation task. Participants are asked to guess how many of the next 10 borrowers' loan repayment outcomes will the AI algorithm correctly predict and can earn an additional $\$0.1$ for answering correctly.

Finally, I ask demographic questions and seven questions to gauge prior knowledge and trust for AI. Finally, to ensure that the participants understood the stakes of their loan allocation decision, I ask participants that if their survey response is randomly selected, what their answers and your answers alone will do. The participants will have to answer correctly (namely, choose the answer "Determine how the private lender will allocate the $\$10,000$ loan between the two real borrowers") to be included in the final analysis sample.

This sample exclusion restriction criteria is pre-registered.

4.1.8 Penalizing Others

I implement a penalty elicitation task designed to measure each participants' willingness to financially penalize another participant ("second participant") based on their loan allocation choices. Specifically, each participant considers a scenario involving a randomly selected second participant, chosen from among those who saw the borrowers' risk scores and whose allocation matched the actual loan distribution implemented by the private lender. This second participant initially receives a bonus of $10, conditional upon successful repayment of 100% of the approved loans. The participant then decides whether, and by how much, to reduce the second participant's bonus at a personal cost of $0.01 per $1.00 reduction, up to the full $10.

I elicit penalty decisions under three distinct scenarios: (1) the second participant was explicitly informed that Borrower B was classified as "high default risk" partly due to being non-White and female, and subsequently denied the loan to Borrower B and approved Borrower A (a White male); (2) the second participant made the same allocation without having any opportunity to learn about the demographic factors contributing to the risk classification; or (3) the second participant had the opportunity but chose not to learn about the role of race and gender in Borrower B's risk assessment, making the same allocation to deny the loan to Borrower B nonetheless. By comparing the penalties that each participant imposes across these scenarios, I directly evaluate how strongly participants wish to sanction behaviors reflecting varying degrees of informed bias or strategic ignorance.

4.1.9 Auditing AI's Predictions vs. Explanations

At the conclusion of the experiment, I implement a within-subject design to evaluate how participants assess the fairness of AI systems when confronted with differences between algorithms based on past algorithmic predictions and explicit use of sensitive attributes as manifested to the participant via an explanation. Participants are presented with two AI algorithms used by a bank to predict loan default risk.

One of the AIs, AI Algorithm 1, may or may not rely on race as an input, but in a finite set of observed cases it has systematically assigned Black borrowers a 10% higher default

risk on average than otherwise similar White borrowers, reflecting historical patterns in past lending decisions. By contrast, AI Algorithm 2 explicitly uses race as an input and consistently assigns Black borrowers a 10% higher default risk than otherwise identical White borrowers across all past, current, and future cases.

Participants are then asked to evaluate how a regulator should respond. Specifically, they indicate whether Algorithm 1 is more unfair and should be prioritized for regulatory action, whether Algorithm 2 is more unfair and should be prioritized, whether both should be treated equally by regulators, or whether neither is unfair and thus no regulatory action is warranted.

Following this fairness assessment, I elicit participants' beliefs about the persistence of these disparities into the future. For each algorithm, participants rate their confidence (on a $0 - 100$ scale) that, across a broad and potentially infinite number of future borrower assessments, the algorithm will continue to assign Black borrowers at least a 10% higher default risk than otherwise identical White borrowers. This design allows me to contrast how participants weigh (i) observed versus explicitly encoded disparities, and (ii) beliefs about the persistence of biased outcomes versus judgments about the legitimacy of algorithmic design.

4.2 Experiment 2: Prediction of Loan Outcome

4.2.1 Prediction Task

The experiment's decision task asks participants how much they are willing to pay (WTP) to see a one-line explanation of why an AI system labeled a loan applicant as high default risk. Using a Becker-DeGroot-Marschak (BDM) procedure, participants state their maximum WTP, with a random price draw determining whether they receive the explanation. They then make a single, incentivized prediction—whether the applicant repays or defaults—earning a bonus based on the real-world repayment outcome, minus any cost paid for the explanation.

Participants, recruited online, must make a single prediction about the applicant's outcome: either "Applicant Defaults" or "Applicant Repays." This prediction is final and cannot be revised. Payoffs are structured as follows: if the participant predicts correctly, they earn a +\$0.50 bonus; if incorrect, they lose −\$0.50. Any price paid to obtain the AI explanation is deducted from this bonus, so participants must weigh the potential value

of the explanation against its cost.

Thus, the choice has two parts: (i) how much to bid ($0 - 150$ cents) for the chance to see the explanation under the BDM rule, and (ii) which of the two discrete outcomes—Defaults or Repays—to predict. The design ensures that both components are incentive-compatible: truthful bidding is the best strategy under BDM, and accurate prediction maximizes expected payoff given the true repayment outcome of the loan applicant

4.2.2 Explanation and Willingness-to-pay

In the task, participants may purchase access to a one-line explanation of the AI's high-risk label. The explanation does not reveal probabilities or the full model but simply identifies the single main feature driving the "High default risk" prediction. As the survey text states: "The Explanation will name the main feature behind this prediction. This will tell you whether a 2-year employment gap is the key reason the applicant was labeled high default risk". In training data, such gaps typically reflected job termination, so knowing whether the gap is the model's driver allows participants to assess whether their private note (indicating the gap was due to a benign certificate program; see below) is directly relevant.

WTP for the explanation is elicited using the BDM mechanism (Becker et al. (1964)). Participants are instructed: "Below, you can state the maximum amount you are willing to pay to see the Explanation (any amount between \$0.00 and \$1.50). ... After you enter your amount, the computer will randomly draw a price ... if your amount is greater than or equal to that price, you pay the drawn price and see the Explanation; if it is lower, you pay nothing and do not see it". The instructions emphasize that the best strategy is to report one's true maximum WTP, making the elicitation incentive-compatible.[11] The description of the BDM mechanism is kept as simple and transparent as possible, emphasizing to participants that "The best strategy is always to stating your most accurate estimate of your maximum you are willing to pay for the chance to

[11] After the pre-registration and before the experiment's launch, we drew a random number among all round number cents (i.e., among \$0.00, \$0.10, \$0.20, \$0.30, \$0.40... , \$1.50) and the number drawn on September 18, 2025 is \$0.00 using a random number generator function on Microsoft Excel. We have then proceeded to use this as the random number to implement the BDM for all the subjects instead of drawing a new number each time. The BDM design is incentive compatible in both the single draw design and the multiple draw design.

review the Explanation" rather than burdening them with the technical details of the incentive structure (a link to a detailed explanation of BDM was provided but can be easily ignored), since over-explaining can actually increase biased or non-truthful reports (Danz et al. (2024)).

4.2.3 Key Treatments

In this experiment, all participants face a within-subjects manipulation of private information. After an initial WTP elicitation when participants have no additional information about the loan applicant beyond what the accurate AI system already knows and has incorporated (The *Pre-information* WTP), they are shown a verified *Private Note* (followed by understanding questions):

- **Private Note:** The applicant's 2-year employment gap is due to a **full-time, fixed-term professional certificate program** (tuition covered; living expenses from savings). It is **not** due to being terminated from a job.

