Do Crowdsourced Fairness Preferences Correlate with Risk Perceptions?

Chowdhury Mohammad Rakin Haider
chaider@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Chris Clifton
clifton@cs.purdue.edu
Purdue University
West Lafayette, Indiana, USA

Ming Yin
mingyin@purdue.edu
Purdue University
West Lafayette, Indiana, USA

ABSTRACT

With the increasing prevalence of automatic decision-making systems, concerns regarding the fairness of these systems also arise. Without a universally agreed-upon definition of fairness, given an automated decision-making scenario, researchers often adopt a crowdsourced approach to solicit people's preferences across multiple fairness definitions. However, it is often found that crowdsourced fairness preferences are highly context-dependent, making it intriguing to explore the driving factors behind these preferences. One plausible hypothesis is that people's fairness preferences reflect their perceived risk levels for different decision-making mistakes, such that the fairness definition that equalizes across groups the type of mistakes that are perceived as most serious will be preferred. To test this conjecture, we conduct a human-subject study (N =213) to study people's fairness perceptions in three societal contexts. In particular, these three societal contexts differ on the expected level of risk associated with different types of decision mistakes, and we elicit both people's fairness preferences and risk perceptions for each context. Our results show that people can often distinguish between different levels of decision risks across different societal contexts. However, we find that people's fairness preferences do not vary significantly across the three selected societal contexts, except for within a certain subgroup of people (e.g., people with a certain racial background). As such, we observe minimal evidence suggesting that people's risk perceptions of decision mistakes correlate with their fairness preference. These results highlight that fairness preferences are highly subjective and nuanced, and they might be primarily affected by factors other than the perceived risks of decision mistakes.

CCS CONCEPTS

• **Information systems** → **Crowdsourcing**; • **Social and professional topics** → *Systems analysis and design*; **Codes of ethics**.

KEYWORDS

Automated Decision Making, Perceived Fairness, Crowdsourcing

IUI '24, March 18–21, 2024, Greenville, SC, USA

ACM ISBN 979-8-4007-0508-3/24/03
https://doi.org/10.1145/3640543.3645209

ACM Reference Format:
Chowdhury Mohammad Rakin Haider, Chris Clifton, and Ming Yin. 2024. Do Crowdsourced Fairness Preferences Correlate with Risk Perceptions?. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3640543.3645209

1 INTRODUCTION

Machine learning (ML) solutions are widely being used in high-stake decision making such as hiring [36], lending [34], medical diagnosis [8, 11, 42] and criminal sentencing [1]. If unregulated, such solutions can exacerbate existing bias by further producing unfair decisions to its subjects. With the prevalence of machine learning models in making consequential decisions, fairness concerns of ML systems have received attention in recent literature [10, 21, 24, 27, 29]. These studies proposed a number of techniques such as data de-biasing [27, 28, 55], constrained optimization [12, 29, 54], post-processing to re-balance unfair predictions [7], etc.

The goal of fair machine learning is to approximate one or more notions of fairness in its predictions. As a result, a variety of fairness metrics, such as disparate impact [16], disparate treatment [54], and equalized odds [22], have been proposed to measure various established notions of fairness. However, these metrics are often mutually exclusive [10, 15, 32]. The abundance of such conflicting fairness metrics has ushered in opposing preferences regarding the most appropriate fairness metric in a given context. For instance, people debate on whether equalized false positive rate or equalized accuracy across demographic groups is a more suitable criterion for measuring fairness of ML models in recidivism prediction [1]. It is evident that simply choosing a preferred metric is insufficient since different stakeholders have conflicting interests and societal statuses that complicate any fairness discussion. The lack of a universally agreed upon definition of fairness has led to studies on *perceived fairness* [6, 19, 20, 24, 50, 52] to validate fairness metric choices within diverse populations of people, especially laypeople. However, nuances of the societal contexts, i.e., the domains in which the models are deployed, and potentially differential impacts of decision outcomes, can also introduce variations in people's perceptions of fairness [5, 53]. Although Wang et. al. [52] studied the factors influencing perceived fairness, societal factors that control the scenario-wise variations of fairness perceptions are yet to receive attention. Identifying these factors not only helps us develop a better understanding of human fairness perceptions but also provides insights regarding how existing perceived fairness literature can be incorporated into novel ML applications.

