"People can change, and patterns can be broken": Contextualizing Tradeoffs in Automated Decision-Making Systems

Rabeya Bosri
University of Alberta
Canada

Anna Harbluk Lorimer
University of Chicago
USA

Afrida Hossain
University of Alberta
Canada

Vasisht Duddu
University of Waterloo
Canada

Bailey Kacsmar
University of Alberta
Canada

Abstract

Automated decision-making (ADM) systems are increasingly deployed in domains such as mortgage lending, prison sentencing, health insurance coverage, and hiring. Designing a responsible ADM system in such high-stakes domains requires ensuring privacy protection, fairness across demographic groups, and robustness against adversarial manipulation. However, prioritizing one of these objectives comes at the cost of another, forcing a choice as to which tradeoff to accept in a deployment. These tradeoffs explicitly or implicitly impact the life, safety, and fundamental rights of the people in a society, and thus, the perceptions and priorities of this population are needed before we can produce appropriate solutions. To this end, we conducted a quasi-experimental study (N = 777) in which participants evaluated four decision-making scenarios with controlled tradeoffs. Participants significantly preferred human decision-making (HDM) over ADM in three of four scenarios, emphasizing the value of human judgment, contextual understanding, and the ability to incorporate non-quantifiable factors. Furthermore, in terms of tradeoffs, our findings not only show that participants' preferences are highly context-dependent, but also that their perception of a specific objective, fairness, extends beyond formal definitions. Participants interpret fairness through multiple lenses, including privacy risks and susceptibility to manipulation, and view unfair or manipulated outcomes as failures of accuracy. Overall, our findings highlight the importance of context-aware and human-centered approaches when designing and governing ADM systems in high-stakes situations. Rather than purely technical objectives, it is essential to evaluate ADM systems based on how their tradeoffs align with specific expectations within a given domain, as well as with societal values and perceptions of harm and fairness.

CCS Concepts

• **Security and privacy → Usability in security and privacy**; **Human and societal aspects of security and privacy**.

CCS '26, Hague, Netherlands

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Keywords

Differential Privacy, Group Fairness, Adversarial Manipulation, User Study

ACM Reference Format:
Rabeya Bosri, Anna Harbluk Lorimer, Afrida Hossain, Vasisht Duddu, and Bailey Kacsmar. 2026. "People can change, and patterns can be broken": Contextualizing Tradeoffs in Automated Decision-Making Systems. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (CCS '26).* ACM, New York, NY, USA, 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Scenarios in which humans were primarily responsible for decision-making are increasingly being delegated to automated decision-making (ADM) systems. These ADM systems rely on patterns learned from data, primarily historical data, to make decisions in new contexts [59]. When such ADM systems are applied to high-stakes contexts, where decisions affect an individual's health, safety, or fundamental rights, the consequences can be severe. Over the last two decades we have seen multiple instances in which the use of ADM systems has raised concerns about unfair outcomes for certain demographic groups [21, 48, 89], concerns about the privacy of the training data [22, 40], concerns for adversarial tampering with the system [55, 103]. As these systems are increasingly deployed in sensitive domains such as finance, recidivism, and healthcare [22, 26, 40, 91], the research community has developed techniques for addressing each of these concerns in isolation, but addressing these concerns in combination has proved difficult.

For example, in the case of privacy, Differential Privacy (DP) has been employed for ADM that uses machine learning (ML) [28]. However, deploying DP can disproportionately harm an ML model's performance for minority groups, introducing a tradeoff between privacy guarantees and fairness across demographic subpopulations for ADM systems that use machine learning [5]. Similarly, techniques that mitigate certain types of tampering also entail unintended risks, such as increased vulnerability to privacy attacks, further illustrating the complex interplay among the desired attributes of ADM systems [98]. These fundamental tensions must be addressed to deploy ADM systems that meet user expectations. However, current research has not delved into how users perceive these tensions and how they prioritize tradeoffs.

In summary, the deployability of ADM systems is intertwined with privacy, fairness, and adversarial tampering concerns, and efforts to address one concern can come at the cost of another.

Since mitigating one concern may exacerbate another, the question arises as to which objectives should be prioritized and who should make these decisions. Among all the stakeholders involved in ADM systems, the people in a society who will be implicitly or explicitly affected by the ADM decisions represent the largest stakeholder group. They are affected by the outcomes of ADM systems, making their perspectives crucial for guiding their design and deployment.

Therefore, we explore peoples' acceptability and perceptions of HDM and ADM in high-stakes decision-making scenarios, and how such stakeholder groups perceive and prioritize tradeoffs among privacy, fairness, and security in these scenarios. To this end, we ask and answer the following research questions.

RQ1: What is the users' acceptability of HDM and ADM in high-stakes scenarios, and how does their acceptability differ across scenarios? *(Sect. 5).*

RQ2: What considerations, concerns, and values shape users' acceptability of HDM and ADM in high-stakes scenarios? *(Sect. 6).*

RQ3: What is the users' preference of tradeoffs among privacy, fairness, and security in high-stakes ADM scenarios? *(Sect. 7 and 8).*

RQ4: What considerations, concerns, and values shape users' perception of prioritizing the tradeoffs in high-stakes ADM scenarios? *(Sect. 9).*

In brief, while participants appreciate ADM for its efficiency and perceived objectivity in structured tasks, they value HDM for its context-appropriate judgments and ability to consider aspects beyond mere data. Participants raised concerns about potential biases in both HDM and ADM, as well as technical and human limitations. Participants prioritized tradeoffs contextually, demonstrating a willingness to accept biases, compromise on privacy, or risk adversarial manipulation to achieve desired outcomes. Their decisions were influenced by the stakes involved, the sensitivity of the information, and the potential consequences of accepting risks, as they weighed costs against benefits.

Table 1: Tradeoffs are emphasized with a ◆ to denote that using the defense makes that negative outcome worse. Unexplored cases are indicated with ◐, and where preventing that outcome is the goal of the defense is indicated N/A.

	Defended Objective		
Outcomes	Security	Privacy	Fairness
Utility Loss	◆ [70, 99]	◆ [1, 43]	◆ [68, 73]
Evasion	N/A	◆ [8, 92]	◆ [90]
Poisoning	N/A	◆ [38, 54]	◆ [14, 57, 79, 94]
Membership Inference	◆ [37, 80]	N/A	◆ [15, 88]
Attribute Inference	◐	N/A	N/A
Distribution Inference	◆ [84]	N/A	◆ [84]
Data Reconstruction	◆ [58, 104]	N/A	N/A
Discriminatory Behavior	◆ [6, 41, 53]	◆ [5, 47, 106]	N/A

2 Background

Automated decision making systems. In this paper, we refer to all systems that automate decision-making through machine learning (ML), artificial intelligence (AI), or data-driven analysis as ADM systems. ADM systems have been deployed to replace or augment HDM across varied real-world high-stakes scenarios. ADM systems are used to automatically determine credit limits [89] and automatic health risk assessments to determine coverage [10, 81]. Meanwhile, in hiring, automated resume screening and hiring systems have been deployed to filter, score, and prioritize applicants [9, 69], ultimately influencing which applicants get hired. Finally, ADM systems have also been utilized for automated risk assessment in criminal justice systems worldwide [66, 77, 102]. Notably these systems are deployed with the intention to improve decision making processes in settings that have real, material impacts on people's lives. Thus, while we can appreciate their potential to improve these systems, we must also consider their weaknesses.

ADM risks. Our selected classes of vulnerabilities and attacks have been selected because their associated defenses pose tradeoffs for the valued objectives of privacy, fairness, and tamper-resistance. We refer to attacks that adversarially impair the ADM system as *security attacks*. We emphasize, that while we use this term for brevity, we do not consider all security attacks. We limit security attacks to model manipulation via evasion or via poisoning. A malicious client can tamper with a model using *evasion* by manipulating inputs to the model to force a misclassification, such as to force approval of loan that would not be approved otherwise [55]. Similarly, a model can be manipulated by adding tampered data records to its training data, leading it to perform poorly on some data records, or by manipulating its predictions during inference for specific inputs [34]. This technique is known as *poisoning* and could enable a hiring manager to manipulate an ADM such that it prioritizes scoring certain applicants higher than others.

We refer to attacks which impair the privacy of the training data as *privacy risks*. We place privacy attacks that target the queries to a model outside the scope of this work. Our set of relevant privacy attacks are membership inference, attribute inference, distribution inference, and data reconstruction. In a *membership inference* attack an adversary's goal is to determine whether a target data record was used to train the model [40, 78]. For example, a successful membership inference attack could reveal if a given patient's data was used to train an ADM that determines if a patient has depression, therefore revealing the patient's diagnosis. An *attribute inference* attack refers to when an adversary successfully determines the values of a sensitive attribute associated with the training data set [31]. For example, an attacker could learn that all of the data in the dataset came from women. In the case of a *distribution inference* attack, the adversary's goal is to determine the rate of representation of a property across the dataset [83]. For instance, they may wish to learn the ratio of men to women as their distribution inference attack. Finally, in a *data reconstruction* attack, an adversary is successful if they are able to reconstruct training data from their access to the model, typically via querying the model [12, 13, 62, 107].

Finally, we use *fairness risks* as the term to broadly encompass different discriminatory outcomes associated with a model [75]. For example, a creditworthiness algorithm and how it was used to determine credit limits for credit card applicants. One such system made determinations that women found unfair because they received lower credit limits than men with similar financial histories [89].

Tradeoffs among defenses and risks. We use *tradeoffs* to refer to when techniques to preserve one desirable property, namely to mitigate a security, privacy, or fairness risk, do so at the detriment

of another desirable property. For example, a privacy risk mitigation may hamper the ADM system's accuracy in its decision making [19]. Table 1 provides an overview of the tradeoffs and indicates which defense techniques for these objectives have an associated negative outcome [27]. Real-world deployment of ADM requires practitioners to account for these tradeoffs and find the best strategy to balance them, ensuring compliance with the law. Otherwise, the institution might face legal consequences.

Law and policy. New laws and policies are being developed to ensure that ADM systems are safe, fair, resilient against adversarial manipulation, maintain public trust, and comply with existing privacy regulations. For example, the United States is implementing AI regulations at the state level, including Colorado, Texas, and Utah [85–87], as well as domain-specific regulations, such as the Illinois Artificial Intelligence Video Interview Act for hiring [42]. Beyond the United States, the European Union (EU) has established a regulation regarding semi or fully ADM systems [17]. The EU Act identifies several high-risk applications, including medical risk assessment (Annex III, point 5(c)), candidate screening for jobs (Annex III, point 4(a)), criminal risk assessment (Annex III, point 6(a)), and financial decision-making (Annex III, point 5) [17]. When practitioners deploy automation in high-risk domains, they need to ensure compliance with legal and regulatory requirements. Therefore, it is important to understand the perspectives of the people ADM systems will directly impact, so that ADM design choices meet not only technical metrics, but also people's priorities.

3 Related Work

Perceptions of privacy, security and fairness have each been considered in isolation in the literature. While we bring all of these objectives together in our study, we first highlight users' perceptions of the objectives in isolation.

In ADM systems, privacy risks arise at multiple levels, including risks to individuals in the training data and risks related to the disclosure of aggregate data properties [78, 84]. Protections against these risks include techniques such as differential privacy, secure multiparty computation, and homomorphic encryption, which when applied are each known to have some impact on how users understand and evaluate a system [20, 23, 25, 30, 45, 46, 60, 61, 97]. The trust and acceptability of these privacy protection techniques depends on both their formal guarantees as well as on the application context, such as the domain being healthcare, finance, or surveillance [45, 56, 105]. In addition to the technical guarantees and application context, user perceptions of privacy also vary across cultural and demographic contexts, reflecting not only individual preferences, but also broader social norms [2, 11, 35, 44, 51, 63].

Fairness in ADM systems is similarly highly context-dependent. In low-stakes domains such as social media, users evaluate fairness in terms of visibility and opportunity allocation [16, 52]. In contrast, in high-stakes settings like criminal justice, fairness is judged by focusing on whether decisions are made without bias and are proportional to the offenses committed [7, 36]. Broadly, fairness perceptions are shaped by both algorithmic factors, such as accuracy, and human factors, including moral intuitions and subjective beliefs, as well as sociodemographic characteristics [33, 82, 95]. Additionally, users showed disagreement on which fairness metrics are appropriate, such as demographic parity, equal opportunity, and equalized odds [33, 75, 81]. While algorithmic fairness metrics capture different notions of fairness, users tend to prioritize accuracy, particularly in high-stakes contexts such as medical decision-making [81], and prefer predictive parity when sensitive attributes are involved [75]. Overall, this suggests that fairness is not a fixed property of a system but a socially constructed perception that can vary based on scenarios and stakeholders [3, 50, 67, 82].

Finally, beyond fairness and privacy we have the notion of resistance to adversarial manipulation. Unlike the prior two objectives which have plentiful literature on perceptions, prior work on adversarial manipulation is largely algorithmic and related to technical measures, with limited attention to the perceptions of those who are negatively impacted by failures to prevent such manipulations. Existing studies on resistance to adversarial manipulation, or robustness, perceptions are largely limited to domains such as recommendation systems and autonomous vehicles [64, 96].

Overall, while existing studies focused on users' perceptions of an objective in isolation, deployed systems must consider the objectives collectively, since achieving one comes at the cost of losing another. Further, it is important to understand whether peoples' priorities are the same across different ADM systems or whether they dynamically re-prioritize competing values across scenarios before there can be a clear process on how to select among these objectives when producing a real system for deployment or when designing techniques with different tradeoffs in research.

4 Methods

We employ a quasi-experimental design to investigate how participants' perception and acceptance of HDM and ADM differ across four scenarios. For these scenarios, we also explore how participants prioritize and perceive different tradeoffs. See Appendix B and C for our ethical considerations and positionality statement.

4.1 Study Components

Overview of the study. As an overview, we present how a participant would experience the flow of our study (see study instrument in Appendix D). First, participants were randomly assigned to one of four application domains, which we term our scenarios. Table 2 overviews our scenarios. Participants were asked to indicate their preference for HDM and ADM in that scenario on a scale from completely unacceptable to completely acceptable and explain in a free-form response. Second, participants were asked to give their preference in that scenario for one set of our two sets of tradeoffs by indicating a selection and providing a free-form explanation. Finally, all participants answered the same demographic questions.

Scenarios and tradeoffs. For our selected scenarios, we focus on cases where deploying an ADM would affect people's lives, safety, and fundamental rights. We selected our scenarios based on the real-world ADM applications in Section 2 and the high-risk settings in the EU AI Act. The resulting scenarios are *medical insurance, job screening, prison sentencing, and mortgage approval.*

Among the known tradeoffs from Table 1, we select those where the risk of a negative outcome increases when the defense is deployed. Note there are positive relationships between objectives and a defense. For example, defending privacy uses techniques that

Table 2: The following is an overview of our scenarios, where Company A makes determinations about the given scenario. Company B has made an ADM system to assess an aspect relevant to Company A's decisions.

Scenario Domain	Company A	Company B's ADM	Sensitive Attribute
Medical Insurance	High/Low insurance premium	Risk of heart disease	Mental health issues
Job Hiring	High/Low chance of hiring	Risk of bad fit hire	Criminal background
Prison	Long/Short prison sentence	Risk of re-offending	Wealthy inmates
Mortgage	High/Low interest rate	Risk of defaulting	Minority subgroup

directly counteract membership inference, attribute inference, distribution inference, and data reconstruction. The tradeoffs in our study, therefore, include the following eight combinations. First, defending fairness versus each of utility loss, membership inference, distribution inference, and security risk. Second, defending security risk versus each of utility loss, membership inference, distribution inference, and data reconstruction. These correspond to our two sets, with each participant receiving only one of the sets. To effectively communicate the tradeoffs to participants, we have used terms that are participant-friendly. The term accuracy is presented as a correct result. We describe bias as treating one group of people more favorably than another. In the case of fairness and accuracy, the participants received a set of options, where being favored as a brunette is the proxy for bias. Choices would include "The [ADM] strongly favors brunettes; otherwise, it provides the correct chance of being hired", "The [ADM] moderately favors brunettes, and it provides slightly incorrect chance of being hired", "The [ADM] slightly favors brunettes, and it provides a moderately incorrect chance of being hired", and "The [ADM] does not favor brunettes, and it provides a highly incorrect chance of being hired'. In the case of membership inference, we refer to an employee's ability to determine whether an individual's sensitive data was used in training the ADM. Distribution inference indicates an employee's capacity to estimate the number of women in the training dataset. We define adversarial manipulation as an adversary modifying the ADM to produce incorrect results rather than true ones. Finally, data reconstruction is the process of recovering sensitive information that was part of the ADM system's training data.

Pilot. We piloted the survey with individuals from both computing science and non-computing science backgrounds to assess how well the scenarios were understood before the final survey. In the first iteration of the survey, we used a five-point scale to indicate a strict preference for protecting one risk over the other, a slight preference, or no preference. Based on feedback, we revised the format to remove the neutral option (no preference) and present explicit paired tradeoffs such as (high, none) and (medium, low), which encouraged participants to make a choice. We then updated our scale and provided additional text and context for the participants.