- A 2-year employment gap caused by a certificate program is considered **benign** and should not increase default risk. Only a 2-year employment gap caused by being **terminated from a job** should increase default risk. The AI system does not distinguish between these two cases and always assumes a 2-year employment gap means termination.

This note makes the explanation instrumentally valuable as it can potentially shed light on whether the AI's label was indeed driven by the 2-year employment gap. The *Post-information* WTP is then elicited.

The experiment varies participants' experience along two between-subjects treatment factors. The first is timing of the explanation. In the *Before* condition, participants are told: "If you bid for and win the chance to see the Explanation, it will be shown BEFORE you make your prediction". This makes the explanation potentially useful for improving predictive accuracy and thus earnings. In contrast, in the *After* condition, participants read: "If you bid for and win the chance to see the Explanation, it will be shown AFTER you make your prediction ... it cannot affect your chance of winning the bonus because it is only shown after your prediction is locked". This manipulation separates curiosity or intrinsic value or instrumental value for the future from instrumental value for the

current task.

The second treatment factor is the presence of a reasoning aid to walk the participants through the contingent reasoning necessary to grasp the value of the explanation as a complement to private information possessed by them. In the *Aided reasoning* arm, participants receive a guided walk-through that spells out the contingent logic (followed by understanding questions):

- **Setup you already know**: The system's current label is **HIGH default risk**. You can bid for a chance to see a **one-line Explanation**: whether or not it is the **2-year employment gap** that drove this label. (It can tell you exactly whether you are in Case 1 or Case 2 below) Your **Private Note** says the applicant's **2-year gap** is due to an **accredited certificate program** (benign), **not** termination.

- Case 1 — The Explanation says the main feature is **"2-year employment gap" (2-year employment gap is THE cause of the High Default Risk label)**: (1) The label is **HIGH** because of the **gap**. (2) Your Private Note says *this specific gap is benign* (certificate program). (3) For this applicant, the system has treated a benign gap **as if** it meant termination. (4) **Therefore (for this study):** the applicant **should have been labeled LOW default risk**. (5) That gives you a **very accurate signal** for your bonus-relevant prediction → predict **Applicant Repays**.

- Case 2 — The Explanation says the main feature is **not** the 2-year employment gap **(2-year employment gap is NOT THE cause of the High Default Risk label)**: (1) The label is **HIGH** because of **another feature**, not the gap. (2) Your Private Note is **about the gap**, so it **does not apply** here. (3) **Therefore (for this study):** the **HIGH default risk** label **stands**. (4) That gives you a **very accurate signal** for your bonus-relevant prediction → predict **Applicant Defaults**.

- **One-line rule:** The Explanation tells you whether your Private Note is **directly relevant**.

  - If **relevant** (gap is the main feature) treat the correct label as **LOW**.

  - If **not relevant** treat the current **HIGH** label as correct.

> In other words, if you bid enough to win the Explanation, you can find out whether you are in Case 1 or Case 2. Either way, seeing the Explanation will get you a very accurate signal for your bonus-relevant prediction!

A one-line rule is also emphasized: "The Explanation tells you whether your Private Note is directly relevant — if relevant, treat the correct label as LOW; if not relevant, treat the current HIGH label as correct". In *No aid* arm, participants see the *Private Note* but not this structured logic.

4.3 Bot Detection and Data Integrity

To maintain data integrity, the experiment employs multiple bot-prevention mechanisms. First, I include reCAPTCHA authentication at multiple stages (once at the start of the experiment and once towards the end). Second, I include a "honey-pot" question designed to be opaque to humans but readable by automated scripts. This is very effective as many participants who managed to complete the reCAPTCHA revealed themselves to be bots or deploying bots via this question (see below). I also time the reCAPTCHA authentication response at later stages of the survey to detect non-human engagement (if the response time was longer than the initial reCAPCHA response by twice). This is to prevent the participants who just manually start the study by passing reCAPTCHAs and then handing it over the bots upon passing the reCAPTCHAs (typically placed by researchers in the beginning of their studies). Finally, attention and comprehension checks are included to ensure participant understanding of task requirements. The participants will have to pass both the bot detection and the attention check to be included in the final analysis sample. This sample exclusion restriction criteria is also pre-registered.

4.4 Summary of Design

4.4.1 Experiment 1

Experiment 1 uses a cross subject design. Experiment 1 was run on Prolific with $2,564$ initial participants. The experiment was done on September 4, 2025. Based on the pre-registered exclusion criteria listed above, 12 were excluded for not passing the bot checks, then 32 were excluded for failing the attention checks above, and then 8 were excluded because they took more than twice longer in the second reCAPTCHA near the end of

the experiment than the first. Main analysis sample based on pre-registered exclusion criteria[12] comes down to 2,512 participants.[13]

The treatment arms described above is summarized in Figure I.

Insert Figure I about here.

4.4.2 Experiment 2

Experiment 2 uses a between-subjects 2×2 design, crossed with a within-subjects manipulation. Experiment 2 was run on Prolific with 1,027 initial participants. The experiment was done on September 18, 2025. Based on the pre-registered exclusion criteria listed above, 11 was excluded for not passing the bot checks, then 19 were excluded for failing the attention checks above, and then 1 was excluded because they took more than twice longer in the second reCAPTCHA near the end of the experiment than the first. Main analysis sample based on pre-registered exclusion criteria comes down to 1,006 participants.

The number of participant in each treatment arm end up being as follows: 503 in the *before* treatment (out of which 254 randomized to receive the tutorial on contingent reasoning [*Aided reasoning*], and 249 randomized to not [*No aid*]) and 503 in the *after* treatment (out of which 241 randomized to receive the tutorial on contingent reasoning [*Aided reasoning*], and 262 randomized to not [*No aid*]).

5 Main Results

5.1 Experiment 1: Strategic Avoidance of Explanations

Empirical results discussed in this Section are based on analytical tables pre-registered in the analysis plan unless otherwise stated.

[12] I kept participants who did not answer "Determine whether the private lender will approve the $10,000 loan(s) for the two real borrowers" in the analysis sample because based on comments and feedback received after the pre-registration. Excluding those participants yield a smaller sample but one that nonetheless yields similar results across.

[13] I encourage experimentalists to adopt of the bot detection techniques described in Subsection 4.3 when running studies on Prolific from 2025 onwards due to the uptick in bot usage in online platforms.

5.1.1 Simultaneous Information Seeking and Information-about-Information Avoiding

On average, the majority of participants prefer more information. With 80.0% opting to see the default risk scores and 46.2%[14] choosing to examine the explanation for the AI-generated default risk scores across all treatments. This is not surprising given that, by design, there is no monetary charge and only the hassle cost of clicking and reviewing an additional page (c_r, the costs of observing the risk signal and c_e, the cost for the explanation are both low).