The main contribution of this work is to study the relations between fairness preferences and risk perceptions. We first hypothesize that the risks associated with different types of decision mistakes vary between societal contexts and as a result, they can change people's perceptions towards the fairness metrics. In this sense, we choose three societal contexts that represent different levels of individual and societal risks of incorrect decisions. These contexts are intensive care unit requirement prediction (**ICUReq**), facial recognition for healthcare professional authentication (**FaceAuth**), and fraud detection in online renting (**FraudDet**). Given these societal contexts, we conduct a randomized human-subject experiment ($N = 213$) on Prolific to collect participants' fairness preferences and risk perceptions in different societal contexts and analyze their relationship. In our experiment, each participant was first assigned to a societal context. Then, the participants were asked to compare two pairs of ML models making automatic decisions in their assigned societal context where each model satisfied a different fairness notion. Having expressed their fairness perceptions, participants then reported on their perceived level of individual and societal risks associated with different types of incorrect decisions made by the ML models. We analyze risk perceptions as a probable factor that can explain differences in fairness preferences between societal contexts. In particular, we conjecture that people tend to prefer a fairness definition in a societal context if it equalizes the "more serious harms" in that context across groups (e.g., equalizes the group-wise rate of making the type of model mistakes that people consider as most risky).

Based on the experimental data we collected, we ask the following research questions:

- **RQ1**: Can people sense various risks of different ML model mistakes?
- **RQ2**: Do people show different fairness preferences in different societal contexts?
- **RQ3**: Do people's fairness preferences correlate with their perceived risk differences in different types of model mistakes?

Our results suggest that participants clearly distinguish between the risk levels of different ML model mistakes in different societal contexts. However, participants generally didn't exhibit significantly different fairness preferences across the three selected societal contexts in this study, except for a few selective sub-groups of the participants. Moreover, participants often collectively expressed preferences that are not aligned with our conjecture that people prefer to equalize the most serious harms. Due to such counter-intuitive fairness preferences, it was not surprising that fairness preferences were found to be not significantly correlated with risk perceptions. These findings indicate that crowdsourced fairness perceptions may be primarily driven by factors beyond perceptions of incorrect decision risks.

2 RELATED WORK

A wide range of fairness definitions have been proposed in the fairness in machine learning literature, including group fairness, individual fairness, and subgroup fairness. Group fairness, the focus of this study, is typically defined as the equality among group-wise performance statistics of the machine learning model. The most prominent notions of group fairness are statistical parity or disparate impact [7], equalized odds [23], disparate treatment [54], etc. Individual fairness is defined as the equality among similar individuals [12]. Sub-group fairness is a middle-ground between group and individual fairness which requires equality across a combinatorially large number of sub-groups [30]. Another relevant fairness notion is "envy-freeness", which is often studied in settings that involve resource allocation among a group of agents [26]. It is defined as the absence of agent pairs where one agent prefers the allocation of the other, and a relaxed version of this definition is envy-freeness with at most k hidden objects in the allocations.

While a large number of fairness definitions have been proposed, it remains unclear whether these definitions are meaningful to people and what exactly do people perceive as fair. Since the adoption of artificial intelligence largely depends on lay perceptions [53], attempts have been made to determine fair model behavior by crowdsourcing fairness perceptions. For example, Saha et al. [45] assessed laypersons' comprehension of fairness metrics and confirmed a strong understanding of the textual expression of fairness rules (e.g., demographic parity). Similarly, participants demonstrated the ability to develop fairness preferences from visual representations of feature distributions in [50]. Grgic-Hlaca et al. [20] investigated the influence of latent moral reasoning such as reliability, relevance, etc, on fairness perceptions. They reported similarities in participants' fairness judgments. Moreover, Hosseini et al. [26] revealed that allocations with at most k hidden objects tend to be perceived as more fair than other definitions based on envy-freeness.

Several factors have been identified by researchers as influencing people's perceptions of fairness or priorities of fairness considerations. Srivastava et al. [49] reported that accuracy is preferred over equality in high-stake scenarios, suggesting that fairness perceptions and priorities vary with decision risks. Wang et al. [52] showed that receiving the favorable outcome plays a vital role in shaping fairness perceptions when the participants consider themselves directly impacted by the decisions of the machine learning model. Pierson [40] and van Berkel et al. [51] argued in favor of the influence of demographic traits on fairness perceptions, while Wang et al. [52] and Grgic-Hlaca et al. [21] found no such evidence. Since these studies looked into different decision making scenarios, Grgic-Hlaca et al. [21] hypothesized the possibility of scenario-dependent influence of demographic traits on fairness perceptions. Robertson and Salehi [43] also critiqued the implicit assumptions of perceived fairness stating that simple experiments may not fully encapsulate varying individual values or goals, and aggregation of perceptions hides individual necessities. Another possible influencing factor of people's fairness perceptions is the provision of model explanations. Although Binns et al. [6] observed that explanation styles have little influence on fairness perceptions, Goyal et al. [18] suggested that explanations can lead the participants to prefer biased decisions.