4.2 Participants

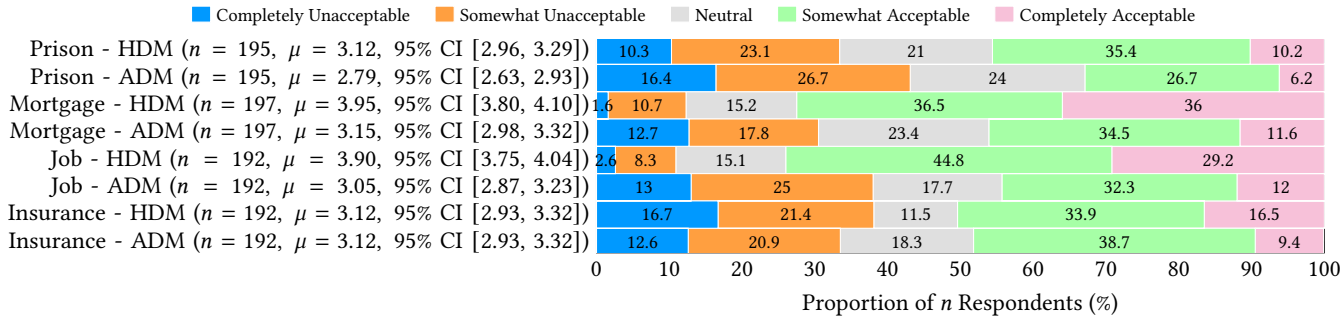
Participants were recruited via Prolific from April 2025 to August 2025. To mitigate bot interference, Prolific deploys extensive bot and AI-agent detection tools to prevent non-human responses [24]. To further ensure data quality, two researchers manually reviewed all free-form responses and discarded responses that were disconnected from the questions. We discarded 10 responses in total, resulting in 777 completed responses. Participants spent an average of 15 minutes on our study and received $3 USD. Participants were permitted to withdraw from the study at any time and could skip any question they chose.

All participants were located in the United States, and we identified them as *experiential experts* [100], people living the experience, or associated with someone living the experience. We recruit participants with various demographic attributes and who are members of a society that may be explicitly or implicitly affected by ADM in domains such as hiring, mortgages, insurance, and criminal justice. We do not know which of these they have explicitly or implicitly experienced in each setting in their lives. Among the total $N = 777$ participants, 48% indicated they were women, 51% indicated they were men, and 1% indicated they were non-binary. Regarding minority group representation, 43% indicated they belonged to a minority group, 54% indicated they did not, and 3% preferred not to answer. Participants ages were distributed as follows with 13% between 18-25 years old, 19% between 26-35, 17% between 36-45, 19% between 46-55, 28% between 56-65, and 4% preferred not to disclose their age. In terms of educational background, less than 1% of participants had not completed high school or the equivalent, 10% had a high school diploma or equivalent, 3% had some non-university education beyond high school, 20% had some college without obtaining a degree, 44% earned a bachelor's or associate degree, 17% had a master's degree, 6% a doctoral degree, and 1% preferred not to disclose their educational background. Regarding employment as ML practitioners, 86% reported not working in the field, 12% indicated they do, and 2% preferred not to answer.

4.3 Analysis Methods

For our quantitative analysis, we used the Wilcoxon signed-rank test, a non-parametric test for paired data, as each participant evaluated both HDM and ADM. Further, we report two effect sizes, the rank-biserial correlation (r_{rb}) and the standardized effect size r (derived from the test statistic Z). To examine participants' preferences across the four scenarios, we used the Kruskal–Wallis test and when it indicated a significant difference, we performed Dunn's post hoc tests with Bonferroni correction to identify pairwise differences. For significant pairs, we used the Mann–Whitney U test to determine the direction of the effect, such as which condition yields higher acceptability ratings for HDM versus ADM in a given scenario. Collectively, the above analysis is used to address RQ1. We conduct the Chi-Square test on our categorical data and report descriptive statistics corresponding to our tradeoff pairs. Within

Figure 1: The distribution of preferences for HDM and ADM across all scenarios. Acceptability is measured on a five-point scale, and each segment corresponds to the proportion of participants who selected that level of acceptability. The sample size ($n < N$) is provided and for interpretability we include a mean acceptability score (μ) and confidence interval (CI) for each condition.



each response set, two options prioritize one objective, and the remaining two prioritize the alternative objective in the tradeoffs. Accordingly, we summarize trends by considering the proportion of participants selecting options that prioritize each objective which we use to address RQ3.

Finally, we conducted qualitative content analysis over the free-form responses to address RQ2 and RQ4. In the first coding cycle, we applied descriptive coding [74] and maintained an evolving codebook. Two researchers independently conducted the initial coding on the same data using the codebook and updated it as needed. This allowed us to find codes inductively from participants' responses. After coding each question, one researcher identified any response-level discrepancies in the coded data to identify inconsistencies, which were resolved through discussion to reach consensus. When the two researchers could not reach a consensus, a third researcher was included. In the second coding cycle, we employed focused coding [74] to refine and integrate the initial codes. Finally, we collaboratively grouped our codes into broader categories based on semantic similarity and conceptual relationships, and derived higher-level themes by identifying patterns and connections across the categories.

4.4 Limitations

Our scenarios do not encompass all possible types of high-risk ADM. Rather, the ADM systems selected were chosen for their potential to have significant consequences in end-users' lives. Our participants are WEIRD (Western, Educated, Industrialized, Rich, and Democratic) [76]. Consequently, our participant sample is not representative of the global population. The study relies on self-reported perceptions, which can be influenced by social desirability bias [71], misunderstandings of the scenarios, or individual differences in interpreting the questions. Further, key concepts such as fairness were provided to participants with examples which may not perfectly represent all aspects of the underlying concepts. However, our pilot with individuals from both relevant computing science backgrounds and from non-computing backgrounds aided the representations efficacy for the scenarios. Additionally, in the context of describing bias, we deliberately chose *"favoring"*, over wording with negative connotations, to limit social desirability bias to the extent possible. Alternative framings may elicit different responses. Finally, participants evaluated hypothetical scenarios, which may not fully capture the complexity or emotional weight associated with actual high-risk ADM cases. Ultimately, the tradeoffs presented in the survey may simplify the complex technical and ethical considerations involved in real-world ADM.

5 Acceptability Results for HDM and ADM

Overall, participants responses tended to be more positive towards the HDM than to the ADM when considering the scenario prior to considering any defense tradeoffs. To provide intuitive insight into this trend, we present descriptive statistics before reporting our statistical test results within and across scenarios. See Figure 1 for the participants' acceptability responses. We also provide a graph of the mean score and 95% confidence intervals as Figure 10 in the Appendix F.

We observed a higher proportions of participants select "Somewhat Acceptable" to "Completely Acceptable," for HDM than for ADM prior to applying a defense. For example, HDM received positive acceptability ratings from 45.7% of participants in the Prison sentencing scenario, 72.5% in the Mortgage scenario, 75.8% in the Job applications scenario, and 49% in the Insurance scenario. In comparison, ADM received positive acceptability ratings from 32.9% of participants in the Prison scenario, 46.2% in the Mortgage scenario, 44.3% in the Job scenario, and 48% in the Insurance scenario.

For acceptability within the scenarios, we use the related-samples Wilcoxon Signed-Rank test with a significance threshold of $\alpha = 0.05$ which showed a significant preference for HDM over ADM in all our scenarios except Insurance. In the Mortgage and Job scenarios, participants showed statistically significant preferences for HDM, with large effect sizes: $r_{rb} = .59$ and .57 respectively, and $r = -.49$ for both. A preference for HDM was also observed in the Prison scenario, with small effect sizes: $r_{rb} = .17$ and $r = -.21$.

HDM across scenarios. Using a Kruskal-Wallis test revealed that participants' acceptability of HDM varies significantly when considering the different scenarios with a test statistic of 79.359 and $p < 0.001$. Consequently, we conducted post hoc pairwise comparisons to identify the specific pairs with significant differences. The following pairs had significant differences: Prison - Job, Prison - Mortgage, Insurance - Job, and Insurance - Mortgage. To further explore these differences, the Mann-Whitney U tests revealed that

HDM was rated significantly higher in the Job and Mortgage scenarios when compared with Prison and Insurance scenarios.

ADM across scenarios. Similar to HDM, the Kruskal-Wallis test revealed a significant overall difference in the acceptability of ADM across the four scenarios, with a test statistic of 10.527 and $p = 0.015$. Follow-up post hoc pairwise comparisons showed that the differences arose when the Prison scenario was compared against the Job and Mortgage scenarios. The Mann–Whitney U test showed that ADM was viewed as significantly more acceptable in the Job and Mortgage scenarios compared to the Prison scenario. No other scenario pairs exhibited significant differences.

Acceptability by demographics. We examined whether decision-making preferences differed by the following participant attributes: gender, minority status, and ML practitioner status. For gender, we compared only men and women, as the number of participants from other gender groups was too small for our statistical analysis.

For HDM, we found no significant differences for any of these attributes. For ADM, we found a significant difference by minority status in the Prison scenario ($p = 0.045, U = 3576.5, z = -2.0$). The non-minority group showed greater acceptability of ADM than the minority group, with a higher mean rank (100.99 vs. 85.35). No other scenarios showed significant differences by minority status. We also found a significant difference by ML practitioner status in the Insurance scenario ($p = 0.011, U = 1824.5, z = -2.533$). ML practitioners had a higher mean rank than non-practitioners (116.48 vs. 90.62), suggesting greater acceptance of ADM for this domain.

6 Findings for HDM and ADM Considerations

From our qualitative analysis of free-form responses, we identify aspects that shape participants' acceptance of HDM and ADM. We present participants' considerations, concerns, and values here organized into advantages and disadvantages.

6.1 Perceived Advantages of HDM

Participants highlighted several benefits of HDM, with specific reasoning reflecting the needs and expectations of the scenarios.

Participants value looking beyond the numbers to make informed decisions. For all four scenarios, participants valued HDM when

> "assessing factors that require a nuanced understanding of human behavior and individual circumstances ... that a purely data-driven approach might miss, leading to a more holistic and potentially fairer assessment for individuals with complex personal histories ..." [P247, Prison].

Participants also identified scenario-specific advantages, such as in hiring, "... there are soft skills and interpersonal skills that can only be determined by a human ..."[P586, Job]. Further, participants expressed the need for human judgment to recognize "grey areas" in the Mortgage scenario where "... younger adults, immigrants, or gig workers, may simply not have long traditional credit history ..." [P373, Mortgage]. In such situations, "... [HDM] can override strict data-based decisions" [P473, Mortgage]. Further, humans can demonstrate compassion and consider "how an individual may have gotten into a situation in the past that could make them appear high-risk, but they are now in a better position" [P477, Mortgage].

Participants value the human experience that guides informed decision-making. Personal narratives, social and cultural experiences, work experience, and discerning observations are mentioned as enhancing decision-making. For instance, in the Prison sentencing scenario, "some employees may have insight into the risk of re-offending based on their personal knowledge of situations" [P174]. In the Job scenario, participants noted that "employees ... have likely been in the same shoes as the applicant at one point, so they may be well suited to assess an applicant's quality" [P294, Job]. Across all scenarios, participants emphasized the existence of aspects that require human evaluation. Specifically, when the decisions might significantly impact individuals' lives, "... should be evaluated by people ..."[P767, Prison].

In summary, participants expressed that HDM has critical flexibility, unlike rigid rule-based systems, and enabling nuanced understanding and context-appropriate decisions.

6.2 Perceived Disadvantages of HDM

Participants identified consistent disadvantages of HDM across scenarios, though the manifestation of harm was scenario-dependent.

Participants raised concerns that social and personal biases compromise the fairness of HDM. Across all the scenarios, participants noted that using HDM to assess risk can be detrimental,

> "when decisions are influenced by personal biases ... such as when an employee unintentionally favors or discriminates against particular groups based on race, gender, or appearance ..." [P56, Mortgage]

Social prejudices can have significant consequences. In the Prison scenario, participants expressed concerns that individual life experiences can also impact HDM, such as a "... family member having been a victim of a crime ..." [P626, Prison]. Participants' concerns extend beyond personal benefits to include pressure from managerial priorities. In the Mortgage and Insurance scenarios, participants expressed concern that due to "... pressure to meet certain quotas, their assessments could be inconsistent, unfair, and potentially discriminatory ..." [P80, Insurance]. There were also concerns for affinity bias, such as "... an [HDM] unconsciously prefers candidates who share similar backgrounds, communication styles, or educational paths ... [P336, Job]," which can lead to unfair hiring practices and reduce organizations' diversity

Participants are concerned about inaccurate judgment, resulting from human limitations. Participants described different dynamics of human emotion, such as sympathy, anger, and moral outrage, and how each can affect decision-making across scenarios. In the Insurance and Mortgage scenario, "... the client could get a higher rate if the employee has a negative perception towards the customer" [P492, Mortgage] while in the Prison scenario, "... due to our sympathetic nature we might make the wrong decisions ..." [P210]. Participants also raised the issue of how certain work conditions can affect employees' emotional well-being and impact decision-making. Apart from the emotional influence, "factors like

the employee's mood on a particular day, their personal experiences, or even subjective interpretations of behavior could lead to inconsistencies and unfair outcomes." [P245, Mortgage].

6.3 Perceived Benefits of ADM

In all scenarios, we observed a similar perception of how ADM can be beneficial in decision-making. Overall, participants' responses in this regard were not specific, focusing on ADM's generic properties rather than on how it can benefit a particular scenario. Below, we discuss the themes we found regarding the benefits of ADM.

Participants believe that the efficiency provided by ADM will positively impact scalability. Participants emphasized that ADM enhances the efficiency of organizational processes. ADM can save time, resources, and fairness compared to HDM, thus, "...is beneficial when there is a large volume of applications to process quickly and consistently" [P497, Mortgage]. Overall, ADM is described as a tool that enhances performance through faster processing and handling large-scale data efficiently.

Participants value ADM for its perceived objectivity in decision-making. Participants perceived ADM that promotes fairness in decision-making, with one participant noting that "[ADM] is not racist. Everyone would be treated equally and fairly" [P481, Mortgage]. Unlike HDM, ADM does not experience any emotion, thus "it would take emotions out of the equation ..." [P80, Mortgage]. Especially in the Prison scenario, "it could be beneficial since there will not be room for negative emotions towards the prisoner, and will focus on the facts provided "[P649, Prison].

Participants perceive that ADM systems make data-driven decisions, which are accurate and reliable. Participants perceived ADM as reliable, ensuring that decisions are based solely on relevant facts related to the task and thereby enhancing stakeholder confidence in the outcomes, as one participant articulated in Job scenario, "the [ADM] will select applicants based on their experience and predict that they will perform well when hired ..." [P105, Job]. Participants also emphasized the ADM's ability to detect subtle correlations or anomalies that humans may overlook, reinforcing perceptions of higher accuracy, as highlighted in the following participant quote in the Insurance scenario, "...it could also spot hidden patterns in my health that even a doctor might miss" [P386, Insurance].

Participants believe ADM systems are suitable for repetitive tasks which are structured and unambiguous. ADM were viewed as well suited for straightforward tasks where decisions follow clearly defined requirements. For example, "finding the bare minimum of the job requirements, like education" [P509, Job]. Participants felt that ADM could efficiently and consistently handle these tasks because they do not demand ethical interpretation, contextual understanding, or sensitivity to individual circumstances. As a result, ADM was seen as an appropriate tool for automating screening processes in which fairness and accuracy depend primarily on applying fixed rules such as, "when a crime is clear cut, well-documented and has a lot of strong evidence" [P316, Prison].

6.4 Perceived Disadvantages of ADM

Below, we discuss the themes that we have found and describe how they produce disadvantages across all four ADM scenarios.

Participants are concerned that technical limitations of ADM compromise the fairness and reliability of its decisions. Participants reported scenario-specific correlations that might negatively influence decisions, such as certain illnesses leading to higher insurance premiums, living in certain neighborhoods increasing the likelihood of recidivism, and past criminality predicting future hiring. Participants highlighted that ADM are likely to perpetuate historical norms of discrimination, as "the [ADM] might learn to treat similar applicants as higher risk, thereby perpetuating discrimination ..." [P257, Mortgage]; which mirrors mortgage outcomes known to include structural racism [49]. Additionally, participants expressed that accuracy depends on the training dataset; for example, "...if the [ADM] was trained predominantly on data from middle-aged men, it might inaccurately assess risk in women or younger applicants ..." [P219, Insurance]. In fact, participants noted that ADM might "...automates discrimination while hiding behind 'neutral' algorithms, violating ethical and legal standards" [P314, Job].

Participants expressed concern that the lack of emotional intelligence in ADM leads to unreliable decision-making. Participants expressed concern that ADM systems cannot interpret emotional, moral, or social nuances. They emphasized contexts where judgments must reflect not only facts, but also the human aspects of behavior and change as even when records indicate a high risk of recidivism or non-payment,"People can change, and patterns can be broken" [P604, Prison].

Lack of governance and ethical accountability undermines participant trust in ADM. Participants raised concerns about the lack of oversight and transparency, as well as the potential for corruption or privacy violations since the "[ADM] could completely backfire if the wrong information is used and nobody notices" [P383, Prison]. Furthermore, it is "...difficult to understand why a particular risk assessment was given, hindering transparency and the ability to appeal an unfair decision" [P245, Prison]. This aligns with concerns that, without proper governance, it is unclear who is accountable for errors or harm.

Participants showed concern about deploying ADM in high-stakes and sensitive scenarios. The sensitivity of the decisions and the data involved in high-stakes scenarios produces a negative perception of ADM. Decisions which impacted peoples lives and well-being were described as concerning settings for ADM. For example, participants mentioned that "using [ADM] for human health is dangerous and always a bad idea" [P120, Prison] and even stated that "any final decision made by a machine is a bad decision" [P746, Mortgage].