Reviewing column (1) of Table I, I find that while 76.5% of participants in the *neutral* treatment who were given an opportunity opted to see the default risk score, and additional 7.1 percentage point more choose to see the risk score in the *lender-aligned* treatment. We conjectured that when the lender places heavier emphasis on monetary returns, there will be *higher demand for the risk signal.* The data seems to bear out this way.

The pattern is different for information that come in form of an explanation of the AI's risk prediction. Here I focus on the participants who were given a choice to choose whether or not to see an explanation, having randomized into reviewing the default risk scores. This helps alleviate selection bias, as those who selected not to see risk scores will not see an explanation either. Columns (2) through (5) shows the information seeking/avoidance results for the full sample of people who were given a choice. Overall, aggregating across all the ways that the option to review the explanation is presented to the participants, there is a significantly higher rate of avoiding the explanation by participants who are *lender-aligned*, by 8.8 percentage points or 19.5% (and significant at the 1%-level). This is show in column (2). As the participants are aware that the signal in the AI risk score is quite accurate in the past, participants might be inclined to avoid explanations to not reveal fairness concerns.

Insert Table I about here.

A quick review of the columns (3) to (5), I notice some heterogeneity. On one hand, in column (3), when the option to review explanation was present making the possibility of the inclusion of both financial information, race and gender as explanatory variable

[14] When given a choice.

salient, relative explanation avoidance by participants who are *lender-aligned* is significant and higher at 10.4 percentage points or 23.4% (and significant at the 1%-level).

On the other hand, the difference in explanation seeking/avoidance between *neutral* and *lender-aligned* drops to levels not significantly different from zero (at conventional levels) if the option to review explanation was presented in ways that only made the possibility of the inclusion of financial information as explanatory variable salient (column (4)).

Next, we review whether participants' explanation avoidance behavior change when fairness concerns are made much less salient. In column (5) in Table I, the basic decision environment and information environment is identical to the main treatment. However, recall that under this *demographic-free borrower* treatment, the fact that the borrowers are of different races and gender is replaced by the fact that one borrower's phone number is an odd/even number. With this small change, the difference between in explanation seeking/avoidance between *neutral* and *lender-aligned* drops to levels not significantly different from zero (at conventional levels). The fact that the coefficient on "*lender-aligned*" both dropped in magnitude and went from significant to not significant indicates that the basically identical decision and information environments are not driving the results of explanation avoidance.

Finally, running the same regression but including the control for being randomized into the "option to view AI recommendation" treatment, the results above stay largely similar (see columns (6) to (9) in Table I).

5.1.2 Explanation Affects Loan Allocation

An *equal allocation* where the AI's recommendation is overridden and both US$10,000 loans are approved yields less expected monetary returns to the lender than the *risk-based allocation* where the AI's recommendation is accepted and only the loan to the low-risk borrower is approved. Overall, *equal allocation* is the less chosen option across the board, with only 18.2% of participants selecting this loan allocation. Even for *neutral* participants, only 23.9% chose *equal allocation* (Cf 12.6% among *lender-aligned* participants who directed monetary gains tied to loan repayment). This could be due to the fact that the participants are aware that the private lender might be harmed by a loan default.

Focusing on the main sample, I evaluate whether the rate of choosing *equal allocation*

for the $10,000-loans is affected by differences in the explanation shown to the participant. If AI explainability might reveal bias and undercut a purely self-interested allocation (lowering the agent's self-image if they invest heavily in one borrower), the finding that *lender-aligned* participants are more likely to avoid explanations is only half the picture. We need to show that participants might have a reasonable prior that an explanation that reveals sensitive traits (e.g., race, gender) underlie the prediction would nudge them towards the less-profitable *equal allocation*.

Column (1) of Table II, shows that when presented with an explanation that reveals sensitive traits (e.g., race, gender) led to less favorable risk predictions by the AI, participants are 7.5 percentage points or 41.7% more likely to shift toward an *equal allocation* of the loans. This result is significant at the 1%-level. This effect is just less than double (10.2 percentage points more likely, significant at the 5%-level; see column (3)) for those who actually made a decision to see the explanation and see that race and gender were included in the AI default risk calculation.

Insert Table II about here.

Explanation avoidance behavior, documented in Subsection 5.1.1, is correlated with the choice of an *equal allocation* as well. In column (2), we see that those who sought AI explainability are 5.7 percentage points *more* likely to choose an *equal allocation* of the $10,000-loans (significant at the 5%-level). In fact, a plurality of the participants who chose to see an explanation, after getting hints that financial information, race and gender might be part of the explanation, stated in a free text response something related to algorithmic fairness (e.g. race, gender, bias; Figure II). On the other hand, as shown in column (5), those who avoid AI explainability are 5.7 percentage points *less* likely to choose an *equal allocation* of the $10,000-loans (significant at the 5%-level).

Insert Figure II about here.

Except for the explanation that reveals sensitive traits (e.g., race, gender) led to less favorable risk predictions by the AI, the choice of *equal allocation* is similar whether or not an explanation was shown (column (4)).

One alternative hypothesis for why an explanation that reveals sensitive traits (e.g., race, gender) underlie the prediction would nudge participants towards *equal allocation*

which is deemed by the AI risk score as being less-profitable, is that the explanation shifted the beliefs about how accurate the AI is. To control for this, I re-run the regressions presented in column (1) to (5) while adding a control that captures the belief of accuracy held by the participants after reviewing the explanations. This measure is based on an incentivized guess about how many of the next 10 borrowers' loan repayment outcomes will the AI algorithm correctly predict. A quick comparison between columns (6) and (10) of Table II will reveal that the impact of explanations on the choice to make an *equal allocation* is not explained away by shifts in beliefs about the AI's accuracy.

5.1.3 Third Parties Punish Information Avoiders

Economists have a blossoming literature on information avoidance as a means to buy moral wiggle room. This Subsection presents evidence on whether third parties penalize "moral wigglers". That is, even if wigglers can excuse themselves, can others excuse them?

When I ask participants to act as a third party and decide whether they want to penalize another randomly selected participant who made the *risk-based allocation* where the approved for the low-risk White male borrower and denied for the high-risk non-White female borrower, and I ask these third parties to do so at a personal cost (they have to pay \$0.01 to inflict each \$1.00 penalty), third parties on average inflict a \$2.19 penalty. But these penalty levels vary drastically, depending how the one who chose the *risk-based allocation* engaged with AI explainability.

In Table III, I show that if the someone made the *risk-based*, "unfair" allocation after having (not chosen to but) seen that the high risk borrower was labeled by AI as such partly due to race and gender, third parties are willing to inflict 66.5% higher penalties or an additional \$1.47 (significant at the 1%-level). What is remarkable, is that the increase in penalties is slightly but significantly lower for "wigglers", those who had the opportunity but chose not to learn about the role of race and gender in the AI risk assessment: third parties are willing to inflict 51.0% higher penalties or and additional \$1.12 (significant at the 1%-level), relative to those who did not have a choice to see an explanation. This is shown in column (1).

