Preference for Explainable AI in Consumer Finance

Alex Chan

Harvard University

(Please click here for the most current version)

September 4, 2025*

Abstract

Around half (48%) prefers to understand how a Black-Box AI operates in a high-stakes decision to approve actual US$10,000-loans. When AI explanation reveals that a racial-minority and female borrower was assigned higher default risk partly due to their race and gender, individuals are 13% more likely to override the AI's recommendation that would have minimized default risk for the lender. This is not driven by shifts in beliefs regarding the AI's accuracy. The plurality of those preferring explainability stated the desire to audit algorithmic fairness as a reason. However, relative to a neutral treatment where payoffs are not tied to actual loan repayment outcome, decision-makers with direct stakes tied to loan repayment are 15% more likely to avoid information that can help them interpret the AI's decision even though they are 15% more likely to choose to know the AI-generated risk score itself. When the potential to audit algorithmic fairness is made unlikely ex ante by framing the explanation only around how financial information affects loan approval, individuals with selfish motivations no longer avoided explainability more. When the perceived borrower pool was distorted to limit algorithmic fairness concerns, individuals with selfish motivations are far less likely to avoid AI explainability. Participants willingly incur personal costs to penalize lender-aligned individuals who deliberately avoid information and profit from unfair loan allocations, although less than those who explicitly know that the AI is not fair.†

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

1 Introduction

People want to know how AI makes decisions—until knowing means they can no longer look away.

AI explainability (the ability to explain "how" a prediction was made)[1] has surged in importance as a research topic. In the recent 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)). However, when the enrichment of oneself and fairness come into conflict, the demand for AI explainability can be decoupled from the demand for AI predictions. For example, a bank lending money might want to minimize

*Email address: achan@hbs.edu. This study was approved by the Institutional Review Board at Harvard University (IRB-25-0040). The study was pre-registered on the AEA RCT registry, ID AEARCTR-0015581. 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]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.

credit risk but disparities in credit risk across demographics might disproportionately limit access to some groups. Economists have long been aware that information is a valuable asset that decision-makers seek (Stigler (1961)). However, self-image concerns could lead individuals to avoid information instrumental to their decisions if such action can help them "excuse selfishness" (Exley and Kessler (2023), Exley (2016), Dana et al. (2007)) or distort their beliefs self-servingly (Saccardo and Serra-Garcia (2023)). In this paper, I investigate whether and why people proactively seek information while avoiding information about that information in a high-stakes AI-supported decision. The aforementioned extant literature on AI explainability primarily explores how to achieve and evaluate it. This paper adds a novel perspective by examining when people might avoid or prefer less AI explainability, and why.

I examine how financial incentives shape engagement with AI-generated information in loan decisions. 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.[3]

Empirical work in consumer lending underscores that explainability tools can improve oversight and regulatory compliance, though usability and fidelity vary widely. This literature (e.g., Davis et al. (2023), Bussmann et al. (2021)) largely assumes that more explainability is always preferred—but what if people sometimes avoid explanations when

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

knowing them risks revealing uncomfortable truths? This paper contributes by showing that, in consumer financial markets, individuals with financial incentives selectively seek or avoid AI explanations depending on whether the explanation threatens or supports their moral and self-image concerns.

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? Most participants seem to believe so—between $57\% - 63\%$ 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—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 the role of financial information in the model's prediction; in another, they were also told it would shed light on the role of race and gender. 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 actively sought out the information. This reversal 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 near 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.

From a policy perspective, the possibility that decision-makers strategically avoid explainable AI (XAI) to maintain moral wiggle room has important implications for regulations governing algorithmic decisions, including "right to explanation" laws. While the supply of XAI technologies continues to increase, the exploitation of AI "black boxes" to avoid confronting algorithmic fairness issues could result in an inefficiently low demand for XAI, particularly by users (such as banks) whose incentives may align with turning a blind eye to fairness concerns. Furthermore, if decision-makers intentionally forgo explanations—even when such information could enhance model reliability and predictive accuracy—to preserve moral wiggle room, this behavior may impede the development and adoption of the most predictive AI technologies more broadly.

Conversely, if strategic avoidance of explainability due to moral wiggle room is rare or nonexistent, excessive regulatory intervention mandating explainability could impose unnecessary costs, potentially discouraging the use of powerful yet opaque models that deliver superior predictive performance. Understanding whether and why users engage in strategic explanation avoidance is therefore essential for calibrating optimal regulatory interventions.