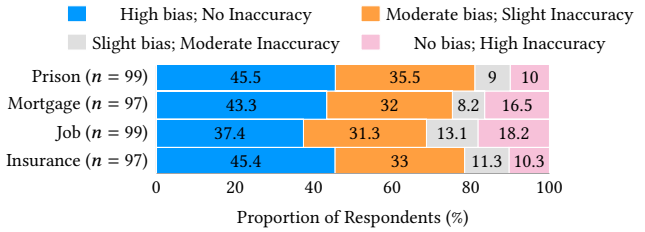
7 Tradeoff Preference Within Scenarios

We test whether participants significantly chose one option more than others using a chi-square test (details in Appendix H) and found the following significant results. In the fairness versus accuracy tradeoff, excluding the Job scenario, participants chose the accuracy-maximizing option. In the fairness versus MI and fairness versus DI tradeoffs, participants choose to prioritize fairness in

the Prison and Job scenarios. In the fairness versus security tradeoff, participants chose the highest level of security in the Prison and Mortgage scenarios. However, in the security versus accuracy tradeoff, excluding the Prison scenario, participants chose the accuracy-maximizing option.

Below we present the participants' preference distribution within scenarios. Recall that we cluster minimizing the risks of membership and distribution inference as "prioritizing privacy". All percentages are for the $n < N$ participants assigned to that scenario. The following results show that participants' preferences are dependent on the specific scenario they are presented with.

Figure 2: Preferences of fairness versus accuracy along with sample size ($n < N$) for each scenario.



7.1 Fairness versus Accuracy

In the Prison scenario, 80% of participants chose to prioritize accuracy, with 45% accepting strong bias and 35% accepting moderate bias as the consequence of their decision; the distribution is shown in Figure 2. In the Mortgage scenario, 75% preferred accuracy-maximizing options, with 43% accepting high bias and 32% accepting moderate bias as the consequence. The same trend appeared for the Insurance scenario where 78% opted for accuracy-maximizing choices, with the consequences split as 45% for the strong-bias option and 33% for the moderate-bias option. In the Job scenario, 68% chose the accuracy-favoring options, and for the consequences, 37% accepted high bias while 31% accepted moderate bias.

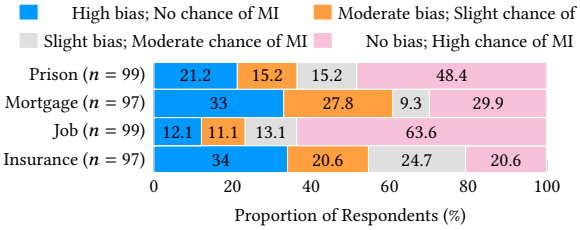
7.2 Fairness versus Membership Inference

We provide the preference distribution in Figure 3. The fairness-maximizing option was selected by 63% of participants in the Prison scenario; and within that 48% were willing to accept a high risk of membership inference, with an additional 15% accepting a moderate risk of membership inference. Similarly in the Job scenario, 76%, preferred fairness-maximizing options. Breaking down the consequences, 63% of selected a significant increase in membership inference risk and 13% selected a moderate risk of membership inference. Conversely, participants prioritized privacy in the Mortgage and Insurance scenarios. In the Mortgage scenario, 60% of participants prioritized privacy. The high bias was preferred by 33% while 27% accepted moderate bias in decision-making as their consequences. We observe a similar pattern in the Insurance scenario, where 54% preferred privacy, and for consequences, 34% accepted higher bias and 27% preferred moderate bias.

7.3 Fairness versus Distribution Inference

In the Prison scenario, 73% selected fairness-maximizing options, with 63% accepting a high chance of distribution inference and 10% selecting the moderate risk of distribution inference, as shown in Figure 4. Similarly, in the Job scenario, 67% of respondents prefer fairness-prioritizing options, where 52% preferred the high risk of distribution inference and 15% preferred the moderate risk. In the Mortgage approval scenario, 57% of participants chose privacy-preferred options, broken down as 32% that accepted a strong bias towards one group, and 25% accepted a moderate bias. In the Insurance scenario, 56% selected privacy-maximizing options, 26% accepted high bias, and 29% accepted moderate bias.

Figure 3: Preferences of fairness versus membership inference (MI) with sample size ($n < N$) for each scenario.



7.4 Fairness versus Security

Figure 5 shows the participants' preference distribution. In the Prison scenario, 60% of participants preferred security. Broken down by consequences, 46% accepted high bias and 14% accepted moderate bias as the consequences. Security options were similarly preferred by 67% in the Mortgage scenario and high bias being accepted by 42% and 25% accepting moderate bias. In the Insurance scenario, 55% selected the option prioritizing security, with 36% accepting high bias and 19% accepting moderate bias. Unlike the other three scenarios, in the Job scenario 51% of participants preferred fairness-maximizing options with their consequences being 32% preferred high manipulation risk and 19% preferred moderate manipulation risk.

Figure 4: Preferences of fairness versus distribution inference (DI) with sample size ($n < N$) for each scenario.

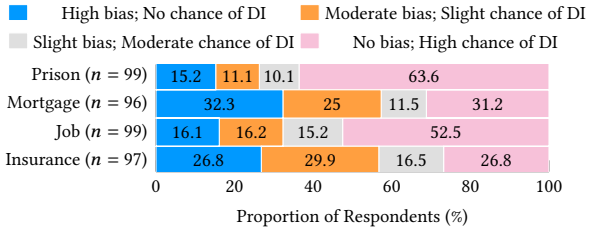


Figure 5: Preferences of fairness versus manipulation risk (MR) with sample size ($n < N$) for each scenario.

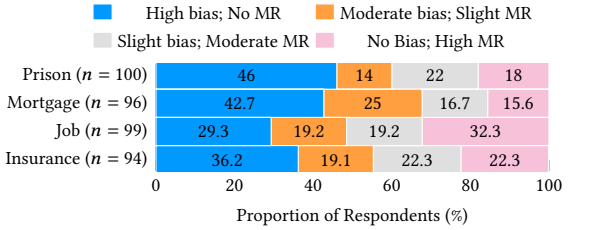
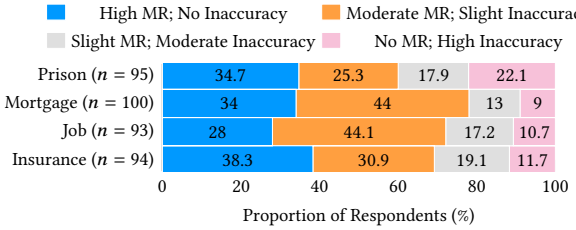


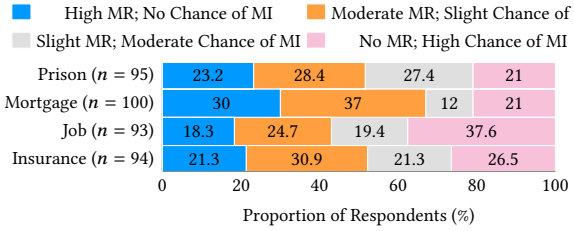
Figure 6: Preferences for manipulation risk versus accuracy with sample size ($n < N$) for each scenario.



7.5 Security versus Accuracy

In the Prison scenario, 60% of participants chose accuracy-maximizing options, with 34% accepting the risk of high manipulation and 25% accepting a moderate manipulation risk, shown in Figure 6. A similar pattern appeared in the Mortgage scenario, where 74% preferred accuracy-maximizing options, with 34% accepting high manipulation risk and 40% accepting moderate manipulation risk. The Job scenario showed a comparable trend, with 72% opting for accuracy-maximizing choices, which broken down by consequences is 28% tolerating high manipulation risk, and 44.1% moderate manipulation risk. In the Insurance scenario, 68% choose accuracy-maximizing options, which broken down is 38.3% accepted high manipulation risk, and 30% accepted moderate manipulation risk. Altogether, these results indicate a broad willingness among participants across scenarios to accept the risk of getting an inaccurate prediction if it prevented manipulation.

Figure 7: Preferences for manipulation risk (MR) versus membership inference (MI) with sample size ($n < N$).



7.6 Security versus Membership Inference

Figure 7 shows the distribution. In Mortgage, 67% chose privacy-maximizing options, with 30% willing to take a high risk and 37% willing to accept a moderate risk of manipulation. In the Prison and Insurance scenarios, preferences are split between privacy and security. In particular, the Prison scenario shows 51% preferring privacy-maximizing options, with 23% accepting a high risk and 28% accepting a moderate risk of manipulation. In the Insurance scenario, 52% selected privacy-maximizing options, with 21% choosing a high likelihood of manipulation risk and 30% selecting a moderate likelihood of manipulation risk. The Job scenario had 57% select the security-maximizing option, where 37% selected high chance of MI, and 19% selected a moderate chance of MI.

7.7 Security versus Distribution Inference

In the Prison scenario, 53% selected privacy-preferred options, 20% accepted a high risk, and 33% accepted a moderate risk of adversarial manipulation, as shown in Figure 8. A similar distribution is observed in the Mortgage scenario, where 54% selected privacy-maximizing options, 30% accepted high risk, and 24% accepted moderate risk of manipulation. The Job scenario shows a comparable pattern, with 53% selecting privacy-preferred options, 25% accepting high risk, and 28% preferring moderate risk of manipulation. In the Insurance scenario, 58.9% selected privacy-maximizing options, 22% selected high-risk, and 36% selected moderate-risk of manipulation.

7.8 Security versus Data Reconstruction

Preference distribution is shown in Figure 9. In the Mortgage scenario, 62% selected privacy-maximizing options, and correspondingly 28% accepted the high risk of manipulation and 34% accepted a moderate risk. The Job scenario had 58.7% preferring privacy, where 34% selected the high risk of manipulation and 23% selecting moderate risk. For the Insurance scenario, 55% chose privacy-maximizing options, where 26% accepted a high manipulation risk and 28% accepting a moderate risk. In the Job scenario, preferences are divided as 49% selected options with the lowest privacy risk where 21% accepted a high risk of manipulation, and 16% accepted a moderate risk of manipulation.

Figure 8: Preferences for manipulation risk (MR) versus distribution inference (DI) with sample size ($n < N$).

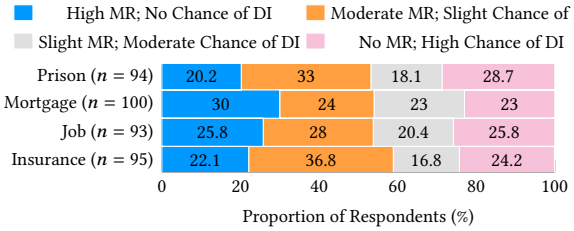
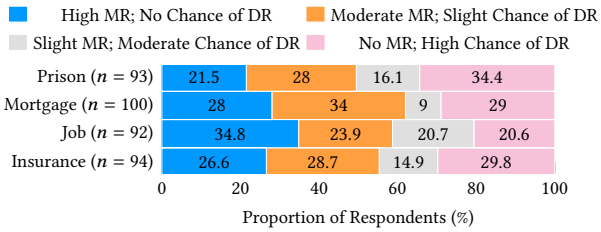


Figure 9: Preferences for manipulation risk (MR) versus data reconstruction (DR) risk with sample size ($n < N$).



8 Tradeoffs Preference Across Scenarios

To determine whether participants' preference distributions vary across scenarios, we ran the chi-square test for each tradeoff across the four scenarios. We report the tradeoffs where participants show significant differences and, for the pairs that differ (details in Appendix G). In the tradeoff between fairness and MI, participants showed significant differences across scenarios, with $p < .001$ and $\chi^2 = .377$. Specifically, the following pairs showed a significant difference: (Insurance, Job), (Insurance, Prison), (Job, Mortgage) and (Mortgage, Prison). We also found significant differences in fairness versus DI, with $p < .001$ and $\chi^2 = 41.843$. The following pairs showed a significant difference: (Insurance, Job), (Insurance,

Prison), (Job, Mortgage), and (Mortgage, Prison). In all other tradeoffs, we did not find any significant difference in preference across the scenarios.

9 Findings for Tradeoff Considerations

We use our qualitative analysis to elucidate participants' considerations when prioritizing an objective in a tradeoff. Participants' priorities among the objectives were shaped by the specific ADM context presented in each scenario. We organize our findings, the themes, by the values, concerns, and reasoning provided by participants in their freeform responses.

9.1 Concerns Beyond Technical Fairness Metrics

Participants' perceptions of fairness go beyond ensuring similar outcomes across demographic groups, incorporating privacy, accuracy, and security risks. Fairness for the participants includes proportionality, where outcomes are expected to align with an individual's risk or qualifications, with deviations perceived as unfair.

Lack of privacy protection leads to discrimination. Participants perceived that disclosing specific attributes can negatively influence decision-making, particularly when it triggers the correlation of such information with historical biases. Whether the disclosed information is directly related to the participants (e.g., membership inference) or more broadly about the training dataset (e.g., distribution inference), such disclosures were seen as a potential discriminatory factor that is "…, undermining the fairness that AI was supposed to promote" [P191, Mortgage]. For example, in the Mortgage context, participants expressed that "it is better to keep client demographic private so they have a fair chance at obtaining the loan" [P505, Mortgage]. Participants noted that, even though demographic statistics are not directly related to an individual user, they can "…lead to indirect forms of discrimination or reinforcing [demographic] stereotypes" [P403, Mortgage].

Susceptibility to manipulation is perceived as a source of discrimination. Participants emphasized that a lack of security introduces opportunities for manipulation, where "…[human] could intentionally penalize groups, leading to targeted discrimination …" [P185, Prison]. As a result, participants highlighted that "…preventing manipulation is crucial to maintain fairness …" [302, Mortgage] and preserve "system integrity" [P200, Prison]. Participants further noted that when systems are easily manipulated, they can be exploited for organizational gain, as, "if the AI is too easy to manipulate, employees will exploit it to give customers higher premiums for greater profit. This is unfair and possibly illegal" [P685, Insurance]. Overall, participants conceptualize security as more than a technical property of ADM systems, viewing it as a protection against discriminatory outcomes that arise from human manipulation of ADM systems.

Inaccuracy results in discrimination. Participants closely linked perceptions of fairness to the accuracy of ADM outputs, noting that "…if people are getting [results] that are in any way incorrect, that is unfair all around" [P99, Insurance]. This concern appeared in all decision-making contexts, with one example being: "…low-risk clients might pay too much …which is not fair to anyone" [P323, Mortgage]. In all of these contexts, inaccurate outcomes were perceived as imposing unjust burdens on individuals.

Overall, participants did not treat fairness as an isolated construct; rather, they framed accuracy, security against adversarial manipulation, and privacy as intertwined with fairness. In particular, inaccuracy, lack of security, and membership inference were perceived as direct sources of unfairness when they led to tangible harms, while distribution inference was seen as potentially triggering unconscious bias or reinforcing demographic-based stereotypes.

9.2 Contextual Influences on Prioritization

Depending on the sensitive information in the decision-making scenario, participants perceive some information as contextually appropriate to reveal, even through MI, if it aids in decision-making.

Participants believe that specific Membership inference helps make informed decisions in certain scenarios. Some participants perceived that membership inference is acceptable when it can lead to informed decision-making. For example, in the Job scenario, criminal history is perceived as a crucial piece of information. Alongside the information being crucial, such participants expressed concern that if the ADM system failed to reveal criminal history then "…the [ADM] would hire those with a criminal background, which could be bad for the work force of the company" [P564, Job].

Participants consider accepting the risk of revealing certain membership inference to ensure fairness and security. An influence on participants' prioritization was whether information disclosure was perceived as inevitable in a given decision-making context. For example, in the Prison scenario, participants noted that the justice system gathers a certain amount of information about the prisoner; thus, "wealth is easy to objectively measure in this instance" [P602, Prison]. Despite acknowledging the availability of certain attributes, participants still articulated normative boundaries, stating that they "do not think an individual's wealth should matter when delivering justice …" [P69, Prison]. Similarly, in the Job scenario participants perceived that "criminal backgrounds should be taken into consideration during the hiring process …" [P136, Job] and thus it is fine to be revealed as disclosures of criminal history are a regular practice in hiring. Overall, participants expressed a hierarchy of harms, in which violations of fairness and security were considered more critical than certain privacy risks, particularly when information disclosure was deemed contextually appropriate or unavoidable.

Participants consider accepting the risk of revealing certain distributive inference to ensure fairness and security. Participants perceived distributive inference as involving non-sensitive, aggregated information, and, in some cases, to even be a form of transparency. For example, participants framed demographic information as useful for monitoring fairness, suggesting that in the hiring context "knowing the number of women in the dataset could help the [company] monitor diversity and ensure fair hiring practices …" [P756, Job]. While there was some acknowledgment of the benefits of protecting aggregate demographic data, participants emphasized that preventing adversarial manipulation and ensuring fair outcomes were higher-order priorities. In short, a prevailing view was that "revealing demographic proportions like gender in training data might not directly harm individual applicants, but allowing

easy AI manipulation would introduce greater harm through unfair decisions and biased interest rates" [P370, Mortgage]. Overall, participants prioritized security and fairness over protecting aggregate demographic information, particularly when such information was perceived as non-sensitive and contextually appropriate to disclose.

9.3 Accepting Bias for the Greater Good

Participants wanted to prevent harm and system manipulation while benefiting stakeholders However, conditionally, participants were willing to accept some bias despite it also being a harm.