Insert Table III about here.

Do third parties who are wigglers themselves act differently? In other words, are they more or less lenient on others who made the *risk-based*, "unfair" allocation? Columns (2) to (5) shows that wiggler third parties penalize others as well. Column (2) shows that wigglers are significantly different (more lenient) in the average penalties they impose. Column (4) shows that if the someone made the *risk-based*, "unfair" allocation after having (not chosen to but) seen that the high risk borrower was labeled by AI as such partly due to race and gender, wiggler third parties are willing to inflict slightly more lenient but significant additional penalties of \$0.87 (significant at the 1%-level). And just like the others, if someone made the same allocation without having any opportunity to learn about the demographic factors contributing to the risk classification, wigglers do not penalize them differently than others (column (5)).

Interestingly, wigglers penalizes other wigglers. Participants who are themselves wigglers, penalize those who had the opportunity but chose not to learn about the role of race and gender in the AI risk assessment at an additional \$0.69 (significant at the 1%-level), relative to those who did not have a choice to see an explanation. This is smaller in magnitudes than non-wigglers but it seems that wigglers are trying to get out of something that they wouldn't let others get out of! The evidence in column (7) confirms these results.

5.1.4 Value of Explanations

In real-world settings, we often observe a stream of AI predictions over time, even if we cannot open the Black Box or are not provided with an explicit explanation of how it works. Policymakers and researchers, for example, frequently audit algorithmic outputs to assess performance and fairness. This naturally raises two questions: what additional value do explanations provide beyond auditing predictions, and do people perceive their understanding of an AI differently when based on observed outcomes versus explicit explanations?

Most respondents (73.1%) judge an algorithm that explicitly and transparently uses race as an input—directly assigning Black borrowers a higher default risk than otherwise identical White borrowers—to be as unfair as an algorithm that may (or may not) use race, but which has produced a pattern of assigning Black borrowers higher risks in the observed cases so far (see columns (4)–(6) in Table IV). Across all participants, 30.8

percentage points more believe that the former should be banned by regulators. However, among those who do perceive a difference in unfairness, nine times more believe that the explicitly race-based algorithm is more unfair and should therefore be subject to stricter regulatory oversight.

Participants also believe that disparities will persist into the future with significantly higher likelihood (22.9% or 15.8 percentage points higher on average) when the algorithm is explicitly using race as an input compared to when disparities are inferred from past predictions alone (see columns (1)–(3) in Table IV). This suggests that explanations shape not only fairness judgments but also expectations about the persistence of algorithmic bias.

5.2 Experiment 2: Undervaluation of Explanations

Explanations can also add value when users or regulators possess contextual information unavailable to the AI. In such cases, the explanation's purpose is not merely to expose the model's reasoning but to serve as a cue, triggering users to recognize when their external knowledge is relevant and should be applied to contextualize—or even override—the model's output.

The second experiment illustrates this dynamic and the behavioral factors that limit how much decision-makers appreciate this value.

Insert Table V about here.

5.2.1 Explanation Valuable but Undervalued in Some Instrumental Aspects

Participants valued explanations provided before making their bonus-earning loan prediction significantly more than those shown only after the prediction was locked in: on average, willingness-to-pay was 14 cents higher (a 54.8% increase, significant at the 1%-level; see Columns (1) and (2) in Table V). This pattern is consistent with the expectation that information with immediate instrumental value for improving prediction accuracy should command a premium. Interestingly, participants still placed a substantial positive value—roughly 64.6% of the before-treatment value—on explanations revealed only after the prediction was submitted (see bars in Figure III). This suggests that explanations also

carry intrinsic consumption value or provide instrumental benefits beyond the immediate prediction task.

Insert Figure III about here.

Next, I examine elicited willingness-to-pay for explanations. In the *No Aid* condition—where participants did not receive any reasoning aid on how explanations interact with private information—participants valued explanations at 42.3 cents when they had no additional information about the applicant and 31.4 cents after receiving the Private Note (see the first two bars in Figure IV). Thus, explanations were valued 34.4% higher (10.8 cents more, significant at the1%-level; see Column (3) in Table V) in the absence of outside information than when the private signal was present—a reversal of the theoretical prediction that explanations should be more valuable when paired with private information.

Insert Figure IV about here.

By contrast, in the *Aided reasoning* condition, where participants were provided with a guided decision aid, the pattern reversed. Here, participants valued explanations about 23.8% lower (9.3 cents less, significant at the 1%-level; see Column (4) in Table V) when they lacked the private signal than when they possessed it. The average willingness-to-pay in this condition was 48.5 cents, just below the theoretical upper bound of 50 cents implied by the expected gain from having a perfectly accurate predictor.

5.2.2 Explanations are Valuable for Prediction Across the Board

Overall, 72.6% of participants correctly predicted that the borrower repaid the loan, a rate significantly better than chance.

A useful comparison is between participants who reduced (or did not increase) their valuation of explanations after receiving private information and those who increased their valuation once the private note was revealed.

Although participants submitted different bids, the BDM random draw was \$0.00, so all participants ultimately received the explanation regardless of their stated WTP. This lets us compare the prediction accuracy between groups submitting different bids for the explanation without worrying about selection.

As shown in Figure V, there are no significant differences in prediction accuracy (at conventional levels) between these groups. This suggests that participants who did not raise their valuation of explanations after receiving private information are not simply less capable of using the explanation effectively.

Insert Figure V about here.

Together with the results from the last subsection, these findings suggest that participants were able to use explanations effectively to improve prediction accuracy and earn bonuses, but were less able to anticipate their instrumental usefulness without explicit guidance on contingent reasoning. This provides clear evidence that explanations are undervalued in some instrumental aspects.

6 Discussion and Conclusion

This paper investigates how economic agents demand, use, and avoid explanations of AI predictions in high-stakes credit settings. Two experiments with real $10,000 loans yield three main findings. First, demand for information is selective: when profit incentives are salient, participants actively seek accurate risk predictions but strategically avoid explanations that could reveal reliance on sensitive attributes. When explanations do expose such reliance, participants are more likely to overturn profit-maximizing recommendations, and third parties impose costly sanctions on deliberate avoidance. Second, explanations are systematically undervalued ex ante, even when they complement private information: willingness-to-pay falls when participants hold private signals that raise the instrumental value of explanations. Third, a minimal tutorial that makes contingent logic explicit reverses this pattern, restoring the theoretically predicted demand. Together, these results imply that observed market demand understates the *social* value of explainability.

The mechanisms are distinct and call for different policy tools. Where avoidance is strategic—rooted in image concerns and moral wiggle room—voluntary markets will underprovide explanations even when they correct socially costly errors. In such settings, targeted mandates or auditor access to minimally sufficient, decision-relevant explanations can improve welfare without requiring full transparency. Where undervaluation

reflects contingent reasoning failures, light-touch interventions such as tutorials, worked examples, or interface cues can substantially raise willingness-to-pay. Importantly, participants also attach positive value to explanations even when they cannot alter the current decision, underscoring that explanations provide both consumption value and dynamic instrumental benefits beyond one-shot accuracy.