Moreover, the existing literature on optimal explainability often highlights an inherent accuracy–explainability trade-off: perfect explainability is typically infeasible because a full explanation of a complex model requires comparable complexity in the explanation itself. Consequently, researchers and practitioners typically settle for explanations that maximize understanding within cognitive constraints (e.g., Lage et al. (2019)). However, if strategic information avoidance regarding explainability is prevalent across various economic contexts, then a purely technical assessment of optimal explainability may misestimate the actual demand and value placed on explainability by decision-makers,

thus necessitating careful reconsideration of the welfare implications of explainability regulations. In short, this study offers timely, policy-relevant insights by exploring a common, policy-relevant scenario that is not over-simplified to assess whether decision-makers deliberately avoid AI explainability as a means to preserve moral wiggle room.

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 (83.6%) 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 (55.3% 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!

Overall, while many seek AI explanations to audit fairness, individuals with financial stakes often avoid them, especially when fairness concerns are made salient. However, when fairness is not a concern, this avoidance diminishes. Moreover, participants are willing to penalize those who strategically avoid explainability while benefiting from biased outcomes. These findings highlight how self-interest shapes AI engagement, raising critical policy implications for algorithmic explainability mandates and the strategic use of black-box AI models in decision-making.

The paper is organized as follows. Section II provides contextual knowledge on XAI. Section III describes the experiment. In Sections IV, V, and VI I describe my empirical specification and empirical findings. Section VII 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.[4] 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

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

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 practice, 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 Experimental Design

3.1 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.[5] The decision task is embedded within a high-stakes AI application in financial decision-making, specifically loan allocations. The lender guarantees a minimum of US$1,000 to each borrower, with the remaining US$8,000 subject to allocation based on participant input.

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 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.

3.2 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

[5]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.

other female.[6] 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.

3.2.1 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.

3.2.2 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 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:

[6] 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.

> "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.

3.2.3 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.[7] Similar to the arms in subsection 3.2.2, I include a *No Explanation*, and, as in subsection 3.2.2, this arm is given only one fourth as much randomization weight.

3.2.4 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

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

is randomized across participants), including:[8]

> "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
> - 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"

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

- **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.

3.2.5 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.

3.2.6 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.

3.2.7 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.

3.2.8 Bot Detection and Data Integrity

To maintain data integrity, the experiment employs multiple bot-prevention mechanisms. First, I include CAPTCHA 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. I also time the CAPTCHA authentication response at later stages of the survey to detect non-human engagement (if the response time was longer than the initial CAPCHA response by twice). 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.

3.3 Summary of Design

My experiment uses a cross subject design. The experiment was run on Prolific with $2,278$ initial participants. The experiment was done on September 4, 2025. Based on the pre-registered exclusion criteria listed above, 0 were excluded for not passing the bot checks, 0 were excluded for failing the attention checks above, and 0 were excluded for not thinking that the stakes of the decisions is that, if chosen, their answer will "Determine whether the private lender will approve the $\$10,000$ loan for the two real borrowers". Main analysis sample based on pre-registered exclusion criteria comes down to $2,278$.[9]

The treatment arms described above is summarized in Figure I. Online Appendix Table 1 summarizes the treatments and the corresponding volume of participants.

Insert Figure I about here.

[9] I encourage experimentalists to adopt of the bot detection techniques described in Subsection 3.2.8 when running studies on Prolific from 2025 onwards due to the uptick in bot usage in online platforms.

4 Main Results

4.1 Short Conceptual Framework Without Formalism

Building on Bénabou and Tirole (2002), Rabin (1995), and related work in behavioral economics, we consider a lender who encounters both pecuniary incentives and moral concerns. The lender can purchase two AI-generated signals: a *risk signal* that predicts borrower creditworthiness (potentially reliably), and an *explanation* indicating whether protected traits (e.g., race, gender) underlie the prediction. Access to these signals modifies the lender's beliefs, monetary returns, and moral costs in a manner consistent with self-signaling and motivated reasoning.

Formally, let the lender's utility be given by

$$U = \alpha\,\mathbb{E}[\Pi] - \beta\,\mathcal{M}(\mu),$$

where $\alpha > 0$ denotes the importance of expected monetary gains $\mathbb{E}[\Pi]$, while $\beta > 0$ represents the salience of moral costs. The term $\mathcal{M}(\mu)$ is an endogenous image penalty that increases in the lender's posterior belief μ about whether their lending rule is "unfair," in the sense of employing sensitive features. If the lender acquires and inspects the explanation, they update μ (possibly according to Bayes' rule); if she remains ignorant, she can maintain a baseline μ (akin to "plausible deniability"). Denote by c_r and c_e the costs of observing the risk signal and the explanation, respectively.