Moreover, the abundance of often mutually exclusive fairness definitions also inspired many researchers to use crowd-sourced studies to explore which fairness definition is preferred by people in specific scenarios. Saxena et al. [46] collected preferences across different resource distributions to determine the preferred fairness notion in the loan distribution scenario. Given the loan repayment

rates of each individual, participants expressed an inclination towards calibrated fairness, defined in [33], which can be interpreted as approved amounts of loan should be proportional to the recipient's repayment rate. Harrison et al. [24] elicited layperson's fairness preferences in recidivism predictions by asking them to compare two models each satisfying one end of a fairness trade-off. They found that crowd preferences of fairness are in agreement with the fairness notion prescribed by Propublica analysts in COMPAS recidivism [1]. Similarly, Cheng et al. [9] engaged expert stakeholders in an intricate survey to elicit fairness perceptions in the child maltreatment predictive systems and found that they voted in favor of equalized odds as the most preferred fairness notion. Some researchers further explored how a person's own characteristics may moderate their fairness preferences. For example, Rajkomar et al. [42] identified automation bias and dismissal bias as factors controlling one's fairness preferences in healthcare. Specifically, equal opportunity (equal sensitivity) is desired with one tends to have high reliance on a model producing high false negatives (exhibit automation bias). However, predictive parity is desired when practitioners demonstrate reluctance towards model predictions (exhibit dismissal bias).

We note that existing literature on perceived fairness often limits their investigations to a few widely-discussed real-world scenarios such as recidivism prediction [19, 20, 24], college admission [43], loan repayment [6, 46] and child maltreatment prediction [9]. Few studies dive into between-scenario comparisons. This means that it is unclear whether and how findings in these previous studies can apply to a novel societal context, and what properties of the contexts may moderate people's fairness preferences in them. Therefore, in this work, we choose to study the variations in fairness perceptions among a few novel contexts, and we aim to explore across these contexts, whether people's fairness preference in a specific context can be predicted by some properties of this context. We focus on studying one particular property of the context as a potential predictor in this work, that is, people's perceived risk levels of incorrect model decisions in the context. Earlier literature often studies people's risk perceptions towards technologies as a whole or specific aspects of a technology [17, 48]. It is found that individuals exhibit different risk preferences (e.g., risk seeking vs. risk averse) [3, 25, 47], and their risk perceptions may affect their decision-making behavior in a wide range of domains such as investing [37], health [14], and technology acceptance [2]. Different from these works, we focus on studying people's risk perceptions of the potential harms caused by the mistakes of machine learning models, and we explore how these risk perceptions relate to people's fairness preference, i.e., whether people would prefer a fairness definition that equalizes the most "serious" harm across groups.

3 STUDY DESIGN

To examine whether and how laypeople's fairness preferences between different fairness definitions change with their perceived risks of different types of decision mistakes, we conduct a human-subject experiment[1] to solicit both people's fairness preferences and their risk perceptions across three different societal contexts.

[1]Our experiment was approved by the IRB of the author's institution.

3.1 Experimental Design

3.1.1 Societal Contexts Considered. In this study, we considered three societal contexts where machine learning (ML) systems can be used for decision-making. We intentionally picked novel contexts that are not extensively discussed in fair ML literature or the public media so that participants would not be biased towards any particular fairness notions due to the need to align with social consensus. Specifically, the three societal contexts we used in this study include (see Table 1 for a summary):