These findings have sharp implications for AI governance. Regulations mandating explainability—such as the EU AI Act or the U.S. Equal Credit Opportunity Act—may constrain the use of complex black-box models, but they also counteract self-serving incentives by collapsing moral wiggle room. At the same time, such mandates can serve as implicit subsidies for explanation, correcting behavioral undervaluation when individuals fail to appreciate complementarities with private signals. Explainability should therefore be viewed not only as a technical attribute of models but also as an *economic good*, with demand shaped by incentives, beliefs, and cognitive constraints. Regulation that combines (i) accountability in settings prone to strategic avoidance with (ii) low-cost scaffolds where undervaluation is behavioral can better align private choices with social value.

Future work should integrate technical XAI development with behavioral insights into information avoidance and undervaluation. The challenge is not merely to design transparent models, but to ensure that transparency is meaningfully engaged with by decision-makers. Mechanisms—monetary, reputational, or legal—that raise the cost of willful ignorance or lower the cognitive cost of contingent reasoning may prove central. More broadly, the welfare analysis of explainability should move beyond accuracy–transparency tradeoffs to account for strategic and behavioral distortions in demand.

Ultimately, these results highlight a broader challenge in the economics of AI: ensuring that algorithmic systems are not only accurate and efficient but also just and transparent. Left unchecked, the strategic use of opaque models risks eroding fairness and public trust in critical markets. If the promise of AI is to enhance both efficiency and equity, future policy and design must address not only the technical feasibility of explainability, but also the behavioral dynamics that determine if and when it is actually used.

References

Bakker, Trevor J, Stefanie DeLuca, Eric A English, James S Fogel, Nathaniel Hendren, and Daniel Herbst, "Credit Access in the United States," Technical Report, National Bureau of Economic Research 2025.

Bartling, Björn, Florian Engl, and Roberto A Weber, "Does willful ignorance deflect punishment?–An experimental study," *European Economic Review*, 2014, *70*, 512–524.

Becker, Gordon M, Morris H DeGroot, and Jacob Marschak, "Measuring utility by a single-response sequential method," *Behavioral science*, 1964, *9* (3), 226–232.

Bénabou, Roland and Jean Tirole, "Self-confidence and personal motivation," *The quarterly journal of economics*, 2002, *117* (3), 871–915.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer, "Salience and consumer choice," *Journal of Political Economy*, 2013, *121* (5), 803–843.

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan, "Productivity and selection of human capital with machine learning," *American Economic Review*, 2016, *106* (5), 124–127.

Chen, Yan and Sherry Xin Li, "Group identity and social preferences," *American Economic Review*, 2009, *99* (1), 431–457.

Consumer Financial Protection Bureau, "Equal Credit Opportunity Act (Regulation B), 12 C.F.R. § 1002.9(b)(2)," 2011. Accessed: March 3, 2025.

Dana, Jason, Roberto A Weber, and Jason Xi Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, 2007, *33* (1), 67–80.

Danz, David, Lise Vesterlund, and Alistair J Wilson, "Evaluating behavioral incentive compatibility: Insights from experiments," *Journal of Economic Perspectives*, 2024, *38* (4), 131–154.

Doshi-Velez, Finale and Been Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

Esponda, Ignacio and Emanuel Vespa, "Hypothetical thinking and information extraction in the laboratory," *American Economic Journal: Microeconomics*, 2014, *6* (4), 180–202.

European Parliament and Council of the European Union, "General Data Protection Regulation (GDPR), Recital 71," 2016. Regulation (EU) 2016/679, enacted 2016, effective 2018. Accessed: March 3, 2025.

_ , "Artificial Intelligence Act, Article 86: Right to Explanation of Individual Decision-Making," 2023. Proposal for a Regulation on Artificial Intelligence (AI Act), Accessed: March 3, 2025.

Exley, Christine L, "Excusing selfishness in charitable giving: The role of risk," *The Review of Economic Studies*, 2016, *83* (2), 587–628.

_ **and Judd B Kessler**, "Information avoidance and image concerns," *The Economic Journal*, 2023, *133* (656), 3153–3168.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably unequal? The effects of machine learning on credit markets," *The Journal of Finance*, 2022, *77* (1), 5–47.

Golman, Russell, David Hagmann, and George Loewenstein, "Information avoidance," *Journal of economic literature*, 2017, *55* (1), 96–135.

Graeber, Thomas, Christopher Roth, and Constantin Schesch, "Explanations," Technical Report, CESifo Working Paper 2024.

Graziani, Mara, Lidia Dutkiewicz, Davide Calvaresi, José Pereira Amorim, Katerina Yordanova, Mor Vered, Rahul Nair, Pedro Henriques Abreu, Tobias Blanke, Valeria Pulignano et al., "A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences," *Artificial intelligence review*, 2023, *56* (4), 3473–3504.

Grossman, Zachary and Joel J Van Der Weele, "Self-image and willful ignorance in social decisions," *Journal of the European Economic Association*, 2017, *15* (1), 173–217.

Halford, Graeme S, Rosemary Baker, Julie E McCredden, and John D Bain, "How many variables can humans process?," *Psychological science*, 2005, *16* (1), 70–76.

Hoong, Ruru and Bnaya Dreyfuss, "Improving AI-Assisted Decision-Making Through Calibrated Coarsening," *Available at SSRN 5286198*, 2025.

IBM Corporation and Morning Consult, "IBM Global AI Adoption Index 2022," 2022. Accessed: March 3, 2025.

Jacovi, Alon, "Trends in explainable ai (xai) literature," *arXiv preprint arXiv:2301.05433*, 2023.

Kleinberg, Jon and Sendhil Mullainathan, "Simplicity creates inequity: implications for fairness, stereotypes, and interpretability," in "Proceedings of the 2019 ACM Conference on Economics and Computation" 2019, pp. 807–808.

_ , Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," *The quarterly journal of economics*, 2018, *133* (1), 237–293.

Krupka, Erin L and Roberto A Weber, "Identifying social norms using coordination games: Why does dictator game sharing vary?," *Journal of the European Economic Association*, 2013, *11* (3), 495–524.

Lage, Isaac, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez, "An evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1902.00006*, 2019.

Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, 2020, *23* (1), 18.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan, "The unreasonable effectiveness of algorithms," in "AEA Papers and Proceedings," Vol. 114 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2024, pp. 623–627.

Lundberg, Scott M and Su-In Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, 2017, *30.*

Luttmer, Erzo FP, Priscila de Oliveira, and Dmitry Taubinsky, "Failures of Contingent Reasoning in Annuitization Decisions," Technical Report, National Bureau of Economic Research 2025.

Martínez-Marquina, Alejandro, Muriel Niederle, and Emanuel Vespa, "Failures in contingent reasoning: The role of uncertainty," *American Economic Review*, 2019, *109* (10), 3437–3474.