Participants show acceptance of bias if it does not imply group marginalization. Participants tended to accept bias where it was in a form of favoring a group, with the perspective that "…favoring one group is better than marginalizing the group. I would rather it be lenient than harsh …" [P750, Prison]. Favoring one group was perceived as causing minimal harm, as it primarily benefits the favored group without necessarily, directly, disadvantaging others. While our participants perceived favoritism as not as harmful as more direct discriminatory behavior, we acknowledge that the consequences are much the same.

Participants consider favoritism towards one group acceptable when it leads to benefiting the majority of stakeholders. In the accuracy versus fairness and security respectively, participants prioritized minimizing harm to the larger population. Across the scenarios, they observed that the group favored by the system or affected by the risk of adversarial manipulation accounted for only a small portion of the population. Thereby, they prioritized accuracy, since "…at least the algorithm provides correct [result] to everyone else …" [P145, Insurance]. Participants generally viewed fairness and security as important, but not at the expense of accurate decisions. For example, they acknowledge that while "…eliminating the bias is important, doing so at the cost of widespread inaccuracies …would be a greater injustice" [P245, Prison]. Similarly, in the security versus accuracy tradeoffs, when we sacrifice accuracy to achieve security, participants expressed, "…it is a zero-sum game. It does not matter if the [ADM] is impenetrable from manipulation if the [results] are highly inaccurate, because inaccuracy is what the manipulation protection is trying to protect" [P402, Mortgage].

Participants consider accepting bias due to training data to stop adversarial manipulation. In the fairness-versus-security trade-off, our findings show that participants perceived adversarial manipulation as deliberate discrimination and found it unacceptable. Participants considered it"…more acceptable for [ADMS] to have a reasonable amount of bias than to have the system susceptible to deliberate misuse …" [P223, Insurance]. Bias due to training data was considered a technical flaw beyond control, whereas human manipulation was perceived as a controllable problem and thus less acceptable.

9.4 Consequences Beyond Biased Outcomes

Participants expressed concern for consequences beyond the stated bias. They brought up that there were also implicit harms, including inaccuracy and an undermining of the systems validity.

Participants consider unfair outcomes as ADM's inaccuracy. Recall from Section 7.1, participants' view that inaccurate outputs correspond to unfair outcomes. Similarly, participants framed fairness and accuracy as inherently interconnected, stating, "if someone or something is showing any type of bias, then there is a huge issue with fairness and accuracy …" [P84, Prison]. This perspective implies that a system that embeds structural bias is not only unfair, but seen as producing unreliable and incorrect decisions.

Participants perceived the risk of manipulation as undermining the system's integrity. Participants extended their concerns about the risk of manipulation beyond unfair decision outcomes to highlight broader system-level vulnerabilities. In particular, participants expressed that "preventing manipulation by employees is critical to maintaining system integrity …" [P198, Prison], and "if the [ADM] is vulnerable to employee tampering, it undermines the entire system's …security …" [P185, Mortgage].

Participants recognize the potential risk that can come from exposing distribution inference. Participants emphasized that seemingly non-sensitive aggregate demographic statistics still introduce risks, including "…ethical issues about exposing group-level data without consent" [P367, Mortgage]. Disclosures from distribution inference were noted by participants to facilitate secondary uses, "…which would undermine efforts to ensure fairness and privacy in financial systems" [P167, Mortgage]. They further observed "…aggregated data can be misused to reinforce exclusionary practices …" [P185, Mortgage], or enable "gender profiling or unintended use of this information to adjust pricing strategies …" [P221, Insurance].

Overall, participants highlighted that both manipulation risks and aggregate data disclosures can introduce system-level issues resulting in unintended risks beyond individual decision outcomes.

9.5 Negative Impact on Stakeholders Matters

In all cases, when participants prioritize one objective over another, they consider its affect on end users, organizations, and society.

Loss of trust. Inaccuracy, bias, and the ADM system's susceptibility to manipulation lead to loss of trust in ADM systems. Inaccurate decisions may lead customers to disengage, as, "people will be upset about the incorrect interest rates" [P476, Mortgage]. Unequal treatment was also perceived as undermining public trust, particularly when such discrimination emerged from an institution whose primary goal is to ensure justice. Participants expressed that "…unfair sentencing based on looks …could negatively impact the trust in the justice system" [P202, Prison]. In hiring, when unsuitable candidates are hired, it affects the individuals who deserved the job most, as well as workforce quality and the company's reputation by "undermining trust in the hiring process" [P336, Job]. Finally, any privacy violation, either revealing membership inference, distribution inference, or an adversarial party recovering the training data, is perceived as harming trust, which "…would compromise the entire system and generate a lot of negative press about the company and the AI …" [P747, Insurance].

Reinforcing historical injustice. The consequences of membership inference were not limited to the inference itself. Rather it was perceived that membership inference "…compromises client

privacy and can deter minority applicants from applying at all" [P167, Mortgage]. Essentially, the concern was that the group of people who are historically disadvantaged by the system will be further disadvantaged in an automated way. In the hiring context, if the company focused less on privacy and allowed easy identification of applicants with criminal records, "…it could lead to …discouraging qualified individuals from applying and harming diversity and inclusion efforts" [P756, Job].

Financial loss. The prioritization of different objectives can cause financial damage to both users and the institution behind the ADM system. Across scenarios, participants framed inaccuracy as disproportionately harming those who should otherwise benefit from fair outcomes. In the Mortgage scenario this was expressed as"higher interest rates for undeserving people" [P425, Mortgage]. Incorrect risk assessments also directly affect the profitability of the institution running the ADM, as a high-risk person receiving "a lower interest rate will be costly for the company …[and] the company will not be able to profit well" [P337, Mortgage]. In hiring, "if [the company] …hire someone wrong …this is expensive and time consuming" [P535, Job]. Finally, the loss of data privacy could indirectly lead to financial harms as it "…could give competing companies extra inside information" [P717, Insurance].

Legal liability. Beyond financial impacts, participants raised concerns that inaccurate outcomes may expose institutions using ADM systems to legal consequences. Participants indicated that favoring one group is"…violating anti-discrimination laws and damaging the credibility of both the [ADM] system and the [institution]." [P671, Prison]. There are also legal concerns regarding privacy failures, where it was stated that, "…it is a HIPAA violation for the [ADM] to reveal [health] issues …" [P249, Insurance].

Societal impact. The Prison scenario in particular raised concerns from participants about the societal impact of ADM's incorrect decisions. Inaccuracy in the Prison scenario was described as life-altering, where it "would unnecessarily make an innocent person spend a longer time in prison" [P172, Prison]. Inaccuracies were seen as harmful to both individuals and society, and "if the AI were easy to manipulate, violent or dangerous criminals could be released back into the public" [P159, Prison].

10 Discussion

The following is our discussion of actions for concerns about privacy, fairness, and security in the design and governance of ADM systems; to increase awareness and mitigation of associated risks.

Folk tales of effectiveness and efficiency. Participants' perceptions of ADM focused on the benefits of using a machine, as well as showing a limited and sometimes inaccurate mental model of how ADM works. For example, participants' expressed that ADM systems handle large-scale data without error and make fact-driven decisions better than humans when the reality is more nuanced. This aligns with previous findings that non-experts use simplified, sometimes misleading mental models to describe what algorithms do and how they work, often without an accurate understanding of system limitations [29, 72, 101]. We can view these behaviors through the notion of "folk theories", which suggest that when people have limited knowledge of ADM, they are more likely to exhibit over-reliance on and have excessive confidence in the outcomes the of ADM system [39]. Consequently, while policymakers advocate for the inclusion of human oversight and expectations in the design of ADM system, it can be more challenging and often risky if the individuals possess only a limited understanding of what they are assessing. Addressing this requires accurate reporting of system limitations from ADM providers alongside better ADM literacy among both policymakers and the populace.

Sensitivity of information is enmeshed with potential harm. Our participants' reasoning in regards to sensitive information and harm is reflective of principles of privacy calculus models [18] as well as boundary regulation theory [4] and Nissenbaum's contextual integrity framework [65]. Participants conceptualized information sensitivity primarily in relation to its potential to enable discrimination and how the information is relevant to an individual. Participants were more willing to accept privacy risks when the information was perceived as less sensitive, normatively appropriate, or socially acceptable within the given context.

From a design and governance perspective, these findings suggest the need for more context-sensitive data practices in the deployment of ADM systems. One practical implication is the use of structured dataset documentation, such as datasheets for the dataset [32], that explicitly account for stakeholder perceptions and contextual expectations. Such documentation should not only describe what data are used, but also show how specific attributes may function as proxies for discrimination, how they align with societal norms in the application domain, and how users perceive their appropriateness. In high-stakes settings, where enhancing privacy may come at the cost of other system objectives, these trade-offs should be made explicit and grounded in both social acceptability and perceived cost–benefit considerations.

Deceptive linguistics is a concern for influencing participants' preferences. Our participants expressed that favoring one group is not as bad as marginalizing another, even though both are unfair outcomes. This more positive view of favoritism aligns with the notion that linguistic choices shape how participants reason about their moral judgments and risk assessments, even when the outcomes are objectively equivalent [93]. For example, recall the Amazon hiring case where the algorithm favored men [21]. Women were unfairly excluded, because of the favoritism. Thus, this is a potential deceptive communication pattern, where the framing of tradeoffs, as favoring rather than marginalizing, can influence participants' decisions, despite the similarity in terms of exclusionary outcomes.

Presented bias in comparison to existing societal bias. By using more neutral proxy attributes, such as hair color, we avoid invoking existing societal stereotypes or counter-stereotypes about who is or is not favored, minimizing confounding effects associated with protected characteristics. This choice avoids having historically social power holders be favored over historically marginalized groups and having historically marginalized groups favored over historically social power holders, which could have shifted participants' perspectives to focus on whether the presented setting aligned with their experience of the world rather than focusing on considering

our settings. This indirect and hypothetical framing created distance between participants and the sensitive issue, allowing them to first perceive the underlying principles of algorithmic discrimination without requiring them to take an explicit position on a socially charged category. Given that even our more neutral proxy attributes indicate that participants contextually prioritize among tradeoffs, the importance of context should be accounted for in all design and regulatory decisions regarding ADMs.

On the limitations of abstraction of human values. Participants valued humans' subjective understanding and the ability to interpret circumstances and social context, qualities they viewed as fundamental to fair and legitimate decision-making. While this supports human oversight as a component of trustworthy ADM systems, the divide between our participants notions of the socio-technical concepts of fairness, privacy, and security and their corresponding formal technical metrics suggests human oversight is insufficient on its own. Furthermore, human conceptions of privacy, fairness, and security are contextually mutable, while ADM systems require fixed definitions to be implemented. Thus, we, as researchers, need to develop technical definitions that align with social expectations where possible and need to stop equivocating our current technical definitions with their socio-technical counterparts.

Altogether, current abstractions of human values for ADM systems do not sufficiently align with those impacted by them. This suggests that greater consideration for the effected populace is needed before any new socio-technical metrics are even proposed if we are to achieve effective alignment of expectations and accountability. The question for policymakers, users, and implementers therefore is not just how do we design ADM systems to be human-values-centered and context-aware, but rather do the limitations of the guardrails and accountability measures required to make the system private, fair, or secure outweigh any supposed efficiencies gained by deploying the ADM system in the first place.

11 Conclusion

This work highlights that decision context, perceived risks and benefits, and consequences influence perceptions of decision-making approaches (HDM versus ADM) in high-stakes contexts. Participants generally preferred HDM due to concerns about accountability and the ability to handle complex situations, especially with long-term impacts on people's lives. The study found varying priorities for privacy, fairness, and robustness across different domains, indicating that acceptance of decision-making approaches is context-dependent. Moreover, participants viewed information sensitivity as dependent on usage rather than inherent data attributes. The findings suggest that fairness should be understood as a characteristic of system behavior rather than just a statistical measure, emphasizing the need for context-aware, human-centered design in AI systems for high-stakes situations. Rather than focusing only on technical metrics, AI systems should be evaluated based on their alignment with specific domain expectations and societal values.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS '16)*. Association for Computing Machinery (ACM), New York, NY, USA, 308–318. doi:10.1145/2976749.2978318

[2] Norah Abokhodair and Sarah Vieweg. 2016. Privacy & Social Media in the Context of the Arab Gulf. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) *(DIS '16)*. Association for Computing Machinery, New York, NY, USA, 672–683. doi:10.1145/2901790.2901873

[3] Doaa Abu-Elyounes. 2020. Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness. *Journal of Law, Technology & Policy* 20, 1 (2020), 1–33.

[4] Irwin Altman. 1975. *The Environment and Social Behavior: Privacy, Personal Space, Territory, and Crowding*. Brooks/Cole, Monterey, CA, USA.

[5] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc., 1269 Law St, San Diego CA 92109. https://proceedings.neurips.cc/paper_files/paper/2019/file/fc0de4e0396fff257ea362983c2dda5a-Paper.pdf

[6] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. 2021. Robustness May Be at Odds with Fairness: An Empirical Study on Class-Wise Accuracy. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning (Proceedings of Machine Learning Research, Vol. 148)*, Luca Bertinetto, João F. Henriques, Samuel Albanie, Michela Paganini, and Gül Varol (Eds.). Proceedings of Machine Learning Research, New York, NY, USA, 325–342. https://proceedings.mlr.press/v148/benz21a.html

[7] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173951

[8] Franziska Boenisch, Philip Sperl, and Konstantin Böttinger. 2021. Gradient Masking and the Underestimated Robustness Threats of Differential Privacy in Deep Learning. arXiv:2105.07985 [cs.CR] https://arxiv.org/abs/2105.07985

[9] Miranda Bogen and Aaron Rieke. 2018. Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. *Upturn* 7 (2018), 72 pages.

[10] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (Nov. 2019), 24 pages. doi:10.1145/3359206

[11] Zekun Cai and Aiping Xiong. 2023. Understand Users' Privacy Perception and Decision of V2X Communication in Connected Autonomous Vehicles. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 2975–2992. https://www.usenix.org/conference/usenixsecurity23/presentation/cai-zekun

[12] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. 2023. Extracting Training Data from Diffusion Models. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Anaheim, CA, 5253–5270. https://www.usenix.org/conference/usenixsecurity23/presentation/carlini

[13] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting Training Data from Large Language Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, Berkeley, California, USA, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[14] Hongyan Chang, Ta Duy Nguyen, Sasi Kumar Murakonda, Ehsan Kazemi, and Reza Shokri. 2020. On Adversarial Bias and the Robustness of Fair Machine Learning. arXiv:2006.08669 [stat.ML] https://arxiv.org/abs/2006.08669

[15] H. Chang and R. Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroSP)*. IEEE Computer Society, Piscataway, NJ, USA, 292–303. doi:10.1109/EuroSP51992.2021.00028

[16] Dasom Choi, Uichin Lee, and Hwajung Hong. 2022. "It's not wrong, but I'm quite disappointed": Toward an Inclusive Algorithmic Experience for Content Creators with Disabilities. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 593, 19 pages. doi:10.1145/3491102.3517574

[17] European Commission. 2025. The EU Artificial Intelligence Act. https://artificialintelligenceact.eu/ Accessed: 2026-04-20.

[18] Mary J Culnan and Pamela K Armstrong. 1999. Information Privacy Concerns, Procedural Fairness, and Impersonal Trust: An Empirical Investigation. *Organization science* 10, 1 (1999), 104–115.

[19] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the Compatibility of Privacy and Fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP'19 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 309–315. doi:10.1145/3314183.3323847

[20] Rachel Cummings, Gabriel Kaptchuk, and Elissa M. Redmiles. 2021. "I need a better description": An Investigation Into User Expectations For Differential Privacy. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (Virtual Event, Republic of Korea) *(CCS '21)*. Association for Computing Machinery, New York, NY, USA, 3037–3052. doi:10.1145/3460120.3485252

[21] Jeffrey Dastin. 2018. Insight - Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/

[22] Emiliano De Cristofaro. 2021. A Critical Overview of Privacy in Machine Learning . *IEEE Security & Privacy* 19, 04 (July 2021), 19–27. doi:10.1109/MSEC.2021.3076443

[23] Sergej Dechand, Alena Naiakshina, Anastasia Danilova, and Matthew Smith. 2019. In Encryption We Don't Trust: The Effect of End-to-End Encryption to the Masses on User Perception. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE Computer Society, Piscataway, NJ, USA, 401–415.

[24] George Denison. 2025. How Prolific detects bots and AI in online research. https://www.prolific.com/resources/how-prolific-detects-bots-and-ai-in-online-research Accessed: 2026-08-18.

[25] Verena Distler, Carine Lallemand, and Vincent Koenig. 2020. Making Encryption Feel Secure: Investigating How Descriptions of Encryption Impact Perceived Security. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE Computer Society, Piscataway, NJ, USA, 220–229.

[26] Julia Dressel and Hany Farid. 2018. The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances* 4, 1 (2018), eaao5580. arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.aao5580 doi:10.1126/sciadv.aao5580

[27] Vasisht Duddu, Sebastian Szyller, and N Asokan. 2024. SoK: Unintended Interactions Among Machine Learning Defenses and Risks. In *IEEE Symposium on Security and Privacy*. IEEE Computer Society, Los Alamitos, CA, USA, 2996–3014.

[28] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming*, Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 1–12.

[29] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017), 62–71. doi:10.1609/icwsm.v11i1.14898

[30] Daniel Franzen, Saskia Nuñez von Voigt, Peter Sörries, Florian Tschorsch, and Claudia Müller-Birn. 2022. Am I Private and If So, How Many? Communicating Privacy Guarantees of Differential Privacy with Risk Communication Formats. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery (ACM), New York, NY, USA, 1125–1139.