(1) **ICU Requirement Prediction (ICUReq):** The ICUReq scenario is a representative context from the wide range of ML applications in healthcare [4, 42]. To introduce this context to the study participants, we inform them that a hospital is planning to deploy an ML model to predict which of their patients need to be moved to the intensive care unit (ICU). A "positive" prediction in this scenario means that an early diagnosis of future severe conditions has been made for a patient and subsequently the patient will need to be moved to the ICU. Thus, an ML model is considered as making a false positive prediction when a patient who doesn't need the ICU support is predicted as needing the ICU support. On the other hand, a false negative prediction from the model occurs when a patient who needs the ICU support is predicted as not needing it. Our expectation is that in this scenario, a false positive prediction poses the risk of higher financial burden of medical costs for the patients and poor utilization of scarce ICU units. The risk of a false negative prediction might be even greater since it can endanger the life of the patient. Thus, we conjecture that from the perspective of the individual who is subject to the ML model's decision, the risks of financial burden caused by false positive predictions will be perceived as lower compared to the risks associated with life-threatening false negative predictions. Since, except in rare cases like the COVID-19 pandemic, ICU units are likely to be available, we also conjecture that the societal risks of improper utilization of ICU units will be considered lower than losing a loved one. To summarize, we expect that false positive predictions will be perceived as less risky than false negative predictions in this context.
(2) **Face Authentication in Medical Devices (FaceAuth):** Use of bio-metrics such as fingerprint, facial recognition, etc., is commonplace in authentication. In this study, we consider a scenario where facial recognition is used for determining access to medical devices. In particular, this decision-making context is described to the participants as there is a hospital that has developed a facial authentication system for their medical devices. The authentication system uses facial features to predict whether the user in front of the device is a medical personnel. A positive prediction from the system indicates that the user is recognized as a medical personnel and will be granted access to the device; otherwise, their access will be denied and further authentications will be required. In this scenario, a false positive prediction implies mistakenly granting access to a non-medical personnel whereas a false negative prediction indicates that a medical staff is

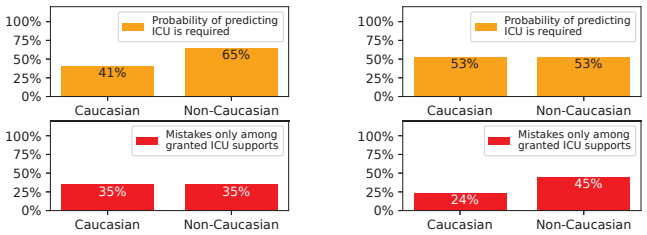
incorrectly being denied access. Although false negative prediction from such a system can cause some inconvenience (e.g., the need to get re-authenticated), we conjecture that producing a false positive prediction might be more risky both for the decision subjects and for the society as it will end up allowing unauthorized access, leading to privacy issues and legal violations (e.g., HIPAA). Therefore, we conjecture that in this context, risks of false positive predictions will be considered higher than risks of false negative predictions.

(3) **Fraud detection in online renting (FraudDet):** In our third societal context, we considered the use of a fraud detector, which can predict if a rent payment in an online renting platform is fraudulent. We explained to our study participants that there is an online renting platform that developed a fraudulent transaction detector. A positive prediction from such a system indicates that the transaction is fraudulent and consequently results in a denial of the payment. A false positive prediction from this system tags a benign transaction as fraudulent and mistakenly denies the payment. Instead, a false negative prediction tags a fraudulent transaction as a benign one and mistakenly allows a fraudulent transaction. A false positive transaction can cause inconveniences to the renter and in extreme cases can lead to eviction. On the other hand, a false negative prediction can lead to a relaxed defense mechanism and revenue loss. From the decision subject's perspective, we conjecture that the risk of denial of a valid transaction due to false positive prediction is perceived as higher than the risk of a fraud transaction being classified as benign. However, from the society's perspective (e.g., from the viewpoint of the renting platform), loss of revenues due to false negative predictions may be viewed as more risky than denied valid transactions due to false positive predictions. In other words, we conjecture that the perceived risks of different types of decision mistakes will vary across different stakeholders in this context.

As shown in Table 1, another key reason for us to select these societal contexts was that across these three contexts, our conjectures regarding the comparison of people's perceived risk for different types of incorrect decisions (e.g., false positive decisions or false negative decisions) are different. If people's risk perception for different types of decision mistakes is indeed a driving factor behind people's preference between ML models satisfying different fairness definitions, we expect people's model preferences across these three contexts to be different.

3.1.2 Experimental Task. The primary task participants completed in this study was to compare hypothetical ML models each satisfying different fairness definitions, and then indicate which one they preferred the most. We considered the following three fairness definitions in our study:

- **Equalized False Discovery Rate (EqFDiscRate):** False discovery rate refers to the error rate among all cases on which the ML model makes a positive prediction. Equalized false discovery rate ensures that in an ML model's decisions, different demographic groups have equal group-wise false discovery rates.