Molnar, Christoph, *Interpretable machine learning*, Lulu. com, 2020.

Mullainathan, Sendhil and Ziad Obermeyer, *A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions*, National Bureau of Economic Research, 2019.

Niederle, Muriel, Alvin E Roth, and M Utku Ünver, "Unraveling results from comparable demand and supply: An experimental investigation," *Games*, 2013, *4* (2), 243–282.

Rabin, Matthew, "Moral preferences, moral constraints, and self-serving biases," 1995.

Roth, Alvin E, *The Shapley value: essays in honor of Lloyd S. Shapley*, Cambridge University Press, 1988.

Saccardo, Silvia and Marta Serra-Garcia, "Enabling or limiting cognitive flexibility? Evidence of demand for moral commitment," *American Economic Review*, 2023, *113* (2), 396–429.

Shafir, Eldar and Amos Tversky, "Thinking through uncertainty: Nonconsequential reasoning and choice," *Cognitive psychology*, 1992, *24* (4), 449–474.

Stigler, George J, "The economics of information," *Journal of political economy*, 1961, *69* (3), 213–225.

Štrumbelj, Erik and Igor Kononenko, "Explaining prediction models and individual predictions with feature contributions," *Knowledge and information systems*, 2014, *41*, 647–665.

Figure I: Number of Participants in Each Condition in Analysis Sample

		The way that the option to see AI explanation is presented to participants:					
		Main	Financial info. only	Demographic-free	Main + No explanation	Demographic-free + No explanation	Always explanation
AI Recommendation mandatory	Lender-aligned	202	210	213	57	52	213
	Neutral	211	200	213	57	52	214
Option to view AI recommendation	Lender-aligned	103	68	0	37	0	103
	Neutral	102	69	0	34	0	102

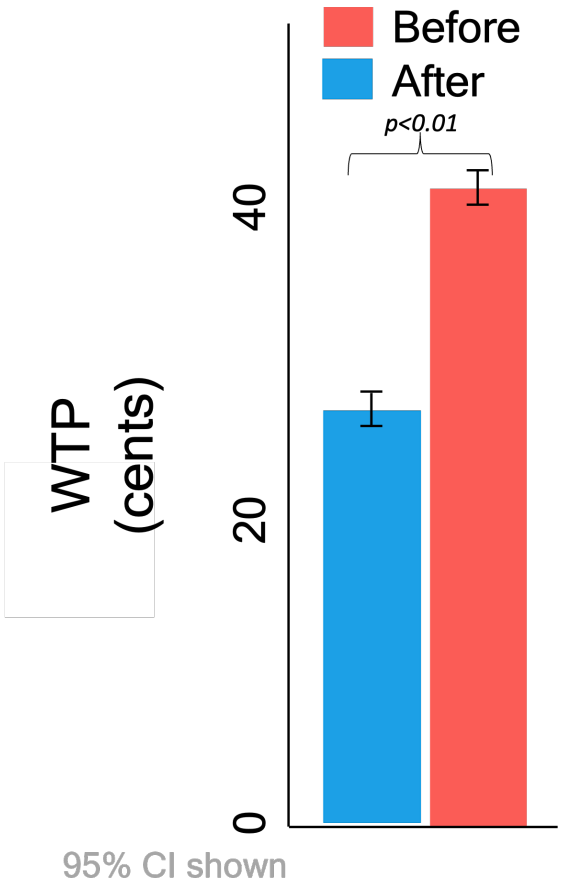
Notes: This figure presents the present the number of subjects in the analysis sample randomized into each of the treatment arms described in Section 4. What is shown here is the sample after implementing the pre-registered exclusion criteria.

Figure II: What Participants Hope to Learn from the Explanation Wordcloud



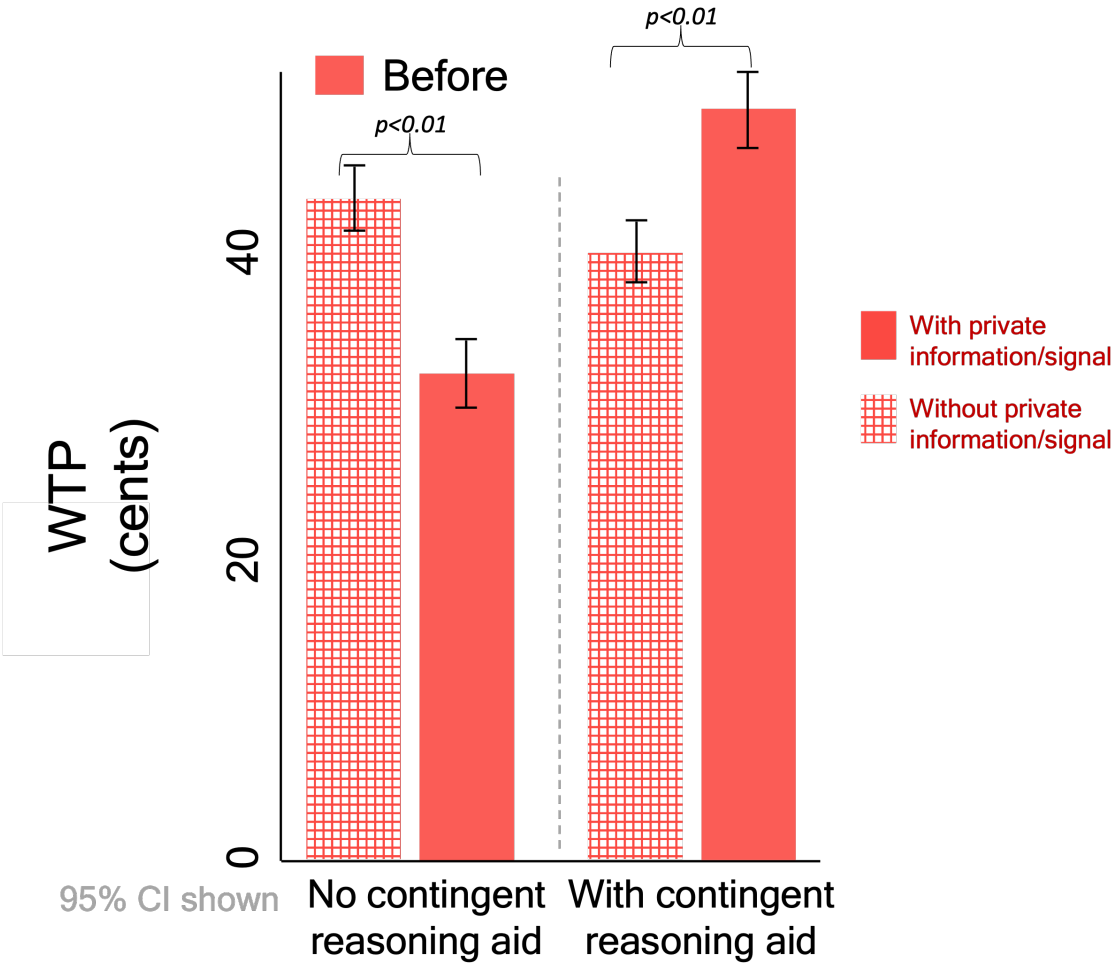
Notes: This figure presents the present the word cloud formed by the responses to the question "You chose to see an explanation of how the AI Algorithm made the default risk predictions BEFORE making the loan decision, what were you hoping to learn from the explanation?", among the participants who was presented and taken up the option to review an explanation where the chance to see the role race and gender play in the default risk calculation by the AI is made salient.