When the risk signal is perceived as reliable, the lender typically purchases it if α is large relative to c_r: precise credit assessments raise expected returns. However, viewing the explanation can reveal that the algorithm relies on sensitive attributes, driving μ upward and amplifying the image penalty. If the lender learns of bias and still allocates loans unequally to maximize returns, they suffer the image penalty, which increases with β. Consequently, a lender anticipating this image cost may forgo the explanation altogether, preserving moral deniability (Exley and Kessler (2023); Saccardo and Serra-Garcia (2023); Dana et al. (2007)).

Tension arises when the explanation also helps verify the *reliability* of the risk signal. If the lender is unsure whether the AI is truly accurate, learning about its methods (e.g., that it is not biased or is indeed correct) can raise expected profits. At the same

time, such information can impose a larger moral cost if it confirms the algorithm is employing objectionable features. Lenders with strong fairness concerns (high β) who discover potential bias might choose more balanced allocations to protect their self-image, while those with high α may avoid learning about the bias and pursue profit-maximizing yet inequitable lending.

In an experimental setting where the risk signal is deliberately perceived as sufficiently accurate, the marginal benefit of obtaining the explanation (in terms of higher expected monetary returns) is small. Hence, as α increases and the lender suspects that maximizing returns may involve prioritizing a single borrower or group, they become more likely to skip the explanation to avoid the moral cost. These comparative statics illustrate how the decision to remain ignorant can facilitate self-serving actions, consistent with motivated beliefs frameworks.

4.1.1 Predicted Comparative Statics (When Weight on Monetary Returns Increases)

When the lender places heavier emphasis on monetary returns (e.g., transitioning from *neutral* to *lender-aligned*), we obtain:

- *Higher demand for the risk signal.* Agents become more inclined to acquire it, particularly when its cost c_r is low and it is presumed accurate, since enhanced predictive precision leads to larger expected profits.

- *Conditional demand for the explanation.* If the reliability of the risk signal is uncertain, the explanation enables the agent to verify its legitimacy, boosting expected returns. Conversely, if the signal's accuracy is already perceived as high, the marginal benefit of the explanation diminishes; as a result, acquiring it becomes less likely, especially if it might reveal fairness issues.

- *Elevated motivation to remain ignorant.* When fairness concerns are salient and the agent believes that protected attributes are integral to the prediction, those with a strong monetary focus may intentionally avoid the explanation to maintain moral deniability. If fairness concerns are less explicitly triggered, avoidance is somewhat weaker.

To evaluate these comparative statics, we implement an experimental design in which

the risk signal is explicitly portrayed as highly accurate, paralleling real-world advances in AI-based lending. By fixing beliefs about predictive quality, we cleanly identify the incremental effect of increasing monetary incentives on agents' propensity to evade the AI explanation. This approach isolates whether participants avoid explainability to preserve self-image, rather than influences stemming from uncertainty about AI performance.

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

4.2 Simultaneous Information Seeking and Information-about-Information Avoiding

On average, the majority of participants prefer more information. With 79.2% opting to see the default risk scores and 47.5%[10] 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 74.7% of participants in the *neutral* treatment who were given an opportunity opted to see the default risk score, and additional 9.4 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 (6) through (9) 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 7.7 percentage points or 14.8% (and significant at the 1%-level). This

[10] When given a choice.

is show in column (6). 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 (7) to (9), I notice some heterogeneity. On one hand, in column (7), when the option to review explanation was present making the possibility of the inclusion of both financial information, race and gender as explanatory variable salient, relative explanation avoidance by participants who are *lender-aligned* is significant and higher at 10.8 percentage points or 18.7% (and significant at the 5%-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 (8)).

Next, we review whether participants' explanation avoidance behavior change when fairness concerns are made much less salient. In column (9) 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 substantially 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 removing the control for being randomized into the "option to view AI recommendation" treatment, the results above stay largely similar (see columns (2) to (5) in Table I).

4.3 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 19.2% of participants selecting this loan allocation. Even for *neutral* participants, only 23.9% chose *equal allocation* (Cf 14.2% 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 20.3% more likely to shift toward an *equal allocation* of the loans. This result is significant at the 5%-level. This effect is more than double (45.8% more likely, marginally significant at the 10%-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 4.2, is correlated with the choice of an *equal allocation* as well. In column (2), we see that those who sought AI explainability are 22.2 percentage points *more* likely to choose an *equal allocation* of the $10,000-loans (significant at the 1%-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 22.6 percentage points *less* likely to choose an *equal allocation* of the $\$10,000$-loans (significant at the 1%-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 (1) 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.