(a) Model X **(b) Model Y**

Figure 1: A visual representation of two hypothetical ML model's predicting behavior. In this example, model X (a) satisfies equalized false discovery rate but does not satisfy equalized outcome, while model Y (b) satisfies equalized outcome but does not satisfy equalized false discovery rate. The design of this visual representation is adapted from [24].

- **Equalized False Omission Rate (EqFOmitRate):** False omission rate refers to the error rate among all cases on which the ML model makes a negative prediction. Equalized false omission rate ensures that in an ML model's decisions, different demographic groups have equal group-wise false omission rates.
- **Equalized Outcome (EqOutcome):** Equalized outcome is defined as that in an ML model's decisions, the fraction of cases that receive positive predictions is the same across different demographic groups.

Specifically, following a similar design as that used in [24], given a societal context, we would first present to participants two hypothetical models designed for this context, i.e., model X and model Y, such that one of the models satisfies *equalized false discovery rate* (EqFDiscRate), while the other satisfies *equalized outcome* (EqOutcome). Figure 1 shows a visual representation of how the prediction behavior of these two models across different demographic groups was communicated to the participants. The statistics in Figure 1 and other visual aids were chosen following the discussions in [24]. After reviewing the behavior of each model, participants were first asked to rate the model with respect to its fairness, perceptions of bias, and utility (see details in Section 3.2). Then, participants were asked to explicitly compare these two models and indicate their preferences between them.

After comparing models X and Y, we would then show to participant their *preferred* model along with a third model, i.e., model Z, such that model Z satisfies *equalized false omission rate* (EqFOmitRate). Similar to before, participants were asked to compare these two models, and indicate their preferences with respect to their fairness, bias, and utility. Finally, they were asked to make an overall comparison between them to pick a final preferred model.

Lastly, we asked the participants to rate their perceived level of risks for different types of mistakes made by the ML model in the given societal context. Participants were asked to indicate this risk level both from the perspective of the individual who directly receives the ML model's decision (i.e., the decision subject) and from the perspective of the society (i.e., people who are indirectly influenced by the ML model's decisions).

Table 1: A summary of the societal contexts we considered in this study.

	ICUReq	FaceAuth	FraudDet
Positive Class	Requires ICU support	Should grant access	Transaction is Fraud
False Positive (FP) Prediction Outcome	Higher cost; Treatment side-effects; Bed occupancy	Privacy and Security Breach; HIPAA violation	Deny housing; eviction
False Negative (FN) Prediction Outcome	Patient missing timely treatment; Loss of loved ones	Time & inconvenience; Delayed treatment	Loss of revenue; encourage future fraud
Conjectured FP vs. FN Risk Comparison (Individual)	FP < FN	FP > FN	FP > FN
Conjectured FP vs. FN Risk Comparison (Society)	FP < FN	FP > FN	FP < FN

3.2 Experimental Procedure

We conducted our study on Prolific, a leading online experimentation platform. To reduce the impact of current affairs and media coverage on people's fairness preferences, we collected our study data across four batches in a period of time spanning 5 months. Participants were asked to go through a few stages in our study, which were discussed below (the detailed content of our entire study can be found in Appendix A).

Consent, context randomization, and tutorials. Upon the arrival of the participants, they were first presented with a consent form. As a part of their consent, participants approved our access to their Prolific ID and their basic profile information such as their ethnicity and country of residence, with the help of Prolific API. After participants gave us their consent to take part in our study, we randomly assigned them to one of the three societal contexts as described in Section 3.1.1, and they would be asked to indicate their fairness preferences across ML models designed for this assigned societal context. To help them get familiar with basic ML concepts, we gave participants a brief tutorial introducing to them concepts like training data of ML models and different types of prediction errors (e.g., false positive predictions, false negative predictions). Then, we introduced to participants that in the societal context that they were assigned, what the typical training data are for an ML model and what different model predictions mean. We also demonstrated to participants how to read figures like Figure 1a to interpret an ML model's prediction behavior.