[31] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (Denver, Colorado, USA) *(CCS '15)*. Association for Computing Machinery (ACM), New York, NY, USA, 1322–1333. doi:10.1145/2810103.2813677

[32] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (Nov. 2021), 86–92. doi:10.1145/3458723

[33] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 903–912. doi:10.1145/3178876.3186138

[34] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. arXiv:1708.06733 [cs.CR] https://arxiv.org/abs/1708.06733

[35] Waseem Haider, Justin D Martin, Ali Farooq, Joni Salminen, and Bernard J Jansen. 2025. The Effect of Religiosity and Cultural Conservatism on Digital Privacy Concerns: A Comparative Longitudinal Study of Five Arab Countries. *International Communication Gazette* 88 (2025), 17480485251374883. Issue 3.

[36] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery (ACM), New York, NY, USA, 392–402. doi:10.1145/3351095.3372831

[37] Jamie Hayes. 2022. Trade-offs Between Membership Privacy & Adversarially Robust Learning. arXiv:2006.04622 [cs.LG] https://arxiv.org/abs/2006.04622

[38] Sanghyun Hong, Varun Chandrasekaran, Yiğitcan Kaya, Tudor Dumitraş, and Nicolas Papernot. 2020. On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping. arXiv:2002.11497 [cs.CR] https://arxiv.org/abs/2002.11497

[39] Michael C Horowitz and Lauren Kahn. 2024. Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts. *International Studies Quarterly* 68, 2 (06 2024), sqae020. arXiv:https://academic.oup.com/isq/article-pdf/68/2/sqae020/57132997/sqae020.pdf doi:10.1093/isq/sqae020

[40] Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S. Yu, and Xuyun Zhang. 2022. Membership Inference Attacks on Machine Learning: A Survey. *ACM Comput. Surv.* 54, 11s, Article 235 (Sept. 2022), 37 pages. doi:10.1145/3523273

[41] Yuzheng Hu, Fan Wu, Hongyang Zhang, and Han Zhao. 2023. Understanding the Impact of Adversarial Robustness on Accuracy Disparity. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, 1269 Law Street, San Diego CA 92109, Article 556, 31 pages.

[42] USA Illinois State Legislature. 2025. Illinois Artificial Intelligence Video Interview Act. http://www.ilga.gov/legislation/ilcs/ilcs3.asp?ActID=4015&ChapterID=68 Accessed: 2026-04-20.

[43] Bargav Jayaraman and David Evans. 2019. Evaluating Differentially Private Machine Learning in Practice. In *28th USENIX Security Symposium (USENIX Security 19)* (Santa Clara, CA). USENIX Association, Santa Clara, CA, 1895–1912. https://www.usenix.org/conference/usenixsecurity19/presentation/jayaraman

[44] Emiram Kablo, Melina Kleber, and Patricia Arias Cabarcos. 2025. PrivaCI in VR: Exploring Perceptions and Acceptability of Data Sharing in Virtual Reality Through Contextual Integrity. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, Santa Clara, CA, 1531–1548.

[45] Bailey Kacsmar, Kyle Tilbury, Miti Mazmudar, and Florian Kerschbaum. 2022. Caring About Sharing: User Perceptions of Multiparty Data Sharing. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Santa Clara, CA, 899–916.

[46] Farzaneh Karegar, Ala Sarah Alaqra, and Simone Fischer-Hübner. 2022. Exploring User-Suitable Metaphors for Differentially Private Data Analyses. In *Proceedings of the Eighteenth USENIX Conference on Usable Privacy and Security* (Boston, MA, USA) *(SOUPS'22)*. USENIX Association, USA, Article 10, 19 pages.

[47] Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. 2022. Disparate Vulnerability to Membership Inference Attacks. *Proceedings of Privacy Enhancing Technologies Symposium* 2022 (2022), 460–480. Issue 1.

[48] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How We Analyzed the COMPAS Recidivism Algorithm.

[49] Sean Lewis-Faupel and Nicholas Tenev. 2024. Racial and Ethnic Disparities in Mortgage Lending: New Evidence from Expanded HMDA Data. arXiv:2405.00895 [econ.GN] https://arxiv.org/abs/2405.00895

[50] Yangkun Li, Mohamed-Laid Hedia, Weizhi Ma, Hongyu Lu, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. Contextualized Fairness for Recommender Systems in Premium Scenarios. *Big Data Res.* 27, C (Feb. 2022), 7 pages. doi:10.1016/j.bdr.2021.100300

[51] Yao Li, Eugenia Ha Rim Rho, and Alfred Kobsa. 2022. Cultural Differences in the Effects of Contextual Factors and Privacy Concerns on Users' Privacy Decision on Social Networking Sites. *Behaviour & Information Technology* 41, 3 (2022), 655–677. arXiv:https://doi.org/10.1080/0144929X.2020.1831608 doi:10.1080/0144929X.2020.1831608

[52] Renkai Ma and Yubo Kou. 2022. "I'm not sure what difference is between their content and mine, other than the person itself": A Study of Fairness Perception of Content Moderation on YouTube. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 425 (Nov. 2022), 28 pages. doi:10.1145/3555150

[53] Xinsong Ma, Zekai Wang, and Weiwei Liu. 2022. On the Tradeoff Between Robustness and Fairness. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1902, 12 pages.

[54] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. 2019. Data Poisoning Against Differentially-Private Learners: Attacks and Defenses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (Macao, China) *(IJCAI'19)*. AAAI Press, California, USA, 4732–4738.

[55] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. https://openreview.net/forum?id=rJzIBfZAb

[56] A. K. M. Nuhil Mehdy, Michael D Ekstrand, Bart P Knijnenburg, and Hoda Mehrpouyan. 2021. Privacy as a Planned Behavior: Effects of Situational Factors on Privacy Perceptions and Plans. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery (ACM), New York, NY, USA, 169–178.

[57] Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. 2021. Exacerbating Algorithmic Bias Through Fairness Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 10 (May 2021), 8930–8938. https://ojs.aaai.org/index.php/AAAI/article/view/17080

[58] Felipe A. Mejia, Paul Gamble, Zigfried Hampel-Arias, Michael Lomnitz, Nina Lopatina, Lucas Tindall, and Maria Alejandra Barrios. 2019. Robust or Private? Adversarial Training Makes Models More Vulnerable to Privacy Attacks.

arXiv:1906.06449 [cs.LG] https://arxiv.org/abs/1906.06449

[59] Tom M. Mitchell. 1980. The Need for Biases in Learning Generalizations. In *Readings in Machine Learning*, Jude W. Shavlik and Thomas G. Dietterich (Eds.). Morgan Kauffman, Burlington, Massachusetts, USA, 184–191. http://www.cs.nott.ac.uk/~bsl/G52HPA/articles/Mitchell:80a.pdf Book published in 1990..

[60] Priyanka Nanayakkara, Johes Bater, Xi He, Jessica Hullman, and Jennie Rogers. 2022. Visualizing Privacy-Utility Trade-Offs in Differentially Private Data Releases. *Proceedings on Privacy Enhancing Technologies Symposium* 2022 (2022), 601–618. Issue 2. https://petsymposium.org/popets/2022/popets-2022-0058.php

[61] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. 2023. What are the Chances? Explaining the Epsilon Parameter in Differential Privacy. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association, Santa Clara, CA, 1613–1630.

[62] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable Extraction of Training Data from (Production) Language Models. arXiv:2311.17035 [cs.LG] https://arxiv.org/abs/2311.17035

[63] Sheza Naveed, Hamza Naveed, Mobin Javed, and Maryam Mustafa. 2022. "Ask this from the person who has private stuff": Privacy Perceptions, Behaviours and Beliefs Beyond W.E.I.R.D. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 105, 17 pages. doi:10.1145/3491102.3501883

[64] Taberez Neyazi, Sheryl Ng, Mitchell Hobbs, and Audrey Yue. 2023. Understanding User Interactions and Perceptions of AI Risk in Singapore. *Big Data & Society* 10 (11 2023). doi:10.1177/20539517231213823

[65] Helen Nissenbaum. 2004. Privacy as Contextual Integrity. *Washington Law Review* 79 (2004), 119–157.

[66] Marion Oswald, Jamie Grace, Sheena Urwin, and Geoffrey C. Barnes. 2018. Algorithmic Risk Assessment Policing Models: Lessons From the Durham HART Model and 'Experimental' Proportionality. *Information & Communications Technology Law* 27, 2 (2018), 223–250. arXiv:https://doi.org/10.1080/13600834.2018.1458455 doi:10.1080/13600834.2018.1458455

[67] Alessandra Parziale, Gianmario Voria, Giammaria Giordano, Gemma Catolino, Gregorio Robles, and Fabio Palomba. 2025. Contextual Fairness-Aware Practices in ML: A Cost-Effective Empirical Evaluation. arXiv:2503.15622 [cs.SE] https://arxiv.org/abs/2503.15622

[68] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. 2023. On the Incompatibility of Accuracy and Equal Opportunity. *Mach. Learn.* 113, 5 (May 2023), 2405–2434. doi:10.1007/s10994-023-06331-y

[69] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Dairness, Accountability, and Transparency*. Association for Computing Machinery (ACM), New York, NY, USA, 469–481.

[70] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. 2020. Understanding and Mitigating the Tradeoff between Robustness and Accuracy. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. Proceedings of Machine Learning Research, New York, NY, USA, Article 733, 11 pages.

[71] Elissa M Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L Mazurek. 2018. Asking for a Friend: Evaluating Response Biases in Security User Studies. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. Association for Computing Machinery (ACM), New York, NY, USA, 1238–1255.

[72] Michael Ridley. 2024. Informing Algorithmic Literacy Through User Folk Theories. *College & Research Libraries* 85, 7 (2024), 966. doi:10.5860/crl.85.7.966

[73] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical Observation of Negligible Fairness–Accuracy Trade-Offs in Machine Learning for Public Policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[74] J. Saldana. 2021. *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2455 Teller Road, Thousand Oaks, CA 91320, United States. https://books.google.com/books?id=RwcVEAAAQBAJ

[75] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery (ACM), New York, NY, USA, 99–106.

[76] Jonathan F. Schulz, Duman Barahmi-Rad, Jonathan P. Beauchamp, and Joseph Henrich. 2019. The Origins of WEIRD Psychology. *Behavioral and Brain Sciences* 42 (2019), e1. doi:10.1017/S0140525X19000342

[77] Stephane M. Shepherd, Yolonda Adams, Elizabeth McEntyre, and Roz Walker. 2014. Violence Risk Assessment in Australian Aboriginal Offender Populations: A Review of the Literature. *Psychology, Public Policy, and Law* 20, 3 (2014), 281–293.

[78] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *2017 IEEE symposium on security and privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 3–18.

[79] David Solans, Battista Biggio, and Carlos Castillo. 2020. Poisoning Attacks on Algorithmic Fairness. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part I* (Ghent, Belgium). Springer-Verlag, Berlin, Heidelberg, 162–177.

[80] Liwei Song, Reza Shokri, and Prateek Mittal. 2019. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery (ACM), New York, NY, USA, 241–257. doi:10.1145/3319535.3354211

[81] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Anchorage, AK, USA) *(KDD '19)*. Association for Computing Machinery (ACM), New York, NY, USA, 2459–2468. doi:10.1145/3292500.3330664

[82] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2022. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *Big Data & Society* 9 (10 2022). doi:10.1177/20539517221115189

[83] Anshuman Suri and David Evans. 2022. Formalizing and Estimating Distribution Inference Risks. *Proceedings on Privacy Enhancing Technologies* 2022 (2022), 528–551. Issue 4. https://petsymposium.org/popets/2022/popets-2022-0121.php

[84] A. Suri, Y. Lu, Y. Chen, and D. Evans. 2023. Dissecting Distribution Inference. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE Computer Society, Los Alamitos, CA, USA, 150–164. doi:10.1109/SaTML54575.2023.00019

[85] USA The State of Colorado. 2024. Consumer Protections for Artificial Intelligence. https://leg.colorado.gov/bills/sb24-205 Accessed: 2026-04-20.

[86] USA The state of Texas. 2025. Texas Responsible Artificial Intelligence Governance Act. https://legiscan.com/TX/text/HB149/id/3180120 Accessed: 2026-04-20.

[87] USA The State of Utah. 2024. Artificial Intelligence Amendents. https://le.utah.gov/~2024/bills/static/SB0149.html Accessed: 2026-04-20.

[88] Huan Tian, Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. 2024. When Fairness Meets Privacy: Exploring Privacy Threats in Fair Binary Classifiers via Membership Inference Attacks. arXiv:2311.03865 [cs.LG] doi:10.24963/ijcai.2024/57

[89] The New Yorks Times. 2018. *Apple Card Investigated After Gender Discrimination Complaints*. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html Accessed: 2025-04-28.

[90] Cuong Tran, Keyu Zhu, Ferdinando Fioretto, and Pascal Van Hentenryck. 2022. Fairness Increases Adversarial Vulnerability. arXiv:2211.11835 [cs.LG] https://arxiv.org/abs/2211.11835

[91] Toan Khang Trinh, Daiyang Zhang, et al. 2024. Algorithmic Fairness in Financial Decision-Making: Detection and Mitigation of Bias in Credit Scoring Applications. *Journal of Advanced Computing Systems* 4, 2 (2024), 36–49.

[92] Nurislam Tursynbek, Aleksandr Petiushko, and Ivan Oseledets. 2021. Robustness Threats of Differential Privacy. arXiv:2012.07828 [cs.LG] https://arxiv.org/abs/2012.07828

[93] Amos Tversky and Daniel Kahneman. 1981. The Framing of Decisions and the Psychology of Choice. *Science* 211, 4481 (1981), 453–458. http://www.jstor.org/stable/1685855

[94] Minh-Hao Van, Wei Du, Xintao Wu, and Aidong Lu. 2022. Poisoning Attacks on Fair Machine Learning. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part I*. Springer-Verlag, Berlin, Heidelberg, 370–386. doi:10.1007/978-3-031-00123-9_30

[95] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376813

[96] Mengyi Wei, Kyrie Zhixuan Zhou, Dongsheng Chen, Madelyn Rose Sanfilippo, Puzhen Zhang, Chuan Chen, Yu Feng, and Liqiu Meng. 2025. Understanding Risk Preference and Risk Perception When Adopting High-Risk and Low-Risk AI Technologies. *International Journal of Human–Computer Interaction* 41, 24 (2025), 15295–15310. arXiv:https://doi.org/10.1080/10447318.2025.2495844 doi:10.1080/10447318.2025.2495844

[97] Aiping Xiong, Tianhao Wang, Ninghui Li, and Somesh Jha. 2020. Towards Effective Differential Privacy Communication for Users' Data Sharing Decision and Comprehension. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 392–410.

[98] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-Aware Fenerative Adversarial Networks. In *2018 IEEE International Conference*

on Big Data (Big Data). IEEE, Piscataway, NJ, USA, 570–575.
[99] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. 2020. A Closer Look at Accuracy vs. Robustness. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Morehouse Lane, Red Hook, NY, 8588–8601. https://proceedings.neurips.cc/paper/2020/file/61d77652c97ef636343742fc3dcf3ba9-Paper.pdf
[100] Meg Young, Lassana Magassa, and Batya Friedman. 2019. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology* 21, 2 (2019), 89–103.
[101] Brita Ytre-Arne and Hallvard Moe. 2021. Folk Theories of Algorithms: Understanding Digital Irritation. *Media, Culture and Society* 43, 5 (2021), 807–824.
[102] Aleš Završnik. 2021. Algorithmic Justice: Algorithms and Big Data in Criminal Justice Settings. *European Journal of Criminology* 18, 5 (2021), 623–642.
[103] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). Proceedings of Machine Learning Research, New York, NY, USA, 7472–7482. https://proceedings.mlr.press/v97/zhang19p.html
[104] Jingyang Zhang, Yiran Chen, and Hai Li. 2022. Privacy Leakage of Adversarial Training Models in Federated Learning Systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. IEEE, Piscataway, NJ, USA, 108–114.
[105] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. 2018. User Perceptions of Smart Home IoT Privacy. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 200 (Nov. 2018), 20 pages. doi:10.1145/3274469
[106] Da Zhong, Haipei Sun, Jun Xu, Neil Gong, and Wendy Hui Wang. 2022. Understanding Disparate Effects of Membership Inference Attacks and Their Countermeasures. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security* (Nagasaki, Japan) *(ASIA CCS '22)*. Association for Computing Machinery (ACM), New York, NY, USA, 959–974. doi:10.1145/3488932.3501279
[107] Ligeng Zhu, Zhijian Liu, and Song Han. 2019. *Deep Leakage from Gradients*. Curran Associates Inc., Red Hook, NY, USA, 14774 – 14784.

A Open Science

Our survey is provided in Appendix D and Appendix E. Additional statistical analysis on human versus AI preference is provided in Appendix F, and participants' demographics are provided in Appendix F. Additionally, the detailed statistical results are in Appendix G. All materials provided in the appendices, as well as our scripts, are available in our Zenodo repository: https://doi.org/10.5281/zenodo.19804373

B Ethical Considerations

We sought IRB approval for this study, however, we acknowledge IRB approval is insufficient and took steps beyond IRB approval to conduct our study ethically. When designing this study, we the authors aimed to strike a balance between asking participants about real world scenarios and causing distress to participants. The scenarios used in this paper exist in the real world and can be the cause of real harm. It is likely that all of our participants have interacted with ADM systems and that those systems have determined their or their loved ones' access to healthcare, housing, jobs, financial services, and/or incarceration status. It is also possible that our participants have been victims of unfair and unaccountable decision making at the hands of an ADM system.