Figure III: WTP for Explanation When Offered Before or After Prediction



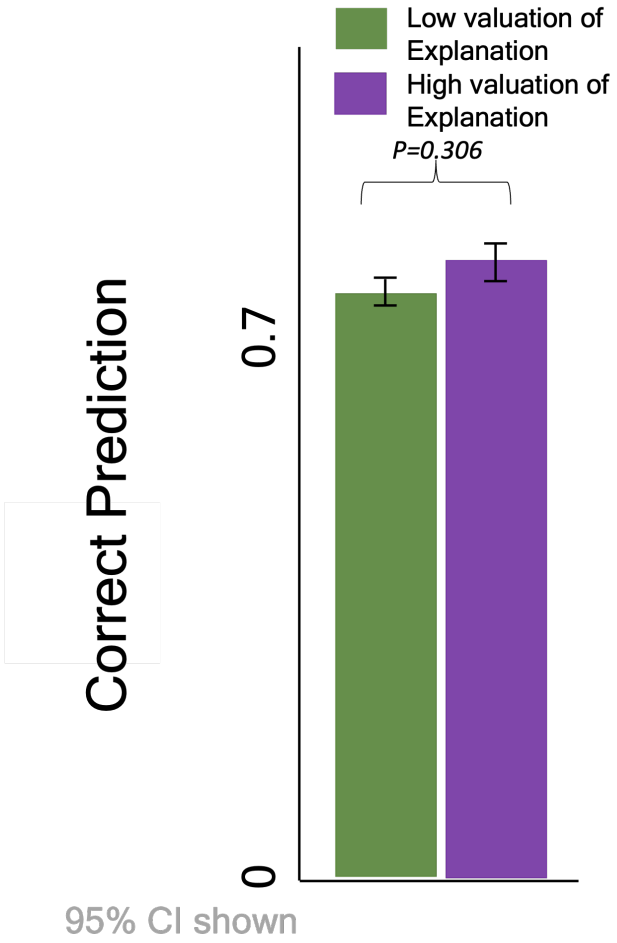
Notes: This figure presents the present WTP estimates (via BDM) estimated with the data from the full sample collected in September 18, 2025, this could range from \$0.00 to \$1.50.

Figure IV: WTP for Explanation With and Without Reasoning Aid



Notes: This figure presents the present WTP estimates (via BDM) estimated with the data from the full sample collected in September 18, 2025, this could range from \$0.00 to \$1.50.

Figure V: Prediction Accuracy Comparison Based on Explanation Valuation



Notes: This figure presents the present likelihood of making a correct prediction of the loan repayment/default outcome estimated with the data from the full sample collected in September 18, 2025. Low valuation of Explanation is defined as participants who submitted a WTP that is the same or lower when they have when they have contextual information unavailable to the AI (v.s. when they do not). High valuation of Explanation indicates the complementary set of participants.

Table I: Information Seeking and Avoidance: Rates of Choosing to See Information

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Default Risk	Explanation for the Default Risk Prediction							
		Without Default Risk Info Choice Control				With Default Risk Info Choice Control			
		All	Main	Financial-info	Demo-free	All	Main	Financial-info	Demo-free
Lender-aligned	0.071*	-0.088**	-0.104*	-0.070	-0.070	-0.086**	-0.101*	-0.069	-0.070
	(0.032)	(0.030)	(0.041)	(0.044)	(0.048)	(0.030)	(0.041)	(0.044)	(0.048)
Constant	0.765**	0.535**	0.545**	0.524**	0.465**	0.555**	0.570**	0.538**	0.465**
	(0.023)	(0.021)	(0.029)	(0.032)	(0.034)	(0.023)	(0.032)	(0.033)	(0.034)
Control for Default Risk	N/A	N	N	N	N	Y	Y	Y	Y
N	618	1097	578	519	426	1097	578	519	426
R^2	0.008	0.008	0.011	0.005	0.005	0.013	0.018	0.008	0.005

Standard errors in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table presents the estimated parameter results for the linear probability model $Y_i = \beta_0 + \beta_1 \textit{Lender-aligned}_i + \epsilon_i$. The data here is from the full sample collected in September 4 2025. We model the outcome variables by the binary variable indicating if the participant chose to reveal the information (default risk, or explanation). The independent variable "Lender-aligned" indicates the treatment where participants will receive a bonus if all of the $\$10,000$ loans are repaid to the private lender. The labels "Main" and "Financial-info" indicate the way that the participant was asked if they would like to see an an explanation of how the AI Algorithm made the default risk predictions, the latter if only the inclusion of financial information as explanatory variables in AI explainability is made salient. "Demo-free" is when the demographic information (race/gender) are replaced throughout by odd/even number phone numbers.

Table II: Impact of Explanations: Loan Allocation and Belief of AI Accuracy

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Chose to Override AI (**1**[Approve loans to both borrowers])									
AI used Race/Gender	0.075**		0.102*			0.076**		0.101*		
	(0.022)		(0.048)			(0.022)		(0.048)		
Revealed Explanation		0.057*					0.059*			
		(0.024)					(0.024)			
No Explanation Option				-0.043					-0.044	
				(0.031)					(0.031)	
Avoided Explanation					-0.057*					-0.059*
					(0.024)					(0.024)
ex post Accuracy Belief						-0.052**	-0.042**	-0.025	-0.052**	-0.042**
						(0.012)	(0.015)	(0.034)	(0.012)	(0.015)
Constant	0.185**	0.179**	0.218**	0.205**	0.236**	0.696**	0.589**	0.466	0.712**	0.648**
	(0.010)	(0.016)	(0.020)	(0.009)	(0.017)	(0.119)	(0.148)	(0.335)	(0.120)	(0.149)
N	1982	1165	538	1982	1165	1982	1165	538	1982	1165
R^2	0.006	0.005	0.009	0.001	0.005	0.015	0.012	0.010	0.010	0.012

Standard errors in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table presents the estimated parameter results for the linear probability model with loan allocation as outcome and for the OLS model with AI accuracy as outcome. The data here is from the full analysis sample in the main arm (not *demographic-free*) collected in September 4, 2025. The outcome of interest reported in this table is "**1**[Approve loans to both borrowers]" (reported in the first five columns). The outcome is indicator variable indicating that to allocate the loan as **Equal Allocation:** where the participant overrode the AI's recommendation and approved both borrowers, instead of only approving the US$10,000-loan to the low-risk borrower, and denying the loan to the high-risk borrower. The independent variable "AI used Race/Gender" indicates that the borrower's race and gender are listed in the explanation that the participant saw as factors that contributed to the high-risk borrower's assessment by the AI. The independent variables "Revealed Explanation" indicates the participant chose to see the explanation when offered a choice, "No Explanation Option" indicates that the participant did not get a chance to choose whether or not to see the explanation, and "Avoided Explanation" indicates the participant not chose to see the explanation when offered a choice. "*ex post* Accuracy Belief" is the participant's incentivized guess about how many of the next 10 borrowers' loan repayment outcomes will the AI algorithm correctly predict. Column (3) uses only the subsample where the participant chose to reveal the explanation.