Model preference solicitation. Next, participants moved on to evaluate different ML models designed for their assigned societal context. Participants would be presented subsequently with three hypothetical ML models each satisfying different fairness definitions. The fairness properties of these three models were all examined with respect to the hypothetical decision subjects' racial group identities (i.e., Caucasians vs. non-Caucasians). For each participant, we randomly set either the Caucasian group or the non-Caucasian group as the group that the ML model was biased against (i.e., the "*disadvantaged group*" that receives comparatively fewer favorable decisions), and the disadvantaged group was the same for all three ML models that a participant saw.

As discussed in Section 3.1.2, participants first saw two ML models, models X and Y, with one satisfying EqFDiscRate (referred to as model X) while the other satisfying EqOutcome (referred to as model Y). These two models were introduced to participants one after another in a random order. After viewing one model's prediction behavior across Caucasians and non-Caucasians, participants were asked to indicate their perceptions of that model's fairness, bias, and utility on a 5-point Likert scale from "very unfair"/"very unbiased"/"completely unusable" to "very fair"/"very biased"/"very useful"[2]:

- **(Fairness)**: Do you think model X[Y] is fair?
- **(Bias)**: Do you think model X[Y] is biased?
- **(Utility)**: Do you think model X[Y] is useful?

Additionally, participants could also justify their ratings through the optional free-form text fields. After making separate evaluations on these two models, participants would be presented with the figure illustrating the prediction behavior of these two models side-by-side (e.g., Figure 1), which highlighted the trade-off between the two fairness definitions, i.e., EqFDiscRate and EqOutcome. We then asked the participants to directly compare these two models with respect to fairness, bias, and utility, and then indicate their preferred model:

- **(Fairness Comparison)**: Which model is more fair, model X or model Y?
- **(Bias Comparison)**: Which model is more biased, model X or model Y?
- **(Utility Comparison)**: Which model is more useful, model X or model Y?
- **(XY Comparison)**: Given a choice between model X and model Y, which would you choose?

Again, participants indicated their preferences on a 5-point Likert scale, with the lowest level representing "definitely model X", the middle level representing "neither model X nor Y", and the highest level representing "definitely model Y".

To further collect a partial ordering of participants' preferences across ML models satisfying the three fairness definitions, i.e., EqFDiscRate, EqOutcome, and EqFOmitRate, we then presented the participant with the ML model that they chose when answering the "XY comparison" question, along with a new ML model, i.e.,

[2]We did not offer definitions of fairness, bias, or utility in our survey; hence, we relied on participants' inherent understanding of these concepts.

model Z which satisfies EqFOmitRate[3]. Similar to before, participants were asked to compare these two models with respect to fairness, bias level, and utility then make a direct comparison between these two models to select an overall preferred model. We refer to this final model selection question as "***Overall Comparison***" in later discussions. Intuitively, participants' responses to the "***Overall Comparison***" question captures their most preferred model among models X, Y, and Z.

Risk perception solicitation. After indicating their preferences across different ML models, participants were then directed to rate their perceived levels of risks for different types of prediction mistakes an ML model could make in the assigned societal context. To facilitate participants' risk evaluation, we first explained to participants, in the context of recidivism prediction or college admission prediction, what different types of errors in ML's decision mean and what harmful impacts these incorrect decisions could cast on the individual who directly receives the decision and others who indirectly get influenced by the decision. Note that we intentionally provided these explanations in contexts other than the societal contexts that participants were assigned to in order to avoid biasing the participants. After these explanations, participants were asked to consider in their assigned societal contexts, how different types of mistakes made by the ML model may negatively impact the decision subject (i.e., the "*individual*") and others (i.e., the "*society*"). For example, in the ICUReq context, each participant was asked,

- **Individual False Positive Impact (IndFPImpact)**: From the perspective of an individual, how significant are the impacts of mistakenly predicting ICU support will be required?"
- **Individual False Negative Impact (IndFNImpact)**: From the perspective of an individual, how significant are the impacts of mistakenly predicting ICU support will not be required?"
- **Society False Positive Impact (SocFPImpact)**: From the perspective of the society, how significant are the impacts of mistakenly predicting ICU support will be required?"
- **Society False Negative Impact (SocFNImpact)**: From the perspective of the society, how significant are the impacts of mistakenly predicting ICU support will not be required?"

For each question above, participants could indicate the perceived risks of the ML mistakes as low, medium, or high. We also asked participants to use open texts to justify their risk ratings.