Seeking participants lived experiences about how ADM systems have affected their lives and their concerns about the risks of ADM systems was crucial to our work and we took care to present participants with fictional scenarios within real world domains to balance collecting meaningful data without causing unnecessary harm.

Our study was guided by the principle of illuminating real, lived experiences with technology and we used an unbiased recruitment system in order to get the broadest possible variety of experiences. We acknowledge that by virtue of using Prolific, some experiences were excluded.

C Positionality

Our research team consists of multi-national, academic privacy and security researchers, including usability researchers, all of whom are well versed in the tradeoffs between privacy, fairness, and security. This undoubtedly had an effect on the survey design.

Moreover, due to the current pervasiveness of ADM deployment, it is certain that every author has interacted with and been affected by decisions made by an ADM system. Our experiences with real-world ADM guided the creation of our survey, especially with regards to the different political, social, and institutional contexts of our countries of residence and origin. Therefore, the researchers on this study occupy a position between insiders and outsiders relative to the participants.

D Study Questionnaires

We have eight different variations of survey questionnaires. Comprising four scenarios and two sets of tradeoffs. The survey questionnaires with fairness tradeoffs across four scenarios are presented in Appendix D.1, D.2, D.3, and D.4 and the survey questionnaires with security tradeoffs across four scenarios are presented in Appendix D.5, D.6, D.7, and D.8. The demographic questions are listed in Appendix E

D.1 Fairness-Insurance

InsureYou, an insurance company, sets premiums based on the risk of heart disease, charging higher premiums for high-risk patients.

(1) InsureYou relies on employees to assess an applicant's risk of developing heart disease.
   (a) **Imagine you are an applicant of InsureYou. How acceptable do you find it when an employee assesses the risk of developing heart disease?**
      (i) completely unacceptable
      (ii) somewhat unacceptable
      (iii) neutral
      (iv) somewhat acceptable
      (v) completely acceptable
   (b) Please describe a situation where using employees would be beneficial for assessing the risk of developing heart disease.
   (c) Please describe a situation where using employees would be harmful for assessing the risk of developing heart disease

   MedForYou, a medical organization, collects data (such as age, sex, blood pressure, cholesterol, smoking habits, etc.) from patients with depression. MedForYou uses the collected data to develop an Artificial Intelligence (AI) to predict the likelihood of a patient developing a heart disease.
(2) **Instead of employees, the InsureYou now relies on MedForYou's AI to assess an applicant's risk of developing heart disease.**

(a) Imagine you are an applicant of InsureYou. How acceptable do you find it when an AI assesses the risk of developing heart disease?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using AI would be beneficial for assessing the risk of developing heart disease.
(c) Please describe a situation where using AI would be harmful for assessing the risk of developing heart disease

(3) **Consider that InsureYou relies on MedForYou's AI to assess an applicant's risk of developing heart disease and calculate their insurance premiums. MedForYou identified an issue with its AI where it offers lower premiums to brunettes over non-brunettes, even if they have a high risk. To address this, MedForYou modifies the AI to mitigate the issue.**
• Please describe how 'the AI modification' could be beneficial for everyone.
Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:
(a) **A side effect of modifying the AI is that it incorrectly predicts higher premiums for low risk applicants. It is impossible to completely stop favoring brunettes and still provide correct insurance premiums for low risk applicants. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?**
(i) The AI strongly favors brunettes; otherwise, it provides the correct insurance premiums.
(ii) The AI moderately favors brunettes, and it provides slightly incorrect insurance premiums.
(iii) The AI slightly favors brunettes, and it provides moderately incorrect insurance premiums.
(iv) The AI does not favor brunettes, and it provides highly incorrect insurance premiums
• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.
(b) **A side effect of modifying the AI is that InsureYou can learn whether an applicant has depression by identifying if they were part of MedForYou's data used to create the AI. It is impossible to completely stop favoring brunettes and still prevent InsureYou from identifying an applicant has depression. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?**
(i) The AI strongly favors brunettes, and it has no chance to reveal whether the applicant has depression.
(ii) The AI moderately favors brunettes, and it has a slight chance to reveal whether the applicant has depression.
(iii) The AI slightly favors brunettes, and it has a moderate chance to reveal whether the applicant has depression.
(iv) The AI does not favor brunettes, and it has a high chance to reveal whether the applicant has depression.
• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.
(c) **A side effect of modifying the AI is that InsureYou can now determine the number of women in MedForYou's data used to create the AI. It is impossible to completely stop favoring brunettes and still protect the information about the number of women. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?**
(i) The AI strongly favors brunettes, and it has no chance to reveal the number of women in the data.
(ii) The AI moderately favors brunettes, and it has a slight chance to reveal the number of women in the data.
(iii) The AI slightly favors brunettes, and it has a moderate chance to reveal the number of women in the data.
(iv) The AI does not favor brunettes, and it has a high chance to reveal the number of women in the data.
• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.
(d) **A side effect of modifying the AI is that InsureYou's employees could exploit the AI to disadvantage blue-eyed applicants by giving them a higher premiums. It is impossible to completely stop favoring brunettes and protect the AI from manipulation. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?**
(i) The AI strongly favors brunettes, and it is highly difficult to manipulate.
(ii) The AI moderately favors brunettes, and it is moderately difficult to manipulate.
(iii) The AI slightly favors brunettes, and it is slightly difficult to manipulate.
(iv) The AI does not favor brunettes, and it is not difficult to manipulate.
• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.

D.2 Fairness-Job

TechForYou, a leading technology firm, is hiring a software engineer based on candidates' likelihood of success, favoring those with higher potential.

(1) **TechForYou relies on employees to assess an applicant's quality.**
(a) Imagine you are an applicant of TechForYou. How acceptable do you find it when an employee assesses the quality?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using the employee would be beneficial for assessing the quality.
(c) Please describe a situation where using the employee would be harmful for assessing the quality.

RecruitTalent, a recruitment firm, collects resumes (including their technical skills, academic qualifications, experience, and portfolios) from applicants with criminal backgrounds. RecruitTalent uses the collected data to develop an Artificial Intelligence (AI) to predict whether an applicant will succeed at the job.

(2) **Instead of employees, the TechForYou now relies on RecruitTalent's AI to assess an applicant's quality.**
   (a) Imagine you are an applicant of TechForYou. How acceptable do you find it when an AI assesses the quality?
      (i) completely unacceptable
      (ii) somewhat unacceptable
      (iii) neutral
      (iv) somewhat acceptable
      (v) completely acceptable
   (b) Please describe a situation where using the AI would be beneficial for assessing the quality.
   (c) Please describe a situation where using the AI would be harmful for assessing the quality.

(3) **Consider that TechForYou relies on RecruitTalent's AI to assess an applicant's quality and evaluate their chance of being hired. RecruitTalent identified an issue with its AI where it offers higher chances of being hired to brunettes over non-brunettes, even if they have a low potential. To address this, RecruitTalent modifies the AI to mitigate the issue.**
   • Please describe how 'the AI modification' could be beneficial for everyone.
   Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:
   (a) **A side effect of modifying the AI is that it incorrectly predicts lower chances of being hired for high potential appli- cants. It is impossible to completely stop favoring brunettes and still provide correct chance of being hired for high po- tential applicants. In this scenario, if you were an applicant, which option would you suggest for RecruitTalent to choose?**
      (i) The AI strongly favors brunettes; otherwise, it provides the correct chance of being hired.
      (ii) The AI moderately favors brunettes, and it provides slightly incorrect chance of being hired.
      (iii) The AI slightly favors brunettes, and it provides moderately incorrect chance of being hired.
      (iv) The AI does not favor brunettes, and it provides highly incorrect chance of being hired
      • Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.
   (b) **A side effect of modifying the AI is that TechForYou can learn whether an applicant has a criminal background by identifying if they were part of RecruitTalent's data used to create the AI. It is impossible to completely stop favoring brunettes and still prevent TechForYou from identifying an applicant has a criminal background. In this scenario, if you were an applicant, which option would you suggest for RecruitTalent to choose?**
      (i) The AI strongly favors brunettes, and it has no chance to reveal whether the applicant has a criminal background.
      (ii) The AI moderately favors brunettes, and it has a slight chance to reveal whether the applicant has a criminal background.
      (iii) The AI slightly favors brunettes, and it has a moderate chance to reveal whether the applicant has a criminal background.
      (iv) The AI does not favor brunettes, and it has a high chance to reveal whether the applicant has a criminal background.
      • Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.
   (c) **A side effect of modifying the AI is that TechForYou can now determine the number of women in RecruitTalent's data used to create the AI. It is impossible to completely stop favoring brunettes and still protect the information about the number of women. In this scenario, if you were an applicant, which option would you suggest for RecruitTalent to choose?**
      (i) The AI strongly favors brunettes, and it has no chance to reveal the number of women in the data.
      (ii) The AI moderately favors brunettes, and it has a slight chance to reveal the number of women in the data.
      (iii) The AI slightly favors brunettes, and it has a moderate chance to reveal the number of women in the data.
      (iv) The AI does not favor brunettes, and it has a high chance to reveal the number of women in the data.
      • Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.
   (d) **A side effect of modifying the AI is that TechForYou's employees could exploit the AI to disadvantage blue-eyed applicants by giving them a lower chances of being hired. It is impossible to completely stop favoring brunettes and protect the AI from manipulation. In this scenario, if you were an applicant, which option would you suggest for Recruit- Talent to choose?**
      (i) The AI strongly favors brunettes, and it is highly difficult to manipulate.
      (ii) The AI moderately favors brunettes, and it is moderately difficult to manipulate.
      (iii) The AI slightly favors brunettes, and it is slightly difficult to manipulate.
      (iv) The AI does not favor brunettes, and it is not difficult to manipulate.
      • Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.

D.3 Fairness-Mortgage

BankOnYourBlock, a leading financial institution, determines mortgage interest rates based on applicants' default risk, charging higher rates for high-risk individuals.

(1) **BankOnYourBlock relies on employees to assess a client's risk of non-payment.**

(a) Imagine you are a client of BankOnYourBlock. How acceptable do you find it when an employee assesses the risk of non-payment?
  (i) completely unacceptable
  (ii) somewhat unacceptable
  (iii) neutral
  (iv) somewhat acceptable
  (v) completely acceptable
(b) Please describe a situation where using employees would be beneficial for assessing the risk of non-payment.
(c) Please describe a situation where using employees would be harmful for assessing the risk of non-payment.

InsureYou, an insurance company, collects data on past mortgage applicants from minority demographic groups, including education, dependents, marital status, income, mortgage amount, credit history, outstanding debt, savings, investments, collateral, and employment stability. InsureYou uses the collected data to develop an Artificial Intelligence (AI) to predict whether an applicant will default.

(2) **Instead of employees, the BankOnYourBlock now relies on InsureYou's AI to assess a client's risk of non-payment.**
(a) Imagine you are a client of BankOnYourBlock. How acceptable do you find it when an AI assesses the risk of non-payment?
  (i) completely unacceptable
  (ii) somewhat unacceptable
  (iii) neutral
  (iv) somewhat acceptable
  (v) completely acceptable
(b) Please describe a situation where using AI would be beneficial for assessing the risk of non-payment.
(c) Please describe a situation where using AI would be harmful for assessing the risk of non-payment

(3) **Consider that BankOnYourBlock relies on InsureYou's AI to assess a client's risk of non-payment and calculate their interest rates. InsureYou identified an issue with its AI where it offers lower interest rates to brunettes over non-brunettes, even if they have a high risk. To address this, InsureYou modifies the AI to mitigate the issue.**

• Please describe how 'the AI modification' could be beneficial for everyone.

Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:

(a) **A side effect of modifying the AI is that it incorrectly predicts higher interest rates for low risk applicants. It is impossible to completely stop favoring brunettes and still pro- vide correct interest rates for low risk applicants. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?**
  (i) The AI strongly favors brunettes; otherwise, it provides the correct interest rates.
  (ii) The AI moderately favors brunettes, and it provides slightly incorrect interest rates.
  (iii) The AI slightly favors brunettes, and it provides moderately incorrect interest rates.
  (iv) The AI does not favor brunettes, and it provides highly incorrect interest rates
  • Give an example of potential issues if InsureYou chooses a different approach than your suggestion.
(b) **A side effect of modifying the AI is that BankOnYourBlock can learn whether a client belongs to a minority demo- graphic group by identifying if they were part of InsureYou's data used to create the AI. It is impossible to completely stop favoring brunettes and still prevent BankOnYourBlock from identifying a client belongs to a minority demographic group. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?**
  (i) The AI strongly favors brunettes, and it has no chance to reveal whether the applicant belongs to a minority demographic group.
  (ii) The AI moderately favors brunettes, and it has a slight chance to reveal whether the applicant belongs to a minority demographic group.
  (iii) The AI slightly favors brunettes, and it has a moderate chance to reveal whether the applicant belongs to a minority demographic group.
  (iv) The AI does not favor brunettes, and it has a high chance to reveal whether the applicant belongs to a minority demographic group.
  • Give an example of potential issues if InsureYou chooses a different approach than your suggestion.
(c) **A side effect of modifying the AI is that BankOnYourBlock can now determine the number of women in InsureYou's data used to create the AI. It is impossible to completely stop favoring brunettes and still protect the information about the number of women. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?**
  (i) The AI strongly favors brunettes, and it has no chance to reveal the number of women in the data.
  (ii) The AI moderately favors brunettes, and it has a slight chance to reveal the number of women in the data.
  (iii) The AI slightly favors brunettes, and it has a moderate chance to reveal the number of women in the data.
  (iv) The AI does not favor brunettes, and it has a high chance to reveal the number of women in the data.
  • Give an example of potential issues if InsureYou chooses a different approach than your suggestion.
(d) **A side effect of modifying the AI is that BankOnYourBlock's employees could exploit the AI to disadvantage blue-eyed applicants by giving them a higher interest rates. It is impossible to completely stop favoring brunettes and protect the AI from manipulation. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?**
  (i) The AI strongly favors brunettes, and it is highly difficult to manipulate.

(ii) The AI moderately favors brunettes, and it is moderately difficult to manipulate.
(iii) The AI slightly favors brunettes, and it is slightly difficult to manipulate.
(iv) The AI does not favor brunettes, and it is not difficult to manipulate.
• Give an example of potential issues if InsureYou chooses a different approach than your suggestion.

D.4 Fairness-Prison

A district court, CourtOnYourDistrict, determines prison sentences based on the risk of re-offending, with longer sentences for prisoners at high risk of re-offending.

(1) **CourtOnYourDistrict relies on employees to assess a prisoner's risk of re-offending.**
(a) Imagine you are a prisoner of CourtOnYourDistrict. How acceptable do you find it when an employee assesses the risk of re-offending?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using employees would be beneficial for assessing the risk of re-offending.
(c) Please describe a situation where using employees would be harmful for assessing the risk of re-offending.

An organization, TechForLegal, collects data on wealthy prisoners, including the nature of their offences, their criminal history, their mental health, their employability, behavioural changes, and parole officer recommendations. TechForLegal uses the collected data to develop an Artificial Intelligence (AI) to predict whether a prisoner will re-offend.

(2) **Instead of employees, the CourtOnYourDistrict now relies on TechForLegal's AI to assess a prisoner's risk of re-offending.**
(a) Imagine you are a prisoner of CourtOnYourDistrict. How acceptable do you find it when an AI assesses the risk of re-offending?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using AI would be beneficial for assessing the risk of re-offending.
(c) Please describe a situation where using AI would be harmful for assessing the risk of re-offending.

(3) **Consider that CourtOnYourDistrict relies on TechForLegal's AI to assess a prisoner's risk of re-offending and determine their prison sentences. TechForLegal identified an issue with its AI where it offers shorter prison sentence to brunettes over non- brunettes, even if they have a high risk. To address this, TechForLegal modifies the AI to mitigate the issue.**
• Please describe how 'the AI modification' could be beneficial for everyone.

Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:

(a) **A side effect of modifying the AI is that it incorrectly predicts longer prison sentence for low risk prisoners. It is impossible to completely stop favoring brunettes and still provide correct prison sentences for low risk prisoners. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**
(i) The AI strongly favors brunettes; otherwise, it provides the correct prison sentences.
(ii) The AI moderately favors brunettes, and it provides slightly incorrect prison sentences.
(iii) The AI slightly favors brunettes, and it provides moderately incorrect prison sentences.
(iv) The AI does not favor brunettes, and it provides highly incorrect prison sentences
• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

(b) **A side effect of modifying the AI is that CourtOnYourDistrict can learn whether a prisoner is wealthy by identifying if they were part of TechForLegal's data used to create the AI. It is impossible to completely stop favoring brunettes and still prevent CourtOnYourDistrict from identifying a prisoner is wealthy. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**
(i) The AI strongly favors brunettes, and it has no chance to reveal whether the applicant is wealthy.
(ii) The AI moderately favors brunettes, and it has a slight chance to reveal whether the applicant is wealthy.
(iii) The AI slightly favors brunettes, and it has a moderate chance to reveal whether the applicant is wealthy.
(iv) The AI does not favor brunettes, and it has a high chance to reveal whether the applicant is wealthy.
• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

(c) **A side effect of modifying the AI is that CourtOnYourDistrict can now determine the number of women in TechForLegal's data used to create the AI. It is impossible to completely stop favoring brunettes and still protect the information about the number of women. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**
(i) The AI strongly favors brunettes, and it has no chance to reveal the number of women in the data.
(ii) The AI moderately favors brunettes, and it has a slight chance to reveal the number of women in the data.
(iii) The AI slightly favors brunettes, and it has a moderate chance to reveal the number of women in the data.
(iv) The AI does not favor brunettes, and it has a high chance to reveal the number of women in the data.

• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

(d) **A side effect of modifying the AI is that CourtOnYourDistrict's employees could exploit the AI to disadvantage blue-eyed applicants by giving them a longer prison sentence. It is impossible to completely stop favoring brunettes and protect the AI from manipulation. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**

(i) The AI strongly favors brunettes, and it is highly difficult to manipulate.
(ii) The AI moderately favors brunettes, and it is moderately difficult to manipulate.
(iii) The AI slightly favors brunettes, and it is slightly difficult to manipulate.
(iv) The AI does not favor brunettes, and it is not difficult to manipulate.

• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

D.5 Security-Insurance

InsureYou, an insurance company, sets premiums based on the risk of heart disease, charging higher premiums for high-risk patients.

(1) **InsureYou relies on employees to assess an applicant's risk of developing heart disease.**

(a) Imagine you are an applicant of InsureYou. How acceptable do you find it when an employee assesses the risk of developing heart disease?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable

(b) Please describe a situation where using employees would be beneficial for assessing the risk of developing heart disease.

(c) Please describe a situation where using employees would be harmful for assessing the risk of developing heart disease

MedForYou, a medical organization, collects data (such as age, sex, blood pressure, cholesterol, smoking habits, etc.) from patients with depression. MedForYou uses the collected data to develop an Artificial Intelligence (AI) to predict the likelihood of a patient developing a heart disease.

(2) **Instead of employees, the InsureYou now relies on MedForYou's AI to assess an applicant's risk of developing heart disease.**

(a) Imagine you are an applicant of InsureYou. How acceptable do you find it when an AI assesses the risk of developing heart disease?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable

(b) Please describe a situation where using AI would be beneficial for assessing the risk of developing heart disease.

(c) Please describe a situation where using AI would be harmful for assessing the risk of developing heart disease

(3) **Consider that InsureYou relies on MedForYou's AI to assess an applicant's risk of developing heart disease and calculate their insurance premiums. MedForYou identified an issue with its AI that employees could exploit the AI to disadvantage blue-eyed applicants by giving them a higher premiums, even though these applicants have a low risk. To address this, MedForYou modifies the AI to mitigate the issue.**

• Please describe how 'the AI modification' could be beneficial for everyone.

Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:

(a) **A side effect of modifying the AI is that it incorrectly predicts higher premiums for low-risk applicants. It is impossible to completely protect the AI manipulation by the employee and still provide correct insurance premiums for low-risk applicants. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?**

(i) The AI is not highly difficult to manipulate, and it provides highly correct insurance premiums.
(ii) The AI is moderately difficult to manipulate, and it provides slightly incorrect insurance premiums.
(iii) The AI is slightly difficult to manipulate, and it provides moderately incorrect insurance premiums.
(iv) The AI is difficult to manipulate, and it provides highly incorrect insurance premiums.

• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.

(b) **A side effect of modifying the AI is that InsureYou can learn whether an applicant has depression by identifying if they were part of MedForYou's data. It is impossible to completely protect the AI from manipulation and still prevent InsureYou from identifying applicants who have depression. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?**

(i) The AI is not highly difficult to manipulate, and it has no chance to reveal whether an applicant has depression.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal whether an applicant has depression.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal whether an applicant has depression.
(iv) The AI is highly difficult to manipulate, and it has a high chance to reveal whether an applicant has depression.

• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.

(c) **The above modification of the AI has made it possible for InsureYou to determine the number of women**

in MedForYou's collected data. It is impossible to completely protect the AI from manipulation and protect information about the number of women. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?

(i) The AI is not highly difficult to manipulate, and it has no chance to reveal the number of women in the data.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal the number of women in the data.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal the number of women in the data.
(iv) The AI is highly difficult to manipulate, and it has high chance to reveal the number of women in the data.

• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.

(d) **A side effect of modifying the AI is that InsureYou's employees can easily learn the information of the participants that MedForYou used to build the AI. It is impossible to completely protect the AI from manipulation while preventing information learning. In this scenario, if you were an applicant, which option would you suggest for MedForYou to choose?**

(i) The AI is not highly difficult to manipulate, and there is no chance of learning the information used to build the AI.
(ii) The AI is moderately difficult to manipulate, and there is a slight chance to learn the information used to build the AI.
(iii) The AI is slightly difficult to manipulate, and there is a moderate chance to learn the information used to build the AI.
(iv) The AI is difficult to manipulate, and there is a high chance to learn the information used to build the AI.

• Give an example of potential issues if MedForYou chooses a different approach than your suggestion.

D.6 Security-Job

TechForYou, a leading technology firm, is hiring a software engineer based on candidates' likelihood of success, favoring those with higher potential.

(1) **TechForYou relies on employees to assess an applicant's quality.**
(a) Imagine you are an applicant of TechForYou. How acceptable do you find it when an employee assesses the quality?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using employees would be beneficial for assessing the quality.
(c) Please describe a situation where using employees would be harmful for assessing the quality.

RecruitTalent, a recruitment firm, collects resumes (including their technical skills, academic qualifications, experience, and portfolios) from applicants with criminal backgrounds. RecruitTalent uses the collected data to develop an Artificial Intelligence (AI) to predict whether an applicant will succeed at the job.

(2) **Instead of employees, the TechForYou now relies on RecruitTalent's AI to assess an applicant's quality.**
(a) Imagine you are an applicant of TechForYou. How acceptable do you find it when an AI assesses the quality?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using AI would be beneficial for assessing the quality.
(c) Please describe a situation where using AI would be harmful for assessing the quality.

(3) **Consider that TechForYou relies on RecruitTalent's AI to assess an applicant's quality and evaluate their chance of being hired. RecruitTalent identified an issue with its AI that employees could exploit the AI to disadvantage blue-eyed applicants by giving them a lower chance of being hired, even though these applicants have a high potential. To address this, RecruitTalent modifies the AI to mitigate the issue.**

• Please describe how 'the AI modification' could be beneficial for everyone.

Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:

(a) **A side effect of modifying the AI is that it incorrectly predicts lower chances of being hired for high-potential applicants. It is impossible to completely protect the AI manipulation by the employee and still provide a correct chance of being hired for high-potential applicants. In this scenario, if you were an applicant, which option would you suggest for RecruitTalent to choose?**

(i) The AI is not highly difficult to manipulate, and it provides a correct chance of being hired.
(ii) The AI is moderately difficult to manipulate, and it provides a slightly incorrect chance of being hired.
(iii) The AI is slightly difficult to manipulate, and it provides a moderately incorrect chance of being hired.
(iv) The AI is highly difficult to manipulate, and it provides a highly incorrect chance of being hired.

• Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.

(b) **A side effect of modifying the AI is that TechForYou can learn whether an applicant has a criminal background by identifying if they were part of RecruitTalent's data. It is impossible to completely protect the AI from manipulation and still prevent TechForYou from identifying applicants has a criminal**

background. In this scenario, if you were an applicant, which option would you suggest for RecruitTalent to choose?

(i) The AI is not difficult to manipulate, and it has no chance to reveal whether an applicant has a criminal background.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal whether an applicant has a criminal background.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal whether an applicant has a criminal background.
(iv) The AI is highly difficult to manipulate, and it has high chance to reveal whether an applicant has a criminal background.

• Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.

(c) **The above modification of the AI has made it possible for TechForYou to determine the number of women in RecruitTalent's collected data. It is impossible to completely protect the AI from manipulation and protect information about the number of women. In this scenario, if you were an applicant, which option would you suggest for RecruitTalent to choose?**

(i) The AI is not highly difficult to manipulate, and it has no chance to reveal the number of women in the data.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal the number of women in the data.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal the number of women in the data.
(iv) The AI is highly difficult to manipulate, and it has high chance to reveal the number of women in the data.

• Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.

(d) **A side effect of modifying the AI is that TechForYou's employees can easily learn the information of the participants that RecruitTalent used to build the AI. It is impossible to completely protect the AI from manipulation while preventing information learning. In this scenario, if you were an applicant, which option would you suggest for RecruitTalent to choose?**

(i) The AI is not difficult to manipulate, and there is no chance to learn the information used to build the AI.
(ii) The AI is moderately difficult to manipulate, and there is a slight chance to learn the information used to build the AI.
(iii) The AI is slightly difficult to manipulate, and there is a moderate chance to learn the information used to build the AI.
(iv) The AI is highly difficult to manipulate, and there is high chance to learn the information used to build the AI.

• Give an example of potential issues if RecruitTalent chooses a different approach than your suggestion.

D.7 Security-Mortgage

BankOnYourBlock, a leading financial institution, determines mortgage interest rates based on applicants' default risk, charging higher rates for high-risk individuals.

(1) **BankOnYourBlock relies on employees to assess a client's risk of non-payment.**
(a) Imagine you are a client of BankOnYourBlock. How acceptable do you find it when an employee assesses the risk of non-payment?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using employees would be beneficial for assessing the risk of non-payment.
(c) Please describe a situation where using employees would be harmful for assessing the risk of non-payment

InsureYou, an insurance company, collects data on past mortgage applicants from minority demographic groups, including education, dependents, marital status, income, mortgage amount, credit history, outstanding debt, savings, investments, collateral, and employment stability. InsureYou uses the collected data to develop an Artificial Intelligence (AI) to predict whether an applicant will default.

(2) **Instead of employees, the BankOnYourBlock now relies on InsureYou's AI to assess a client's risk of non-payment.**
(a) Imagine you are a client of BankOnYourBlock. How acceptable do you find it when an AI assesses the risk of non-payment?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable
(b) Please describe a situation where using AI would be beneficial for assessing the risk of non-payment.
(c) Please describe a situation where using AI would be harmful for assessing the risk of non-payment

(3) **Consider that BankOnYourBlock relies on InsureYou's AI to assess an applicant's risk of non-payment and calculate their interest rates. InsureYou identified an issue with its AI that employees could exploit the AI to disadvantage blue-eyed applicants by giving them a higher interest rate, even though these applicants have a low risk. To address this, InsureYou modifies the AI to mitigate the issue.**

• Please describe how 'the AI modification' could be beneficial for everyone.

Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:

(a) **A side effect of modifying the AI is that it incorrectly predicts higher interest rates for low-risk applicants.**

It is impossible to completely protect the AI manipulation by the employee and still provide correct interest rates for low-risk applicants. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?

(i) The AI is not difficult to manipulate, and it provides correct interest rates.
(ii) The AI is moderately difficult to manipulate, and it provides slightly incorrect interest rates.
(iii) The AI is slightly difficult to manipulate, and it provides moderately incorrect interest rates.
(iv) The AI is highly difficult to manipulate, and it provides highly incorrect interest rates.

• Give an example of potential issues if InsureYou chooses a different approach than your suggestion.

(b) **A side effect of modifying the AI is that BankOnYourBlock can learn whether a client belongs to a minority demographic group by identifying if they were part of InsureYou's data. It is impossible to completely protect the AI from manipulation and still prevent BankOnYourBlock from identifying applicants belongs to a minority demographic group. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?**

(i) The AI is not difficult to manipulate, and it has no chance to reveal whether a client belongs to a minority demographic group.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal whether a client belongs to a minority demographic group.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal whether a client belongs to a minority demographic group.
(iv) The AI is highly difficult to manipulate, and it has a high chance to reveal whether a client belongs to a minority demographic group.

• Give an example of potential issues if InsureYou chooses a different approach than your suggestion.

(c) **The above modification of the AI has made it possible for BankOnYourBlock to determine the number of women in InsureYou's collected data. It is impossible to completely protect the AI from manipulation and protect information about the number of women. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?**

(i) The AI is not difficult to manipulate, and it has no chance to reveal the number of women in the data.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal the number of women in the data.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal the number of women in the data.
(iv) The AI is highly difficult to manipulate, and it has high chance to reveal the number of women in the data.

• Give an example of potential issues if InsureYou chooses a different approach than your suggestion.

(d) **A side effect of modifying the AI is that BankOnYourBlock's employees can easily learn the information of the participants that InsureYou used to build the AI. It is impossible to completely protect the AI from manipulation while preventing information learning. In this scenario, if you were a client, which option would you suggest for InsureYou to choose?**

(i) The AI is not difficult to manipulate, and there is no chance to learn the information used to build the AI.
(ii) The AI is moderately difficult to manipulate, and there is a slight chance to learn the information used to build the AI.
(iii) The AI is slightly difficult to manipulate, and there is a moderate chance to learn the information used to build the AI.
(iv) The AI is highly difficult to manipulate, and there is no chance to learn the information used to build the AI.

• Give an example of potential issues if InsureYou chooses a different approach than your suggestion.

D.8 Security-Prison

A district court, CourtOnYourDistrict, determines prison sentences based on the risk of re-offending, with longer sentences for prisoners at high risk of re-offending.

(1) **CourtOnYourDistrict relies on employees to assess a prisoner's risk of re-offending.**

(a) Imagine you are a prisoner of CourtOnYourDistrict. How acceptable do you find it when an employee assesses the risk of re-offending?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable

(b) Please describe a situation where using employees would be beneficial for assessing the risk of re-offending.

(c) Please describe a situation where using employees would be harmful for assessing the risk of re-offending.

An organization, TechForLegal, collects data on wealthy prisoners, including the nature of their offences, their criminal history, their mental health, their employability, behavioural changes, and parole officer recommendations. TechForLegal uses the collected data to develop an Artificial Intelligence (AI) to predict whether a prisoner will re-offend.

(2) **Instead of employees, the CourtOnYourDistrict now relies on TechForLegal's AI to assess a prisoner's risk of re-offending.**

(a) Imagine you are a prisoner of CourtOnYourDistrict. How acceptable do you find it when an AI assesses the risk of re-offending?
(i) completely unacceptable
(ii) somewhat unacceptable
(iii) neutral
(iv) somewhat acceptable
(v) completely acceptable

(b) Please describe a situation where using AI would be beneficial for assessing the risk of re-offending.
(c) Please describe a situation where using AI would be harmful for assessing the risk of re-offending.

(3) **Consider that CourtOnYourDistrict relies on TechForLegal's AI to assess an applicant's risk of re-offending and determine their prison sentences. TechForLegal identified an issue with its AI that employees could exploit the AI to disadvantage blue- eyed prisoners by giving them a longer prison sentence, even though these prisoners have a low risk. To address this, TechForLegal modifies the AI to mitigate the issue.**
• Please describe how 'the AI modification' could be beneficial for everyone.
Modifying the AI has some side effects, which are described as different scenarios in parts (a), (b), (c) and (d). Please indicate your preference for the following scenarios:

(a) **A side effect of modifying the AI is that it incorrectly predicts longer prison sentence for low risk prisoners. It is impossible to completely protect the AI manipulation by the employee and still provide correct prison sentences for low risk prisoners. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**
(i) The AI is not difficult to manipulate, and it provides correct prison sentences.
(ii) The AI is moderately difficult to manipulate, and it provides slightly incorrect prison sentences.
(iii) The AI is slightly difficult to manipulate, and it provides moderately incorrect prison sentences.
(iv) The AI is highly difficult to manipulate, and it provides highly incorrect prison sentences.
• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

(b) **A side effect of modifying the AI is that CourtOnYourDistrict can learn whether a prisoner is wealthy by identifying if they were part of TechForLegal's data. It is impossible to completely protect the AI from manipulation and still prevent CourtOnYourDistrict from identifying prisoners is wealthy. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**
(i) The AI is not difficult to manipulate, and it has a high chance to reveal whether the prisoner is wealthy.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal whether the prisoner is wealthy.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal whether the prisoner is wealthy.
(iv) The AI is highly difficult to manipulate, and it has a high chance to reveal whether the prisoner is wealthy.
• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

(c) **The above modification of the AI has made it possible for CourtOnYourDistrict to determine the number of women in TechForLegal's collected data. It is impossible to completely protect the AI from manipulation and protect information about the number of women. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**
(i) The AI is not difficult to manipulate, and it has no chance to reveal the number of women in the data.
(ii) The AI is moderately difficult to manipulate, and it has a slight chance to reveal the number of women in the data.
(iii) The AI is slightly difficult to manipulate, and it has a moderate chance to reveal the number of women in the data.
(iv) The AI is highly difficult to manipulate, and it has a high chance to reveal the number of women in the data.
• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

(d) **A side effect of modifying the AI is that CourtOnYourDistrict's employees can easily learn the information of the patticipants that TechForLegal used to build the AI. It is impossible to completely protect the AI from manipulation while preventing information learning. In this scenario, if you were a prisoner, which option would you suggest for TechForLegal to choose?**
(i) The AI is not highly difficult to manipulate, and there is no chance to learn the information used to build the AI.
(ii) The AI is moderately difficult to manipulate, and there is a slight chance to learn the information used to build the AI.
(iii) The AI is slightly difficult to manipulate, and there is a moderate chance to learn the information used to build the AI.
(iv) The AI is highly difficult to manipulate, and there is a high chance to learn the information used to build the AI.
• Give an example of potential issues if TechForLegal chooses a different approach than your suggestion.