4.4 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 $\$2.20$ 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 107.9% higher penalties or an additional \$1.45 (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 83.6% 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 similarly to the others! Column (2) shows that wigglers are significantly different 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 \$0.86 (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 \$0.70 (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.

4.5 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 (53.8%) 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, 30.9 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 (25.9% or 15.7 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.

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.

A secondary experiment (see Online Appendix) illustrates this point. Participants were asked (and incentivized) to predict the actual default outcome of a loan applicant, with some randomly assigned private, outside knowledge that the AI did not possess. Using the Becker–DeGroot–Marschak method, I elicited their willingness-to-pay for an

explanation. In the "simple" condition, where participants did not receive a tutorial on contingent reasoning, they valued explanations about 20% higher when they lacked outside knowledge than when they possessed it—a reversal of what theory would predict. In contrast, in the "contingent reasoning aided" treatment, the relative valuation reversed: participants valued explanations about 17% lower when they lacked outside knowledge than when they possessed it.

These findings highlight that failures of contingent reasoning can cause individuals to undervalue explanations, thereby complicating the policy debate around mandating explainability and interpretability requirements in AI systems.

5 Discussion and Conclusion

This paper provides novel experimental evidence, relevant to both economic theory and policy, demonstrating how self-interest significantly influences engagement with AI systems, particularly regarding explainability. It contributes to the economics literature on information avoidance and moral wiggle room by showing these and other richer versions of these dynamics operate strategically in the context of AI adoption. And it also adds an important perspective to the consumer financial markets where regulation of AI and algorithmic applications is prevalent. While a majority of individuals initially express a preference for AI transparency—especially to audit algorithmic fairness —this demand diminishes markedly when fairness considerations conflict with self-serving incentives. Participants with lender-aligned incentives not only exhibited a higher tendency to avoid explainable AI (XAI) when fairness-related variables such as race and gender were salient, but also disproportionately favored decisions maximizing financial gain, even at the cost of reinforcing potential algorithmic bias.

Crucially, individuals do not avoid information regarding AI altogether. They actively seek AI-generated predictions when those serve their monetary interests. However, they simultaneously avoid information about how these predictions were derived when it risks revealing demographically sensitive information. This asymmetric pattern—seeking predictive output while avoiding explainability—suggests a strategic form of ignorance designed to preserve plausible deniability. This behavior presents a subtle but powerful challenge to emerging regulatory efforts mandating AI explainability, which often as-

sume transparency directly leads to fairer decision-making. Furthermore, the study finds that such strategic ignorance is recognized and penalized; third parties willingly incur personal costs to punish decision-makers who avoided explanations potentially revealing bias before making profit-maximizing, inequitable allocations, even when the punishers themselves had engaged in similar avoidance.

These findings have sharp implications for AI governance, rooted in behavioral economics. Regulations mandating algorithmic explainability—like the EU AI Act or the U.S. Equal Credit Opportunity Act—may constrain the use of highly complex, opaque models. Yet, such mandates serve a critical function: counteracting self-serving incentives by making the moral costs of potentially biased decisions salient. When these costs are hidden, individuals exploit "moral wiggle room" to justify inequitable actions; making algorithmic fairness salient through explainability can collapse this wiggle room and shift choices, even if predictive efficiency is somewhat impacted.

This underscores the importance of treating explainability as both a technical and behavioral policy challenge. Algorithmic transparency is not simply a supply-side issue of designing explainable/interpretable models; it is also a demand-side issue of incentivizing end-users to engage meaningfully with explanations. Without aligning the incentives of users—especially those with decision-making power—regulatory mandates risk fostering symbolic compliance rather than substantive fairness.

Therefore, closer integration between technical XAI development and behavioral research on AI adoption and use is essential. Together, they must inform the design of markets and regulations that do not merely require AI to be transparent, but ensure that such transparency is functionally engaged with. Future research and policy should explore mechanisms—monetary, reputational, or legal—that increase the cost of willful ignorance, particularly when it facilitates discriminatory or inequitable outcomes.

Ultimately, these findings point to a broader challenge in the age of algorithmic decision-making: ensuring that AI systems are not only accurate and efficient but also just and transparent. The strategic use of black-box AI models, when left unchecked, threatens to undermine public trust and fairness in critical markets. If the promise of AI is to be fully realized—enhancing both efficiency and equity—it is imperative that the behavioral dynamics of information engagement, not just technical transparency, shape the next generation of AI policy and design.