Table III: Third Party Penalty on Deliberate Attempts by to Avoid AI Explainability

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
			Penalty in \$ (max = \$10)				
			Full Sample				Wiggler
			3^{rd} *Party:*				
			Avoider	*Knew*	*Ignorant*		
3P Avoided Explanation	1.121**					1.186**	0.692**
	(0.053)					(0.058)	(0.125)
3P Knew AI used Race/Gender	1.467**					1.568**	0.808**
	(0.063)					(0.070)	(0.133)
1^{st}P Wiggler		-0.524**	-0.600**	-0.866**	-0.105	-0.105	
		(0.147)	(0.174)	(0.185)	(0.142)	(0.142)	
Wiggler×3P Avoided Explanation						-0.494**	
						(0.138)	
Wiggler×Knew AI used Race/Gender						-0.761**	
						(0.150)	
Constant	1.332**	2.264**	2.533**	2.914**	1.346**	1.346**	1.241**
	(0.049)	(0.057)	(0.070)	(0.077)	(0.053)	(0.053)	(0.132)
N	7536	7536	2512	2512	2512	7536	1002
R^2	0.039	0.003	0.004	0.007	0.000	0.043	0.016

Standard errors (Robust, clustered by participant) in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table presents the estimated parameter results for the OLS estimation model $Y_i = \beta_0 + \beta_1 \mathbf{1}[ExplanationAvoided]_i + ... + \epsilon_i$. The data here is from the full sample collected in September 4, 2025. The column labels "Wiggler", "*Avoider*", "*Knew*", and "*Ignorant*" indicate the subsample used for that column. We model the outcome variable as the how much the participant, acting as a third party, want to reduce the second participant's bonus at a personal cost of \$0.01 per \$1.00 reduction, this could range from \$0 to \$10. The independent variable "3P Avoided Explanation" indicates the target of punishment had the opportunity but chose not to learn about the role of race and gender in Borrower B's risk assessment and subsequently denied the loan to Borrower B and approved Borrower A (a White male). The independent variable "3P Knew AI used Race/Gender" indicates the target of punishment was explicitly informed that Borrower B was classified as "high default risk" partly due to being non-White and female, and subsequently denied the loan to Borrower B and approved Borrower A (a White male). The last independent variable "1^{st}P Wiggler" is an indicator variable for participants, as the decision-maker whether to and how much to punish as a third party, who themselves were assigned to the *lender-aligned* treatment, and had a choice, knew that race/gender of borrowers are different, and chose to avoid seeing the explanation.

Table IV: AI Interpretation for Future Based on Past Outputs vs. Explanation

	(1)	(2)	(3)	(4)	(5)	(6)
	Belief that AI Continue Being Racist			AI Should be Banned		
Explanation Explicitly Use "Race"	15.778**		18.463**	0.308**		0.327**
	(0.578)		(1.175)	(0.011)		(0.019)
Avoided Explanation		-3.713**	-0.935		-0.032*	-0.009
		(1.292)	(1.460)		(0.016)	(0.025)
Exp. "Race"×Avoided Exp		-5.556**			-0.045+	
			(1.528)			(0.027)
Constant	61.091**	69.763**	60.531**	0.577**	0.756**	0.592**
	(0.565)	(0.925)	(1.082)	(0.010)	(0.011)	(0.018)
N	5024	3182	3182	5024	3182	3182
R^2	0.069	0.004	0.068	0.120	0.001	0.121

Standard errors (Robust, clustered by participant) in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table presents the estimated parameter results for the OLS estimation model $Y_i = \beta_0 + \beta_1 \mathbf{1}[ExplanationRacist]_i + ... + \epsilon_i$. The data here is from the full sample collected in September 4, 2025. The column labels "Belief that AI Continue Being Racist" and "AI Should be Banned" indicate the outcome used for that column. The independent variable "Explanation Explicitly Use "Race" indicates the AI Algorithm thay explicitly and clearly uses race as an input and directly assigns Black borrowers a higher default risk than identical White borrowers. "Avoided Explanation" indicates the participant not chose to see the explanation when offered a choice.

Table V: Willingness-to-pay for Explanations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Both *Before* and *After*			*Before*			*After*	
	All	No Private Info.	*No aid*	*Aided reasoning*	Both	*No aid*	*Aided reasoning*	Both
Before	14.295**	13.527**						
	(2.043)	(2.125)						
w/ Private Info.			-10.823**	9.311**	-10.823**	-4.122*	-0.095	-4.122*
			(1.965)	(1.793)	(1.964)	(1.701)	(1.585)	(1.700)
w/ Private Info.×Aided reasoning					20.134**			4.027+
					(2.659)			(2.324)
Constant	26.067**	27.163**	42.261**	39.150**	42.261**	27.385**	26.921**	27.385**
	(1.451)	(1.524)	(2.162)	(2.031)	(2.161)	(2.028)	(2.303)	(2.027)
N	2012	1006	498	508	1006	524	482	1006
R^2	0.039	0.039	0.024	0.016	0.029	0.004	0.000	0.002

Standard errors (Robust, clustered by participant) in parentheses
$^{+}$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$

Notes: This table presents the estimated parameter results for the OLS estimation model $WTP_i = \beta_0 + \beta_1 \mathbf{1}[X]_i + ... + \epsilon_i$. The data here is from the full sample collected in September 18, 2025, where X is either *Before* (for columns (1) and (2)) or *w/ Private Info.* (for columns (3) to (8)). The column labels indicate the subsample used for that column. We model the outcome variable as the how much the participant submitted as their willingness-to-pay (via BDM), this could range from \$0.00 to \$1.50. The independent variable "*Before*" indicates the *Before* treatment where participants were bidding on an Explanation to be revealed to them BEFORE they make a prediction of loan outcome. The independent variable "*Before*" indicates the *w/ Private Info.* (within-subject) treatment where participants were bidding on an Explanation when they have contextual information unavailable to the AI9.31. Finally, "*w/ Private Info.×Aided reasoning*" is the interaction term between the treatment where participants were bidding on an Explanation when they have contextual information unavailable to the AI and the treatment where participants were exposed to a reasoning aid to walk the participants through the contingent reasoning necessary to grasp the value of the explanation