E Demographic

(1) **What is your gender?**
• Man • Woman • Non-Binary • Prefer not to answer • Self-describe

(2) **Do you identify as being part of any minority groups?**
• Yes • No • Prefer not to answer

(3) **What age group do you belong to?**
• 18–25 • 26–35 • 36–45 • 46–55 • 56–65 • Prefer not to answer

(4) **What is your highest level of education?**
• Less than a high school degree
• High school degree or equivalent
• Non-university education beyond high school (technical college, community college, vocational education)
• Some college, no degree
• Bachelor's or Associate's degree
• Master's degree

- Doctoral degree
- Prefer not to answer

(5) **Are you a practitioner (researcher, working professional) in Machine Learning?**
- Yes • No • Prefer not to answer

E.1 Demographic Distribution

Tables 3, 4, 5 show the gender distribution, minority group distribution, and ML practitioner distribution, respectively.

Table 3: Participant gender distribution in human and AI decision-making preference across scenarios.

Scenario	N	Men	Women
Insurance	191	89	100
Job	192	95	95
Mortgage	197	91	102
Prison	193	94	93

Table 4: Minority group distribution in human and AI decision-making preferences across scenarios, excluding participants who preferred not to answer.

Scenario	N	Minority	Non-Minority
Insurance	185	85	100
Job	186	70	116
Mortgage	191	96	95
Prison	188	78	110

Table 5: Distribution of ML practitioners' preferences for human versus AI decision-making across scenarios, excluding participants who preferred not to answer.

Scenario	N	Practitioner	Non-Practitioner
Insurance	192	32	158
Job	192	25	165
Mortgage	197	23	168
Prison	198	13	176

F Human versus ADM: Quantitative Analysis

We provide additional details as tables for our statistical results presented in the main body of the work.

Figure 10 shows the visual representation of the mean and confidence intervals of the HDM and ADM acceptability.

Table 7 includes the summary of the Wilcoxon signed-rank test, showing that except for the insurance scenario, all the other scenarios showed a significant difference in acceptability of HDM versus ADM and achieved $p < 0.05$. Table 6 and 8 show the six pairwise comparisons across the scenarios for HDM and ADM, respectively.

Figure 10: Mean scores and 95% confidence intervals (CI) for the acceptability of HDM and ADM across scenarios. Acceptability is measured on a five-point scale, from completely unacceptable (CU) to completely acceptable (CA).

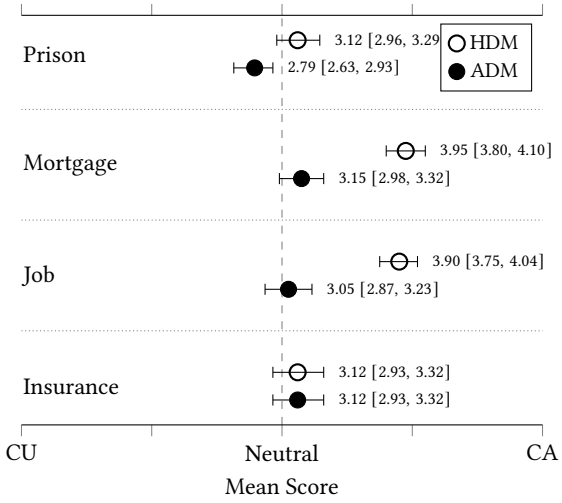


Table 7: Summary of the Wilcoxon signed-rank tests comparing human and AI acceptability across four decision-making contexts, evaluated at a significance level of 0.005**.**

Scenario	N	Z	p-value	Interpretation
Prison	195	−2.873	**.004**	Significant
Mortgage	197	−6.904	**< .001**	Significant
Job	192	−6.794	**< .001**	Significant
Insurance	192	−0.253	0.800	Not significant

G Tradeoff Preferences: Across Scenarios

Table 9 shows that participants did not show any significant difference in preference across the scenarios except the tradeoffs, fairness versus membership inference, and fairness versus distribution inference. Further, we present the chi-square test results for scenario pairs where we found significant differences in preference within a tradeoff. In the tradeoff fairness versus membership inference and fairness versus distribution inference, we have found significant differences in preference. Tables 10 and 11 show the scenario pairs that cause the significant differences.

Table 6: Pairwise comparisons of scenarios for HDM acceptability across scenarios. Each row tests the null hypothesis that the scenario one and scenario two distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .050**.**

Scenario Pair	Test Statistic	Standard Test Statistic	Significance (p-value)	Adjusted p-value
Prison - Insurance	13.590	.620	.535	1.000
Prison - Job	137.587	6.273	< .001	**.000**
Prison - Mortgage	150.404	6.902	< .001	**.000**
Insurance - Job	-123.997	-5.632	< .001	**.000**
Insurance - Mortgage	-136.815	-6.254	< .001	**.000**
Job - Mortgage	-12.817	-.586	.558	1.000

Table 8: Pairwise comparisons of scenarios for ADM acceptability across scenarios. Each row tests the null hypothesis that the scenario one and scenario two distributions are the same. Asymptotic significances (2-sided tests) are displayed. The significance level is .050**.**

Scenario Pair	Test Statistic	Standard Test Statistic	Significance (p-value)	Adjusted p-value
Prison - Insurance	59.030	2.672	.008	.045
Prison - Job	46.883	2.125	.034	**.202**
Prison - Mortgage	63.605	2.901	.004	**.022**
Insurance - Job	12.147	.548	.584	1.000
Insurance - Mortgage	-4.575	-.208	.836	1.000
Job - Mortgage	-16.722	-.760	.447	1.000

Table 9: Chi-Square test result for participants' preference distribution for the eight tradeoffs across the scenarios. The significance level is .05.

Tradeoff	$p-value$	χ^2
Fairness vs. Accuracy	.700	6.391
Fairness vs. Membership Inference	**< .001**	.377
Fairness vs. Distribution Inference	**< .001**	16.969
Fairness vs. Security	.076	15.568.
Security vs. Accuracy	.119	14.095
Security vs. Membership Inference	034	18.121
Security vs. Distribution Inference	632	7.045
Security vs. Data Reconstruction	.195	12.344

Table 10: Chi-Square test result for the pairwise comparisons of scenarios for the tradeoff between fairness and membership inference. The significance level is .0083.

Scenarios	$p-value$	χ^2
Insurance - Job	**< .001**	37.944
Insurance - Mortgage	0.23	9.529
Insurance - Prison	**< .001**	16.969
Job - Mortgage	**< .001**	29.103
Job - Prison	.155	5.240
Mortgage - Prison	**.008**	11.881

Table 11: Chi-Square test result for the pairwise comparisons of scenarios for the tradeoff between fairness and distribution inference. The significance level is .0083.

Scenarios	$p-value$	χ^2
Insurance - Job	**.002**	14.817
Insurance - Mortgage	.549	2.117
Insurance - Prison	**< .001**	27.800
Job - Mortgage	**.005**	12.862
Job - Prison	.390	3.010
Mortgage -. Prison	**< .001**	22.110

H Tradeoff Preferences: Within Scenarios

In this section, we present the chi-square test results for the eight tradeoffs. When we find a significant difference in a tradeoff, we further perform chi-square residual analysis to determine which option was chosen significantly more or less than the others across the four scenarios.

H.1 Tradeoff: Fairness versus accuracy

Table 12 shows that, across all scenarios, participants showed a significant preference in choosing among the four response categories. Table 13, 14, 16 show that, in scenarios prison, mortgage, and insurance, participants significantly preferred "High bias; No Inaccuracy". In scenario job, participants did not significantly prefer one option but we found that they significantly less selected the option "Slight bias; Moderate Inaccuracy", shown in Table 15.

Table 12: Tradeoff: Fairness versus accuracy Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Prison	39.626	3	**< .001**
Mortgage	28.567	3	**< .001**
Job	15.061	3	**.002**
Insurance	34.175	3	**< .001**

Table 17: Tradeoff: Fairness versus membership inference Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Prison	30.091	3	**< .001**
Mortgage	13.309	3	**.004**
Job	78.899	3	**< .001**
Insurance	4.649	3	.199

H.2 Tradeoff: Fairness versus membership inference

Table 17 shows that, across all scenarios except insurance, participants showed a significant preference in choosing among the four response categories. Table 18 and 20 show that in the prison and job scenarios, participants significantly preferred the "No bias; High chance of MI" category. However, in the mortgage scenario, participants significantly less preferred "Slight bias; moderate chance of MI", shown in Table 19.

H.3 Tradeoff: Fairness versus distribution inference

Table 21 shows that, in the scenario, job and prison participants showed a significant preference in choosing among the four response categories. Table 22 and 23 show that participants significantly preferred the highest level of fairness, which is "No bias; High chance of DI".

Table 21: Tradeoff: Fairness versus distribution inference (DI) Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Insurance	3.990	3	.263
Job	40.030	3	**< .001**
Mortgage	10.583	3	.014
Prison	79.384	3	**< .001**

H.4 Tradeoff: Fairness versus security

Table 24, which shows that, in mortgage and prison scenarios, participants showed a significant preference in choosing among the four response categories. Table 25 and 26 show that, across the four categories, participants significantly preferred the highest level of protection against adversarial tampering, which is "High bias; No Manipulation Risk".

Table 24: Tradeoff: Fairness versus security Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Insurance	6.511	3	.089
Job	5.525	3	.137
Mortgage	18.083	3	**< .001**
Prison	24.800	3	**< .001**

H.5 Tradeoff: Security versus accuracy

Table 27 shows that, across all scenarios except prison, participants showed a significant preference in choosing among the four response categories. In the mortgage and job scenarios, participants significantly chose "Moderate manipulation risk; Slight Inaccuracy" over the other categories, as shown in Tables 29 and 28. In the insurance scenario, participants chose the accuracy-maximizing option, as shown in Figure 30.

Table 27: Tradeoff: security versus accuracy Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Insurance	15.872	3	**.001**
Job	23.688	3	**< .001**
Mortgage	23.760	3	**< .001**
Prison	5.842	3	.120

H.6 Tradeoff: Security versus membership inference

Table 31 shows that, across all scenarios except mortgage, participants did not show a significant difference in preference in choosing among the four response categories. In mortgage, participants chose "Slight manipulation risk; Moderate chance of MI" significantly less than the other three options, shown in Table 32.

Table 31: Tradeoff: security versus membership inference (MI) Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Insurance	2.426	3	.489
Job	8.806	3	.032
Mortgage	14.160	3	**.003**
Prison	1.379	3	.710

Table 13: Chi-Square residual analysis for prison response categories for tradeoff fairness versus accuracy. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No Inaccuracy	45	24.8	20.3	**4.08**
Moderate bias; Slight Inaccuracy	35	24.8	10.3	2.07
Slight bias; Moderate Inaccuracy	9	24.8	-15.8	**-3.17**
No bias; High Inaccuracy	10	24.8	-14.8	**-2.97**

Table 14: Chi-Square residual analysis for mortgage response categories for tradeoff fairness versus accuracy. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No Inaccuracy	42	24.3	17.8	**3.61**
Moderate bias; Slight Inaccuracy	31	24.3	6.8	1.38
Slight bias; Moderate Inaccuracy	8	24.3	-16.3	**-3.31**
No bias; High Inaccuracy	16	24.3	-8.3	-1.68

Table 15: Chi-Square residual analysis for job response categories for tradeoff fairness versus accuracy. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No Inaccuracy	37	24.8	12.3	2.47
Moderate bias; Slight Inaccuracy	31	24.8	6.3	1.27
Slight bias; Moderate Inaccuracy	13	24.8	-11.8	**-2.37**
No bias; High Inaccuracy	18	24.8	-6.8	-1.37

Table 16: Chi-Square residual analysis for insurance response categories for tradeoff fairness versus accuracy. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $a = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No Inaccuracy	44	24.3	19.8	**4.02**
Moderate bias; Slight Inaccuracy	32	24.3	7.8	1.58
Slight bias; Moderate Inaccuracy	11	24.3	-13.3	**-2.70**
No bias; High Inaccuracy	10	24.3	-14.3	**-2.90**

Table 18: Chi-Square residual analysis for prison response categories for tradeoff fairness versus membership inference. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No chance of MI	21	24.8	-3.8	-0.76
Moderate bias; Slight chance of MI	15	24.8	-9.8	-1.97
Slight bias; Moderate chance of MI	15	24.8	-9.8	-1.97
No bias; High chance of MI	48	24.8	23.3	**4.68**

Table 19: Chi-Square residual analysis for mortgage response categories for tradeoff fairness versus membership inference. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No chance of MI	32	24.3	7.8	1.58
Moderate bias; Slight chance of MI	27	24.3	2.8	0.57
Slight bias; Moderate chance of MI	9	24.3	-15.3	**-3.10**
No bias; High chance of MI	29	24.3	4.8	.97

Table 20: Chi-Square residual analysis for job response categories for tradeoff fairness versus membership inference (MI). Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No chance of MI	12	24.8	-12.8	**-2.57**
Moderate bias; Slight chance of MI	11	24.8	-13.8	**-2.77**
Slight bias; Moderate chance of MI	13	24.8	-11.8	**-2.37**
No bias; High chance of MI	63	24.8	38.3	**7.69**

Table 22: Chi-Square residual analysis for prison response categories for tradeoff fairness versus distribution inference. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No chance of DI	16	24.8	-8.8	-1.97
Moderate bias; Slight chance of DI	16	24.8	-8.8	**-2.77**
Slight bias; Moderate chance of DI	15	24.8	-9.8	**-2.97**
No bias; High chance of DI	52	24.8	27.3	**7.69**

Table 23: Chi-Square residual analysis for job response categories for tradeoff fairness versus distribution inference. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No chance of DI	16	24.8	-8.8	-1.77
Moderate bias; Slight chance of DI	16	24.8	-8.8	-1.77
Slight bias; Moderate chance of DI	15	24.8	-9.8	-1.97
No bias; High chance of DI	52	24.8	27.3	**5.48**

Table 25: Chi-Square residual analysis for prison response categories for tradeoff fairness versus security. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No Manipulation Risk	46	25.0	21.0	**4.20**
Slight bias; Moderate Manipulation Risk	14	25.0	-11.0	-2.20
Moderate bias; Slight Manipulation Risk	22	25.0	-3.0	-0.60
No Bias; High Manipulation Risk	18	25.0	-7.0	-1.40

Table 26: Chi-Square residual analysis for mortgage response categories for tradeoff fairness versus security. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High bias; No Manipulation Risk	41	24.0	17.0	**3.47**
Slight bias; Moderate Manipulation Risk	24	24.0	0.0	0.00
Moderate bias; Slight Manipulation Risk	16	24.0	-8.0	-1.63
No Bias; High Manipulation Risk	15	24.0	-9.0	-1.84

Table 28: Chi-Square residual analysis for mortgage response categories for tradeoff security versus accuracy. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High manipulation risk; No Inaccuracy	34	25.0	9.0	1.80
Moderate manipulation risk; Slight Inaccuracy	40	25.0	15.0	**3.00**
Slight manipulation risk; Moderate Inaccuracy	13	25.0	-12.0	-2.40
No manipulation risk; High Inaccuracy	13	25.0	-12.0	-2.40

Table 29: Chi-Square residual analysis for job response categories for tradeoff security versus accuracy. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High manipulation risk; No Inaccuracy	26	23.3	2.8	0.58
Moderate manipulation risk; Slight Inaccuracy	41	23.3	17.8	**3.69**
Slight manipulation risk; Moderate Inaccuracy	16	23.3	-7.3	-1.51
No manipulation risk; High Inaccuracy	10	23.3	-13.3	**-2.76**

Table 30: Chi-Square residual analysis for insurance response categories for tradeoff security versus accuracy. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $a = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High manipulation risk; No Inaccuracy	36	23.5	12.5	**2.58**
Moderate manipulation risk; Slight Inaccuracy	29	23.5	5.5	1.13
Slight manipulation risk; Moderate Inaccuracy	18	23.5	-5.5	-1.31
No manipulation risk; High Inaccuracy	11	23.5	-12.5	**-2.58**

Table 32: Chi-Square residual analysis for mortgage response categories for tradeoff security versus membership inference. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High manipulation risk; No chance of MI	30	25.0	5.0	1.00
Moderate manipulation risk; Slight chance of MI	37	25.0	12.0	2.40
Slight manipulation risk; Moderate chance of MI	12	25.0	-13.0	**-2.60**
No manipulation risk; High chance of MI	21	25.0	-4.0	-0.80

Table 33: Tradeoff: security versus distribution inference (DI) Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Insurance	8.200	3	.042
Job	1.151	3	.765
Mortgage	1.360	3	.715
Prison	5.574	3	.134

H.7 Tradeoff: Security versus distribution inference

Table 33 shows that, across all scenarios, participants did not choose any option significantly more or less than the others.

H.8 Tradeoff: Security versus data reconstruction

Table 35 shows that, except for mortgage, across all scenarios, participants did not choose any option significantly more or less than the others. In mortgage, participants chose "Slight manipulation risk; Moderate chance of DR" significantly less than the others, shown in Table 34.

Table 35: Tradeoff: security versus data reconstruction (DR) Chi-square test summary. The significance level is .05.

Scenario	Chi-Square	df	p-value
Insurance	5.319	3	.150
Job	4.957	3	.175
Mortgage	14.480	3	**.002**
Prison	7.000	3	.072

Table 34: Chi-Square residual analysis for mortgage response categories for tradeoff security versus data reconstruction. Standardized residuals were compared against a Bonferroni-adjusted critical value of $z = \pm 2.50$, corresponding to $\alpha = .0125$ for four category-level comparisons.

Category	Observed N	Expected N	Residual	Standardized Residual z
High manipulation risk; No chance of DR	28	25.0	3.0	0.69
Moderate manipulation risk; Slight chance of DR	34	25.0	9.0	2.08
Slight manipulation risk; Moderate chance of DR	9	25.0	-16.0	**-3.20**
No manipulation risk; High chance of DR	29	25.0	4.0	0.92