References

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

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.

Bussmann, Niklas, Paolo Giudici, Dimitri Marinelli, and Jochen Papenbrock, "Explainable machine learning in credit risk management," *Computational Economics*, 2021, *57* (1), 203–216.

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.

Davis, Randall, Andrew W Lo, Sudhanshu Mishra, Arash Nourian, Manish Singh, Nicholas Wu et al., "Explainable Machine Learning Models of Consumer Credit Risk.," *Journal of Financial Data Science*, 2023, *5* (4).

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

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.

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.

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*.

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.

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.

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	186	193	191	50	47	195
	Neutral	187	180	192	51	48	194
Option to view AI recommendation	Lender-aligned	94	61	0	33	0	95
	Neutral	92	65	0	29	0	95

Notes: This figure presents the present the number of subjects in the analysis sample randomized into each of the treatment arms described in Section 3. 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.

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	0.094**	-0.091**	-0.109*	-0.071	-0.045	-0.090**	-0.108*	-0.069	-0.045
	(0.034)	(0.032)	(0.044)	(0.046)	(0.051)	(0.032)	(0.044)	(0.046)	(0.051)
Constant	0.747**	0.541**	0.558**	0.522**	0.448**	0.558**	0.579**	0.534**	0.448**
	(0.024)	(0.023)	(0.031)	(0.033)	(0.036)	(0.024)	(0.034)	(0.035)	(0.036)
Control for Default Risk	N/A	N	N	N	N	Y	Y	Y	Y
N	564	993	521	472	383	993	521	472	383
R^2	0.013	0.008	0.012	0.005	0.002	0.012	0.017	0.008	0.002

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	0.039*		0.088+			0.040*		0.088+		
	(0.020)		(0.049)			(0.020)		(0.049)		
Revealed Explanation		0.226**					0.226**			
		(0.018)					(0.018)			
No Explanation Option				0.008					0.008	
				(0.028)					(0.028)	
Avoided Explanation					-0.226**					-0.226**
					(0.018)					(0.018)
ex post Accuracy Belief						-0.021*	-0.005	-0.019	-0.020+	-0.005
						(0.010)	(0.011)	(0.034)	(0.010)	(0.011)
Constant	0.126**	0.000	0.210**	0.133**	0.226**	0.327**	0.044	0.396	0.330**	0.271*
	(0.009)	(0.012)	(0.021)	(0.008)	(0.013)	(0.103)	(0.106)	(0.335)	(0.103)	(0.107)
N	1800	1058	491	1800	1058	1800	1058	491	1800	1058
R^2	0.002	0.135	0.007	0.000	0.135	0.004	0.135	0.007	0.002	0.135

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
			Avoider	*Knew*	*Ignorant*		
Explanation Avoided	1.124**					1.189**	0.698**
	(0.056)					(0.061)	(0.129)
Knew AI used Race/Gender	1.450**					1.552**	0.790**
	(0.066)					(0.073)	(0.144)
3^{rd}P Wiggler		-0.515**	-0.588**	-0.859**	-0.097	-0.097	
		(0.159)	(0.187)	(0.199)	(0.152)	(0.152)	
Wiggler×Explanation Avoided						-0.492**	
						(0.142)	
Wiggler×Knew AI used Race/Gender						-0.763**	
						(0.161)	
Constant	1.345**	2.271**	2.547**	2.910**	1.358**	1.358**	1.261**
	(0.052)	(0.060)	(0.074)	(0.080)	(0.056)	(0.056)	(0.142)
N	6834	6834	2278	2278	2278	6834	912
R^2	0.038	0.003	0.004	0.007	0.000	0.042	0.015

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 "Explanation Avoided" 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 "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 "3^{rd}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 Racist	15.862**		18.422**	0.309**		0.320**
	(0.606)		(1.235)	(0.011)		(0.020)
Avoided Explanation		-2.840*	-0.412		-0.037*	-0.025
		(1.347)	(1.535)		(0.016)	(0.026)
Exp. Racist×Avoided Exp		-4.857**			-0.025	
			(1.608)			(0.028)
Constant	61.212**	69.694**	60.483**	0.577**	0.759**	0.599**
	(0.593)	(0.972)	(1.141)	(0.010)	(0.012)	(0.019)
N	4556	2882	2882	4556	2882	2882
R^2	0.070	0.002	0.069	0.122	0.002	0.123

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. .