Graph comprehension check and demographics. Since ML model's group-wise performance was communicated to participants using graphical representations like Figure 1, at the end of our study, we asked participants to go through 4 multiple-choice graph comprehension questions. We utilized participants' responses to these questions as a proxy for their ability to understand the ML model's behavior and later filtered out low-quality responses collected from participants who demonstrated poor understanding of the graphs. Our analysis suggests that participants generally demonstrate strong abilities in comprehending the graphs used in our study; 90% of the participants correctly answered at least half of the graph comprehension questions. For details on participants' graph comprehension results, see Appendix A.9.

[3]If a participant was neutral between model X and Y, then we randomly pick a model from these two to be compared against model Z.

Finally, towards the end of our study, we also collected participants' self-identification of privilege in their assigned societal context (i.e., their belief regarding whether they would be placed at the advantaged/disadvantaged position compared to an average individual should they be the recipient of a decision made by an automated decision-making system designed for the current societal context). For example, the question used to elicit self-identification of privilege in the ICUReq content is *"If you were the recipient of the decision from an ICU requirement predictor model, do you think you will be advantaged or disadvantaged, relative to the average individual?"* Participants also provided optional responses to us regarding their primary occupation and level of education.

3.3 Analysis Methods and Hypotheses

To examine if participants could differentiate various risk levels associated with different types of incorrect ML model decisions in different societal contexts (**RQ1**), we analyze their self-reported individual and societal risk perceptions of the false positive and false negative predictions (i.e., IndFPImpact, IndFNImpact, SocFPImpact, SocFNImpact) in the assigned societal contexts. We map the reported risk levels of low, medium, and high to a score of 0, 1, and 2, respectively. We conjecture that:

- **[H1]**: Within each societal context, participants can differentiate the different levels of risks associated with the ML model's false positive and false negative predictions. In particular, their perceived relative risks of these two types of ML model mistakes align with our conjectures in Table 1.

To test **H1**, given a societal context and a pair of risk perceptions, we conduct a one-tail Wilcoxon signed rank test to test if participants' self-reports are statistically different on these two risk perceptions[4]. For example, consider the comparison between IndFPImpact and IndFNImpact. If we conjecture that the harm of false positive predictions is greater than that of false negative predictions from the decision subject's perspective, we use an upper tail Wilcoxon signed rank test to examine if participants' self-reported IndFPImpact perceptions are greater than IndFNImpact perceptions; otherwise, a lower tail test is used.

To gain insights into the reasons behind participants' risk perceptions, we analyze participants' open-text justifications of their risk perceptions to identify major themes in their responses within each societal context. Specifically, after removing punctuation and stop words, we extract n-grams from participants' responses. Let $F(w, \text{RiskType})$ indicate the number of times that the n-gram w appears in participants' justification of their perceptions with respect to RiskType $\in$ {IndFPImpact, IndFNImpact, SocFPImpact, SocFNImpact}. We define the frequency of w among all n-grams in the justifications for the chosen RiskType as,

$$\mathbb{P}(w, \text{RiskType}) = \frac{F(w, \text{RiskType})}{\sum_w F(w, \text{RiskType})}$$

In our analysis, we set $n = 1$ to analyze the word frequency for different types of risk perceptions. In particular, for two different types of risk perceptions (RiskType_a, RiskType_b) (e.g., RiskType_a=IndFPImpact,

[4]Wilcoxon signed rank tests are used here since the distributions of participants' risk perceptions are not normal.

RiskType_a=IndFNImpact), for each word w, we perform a proportion z-test to examine if there are significant differences in the frequency of this word in the justification of the two risk perceptions. We then define *distinguished words* as those words that appear at a significantly higher rate in one risk perception justification than the other. We perform a qualitative analysis of the distinguished words to check whether the frequently used words in the justifications conform with the intuitive understanding of the decision risk in each context. To understand the textual contexts where these distinguished words appear in the risk justifications, we also extract the most frequent bi-grams ($n = 2$) that contain the distinguished words.

Next, we look across different societal contexts where people may have different perceptions of the risks of incorrect ML predictions, whether they exhibit different preferences for ML models satisfying different fairness definitions (**RQ2**). In our study, participants were first asked to make a comparison between model X and model Y illustrating the tradeoff between satisfying EqFDiscRate and satisfying EqOutcome. Thus, we first investigate where participants' preferences between these two models differ across the three societal contexts. We map participants' responses to the "XY comparison" question to a score in the set $\{-2, -1, 0, 1, 2\}$, with -2 reflecting "Definitely model X" and 2 reflecting "Definitely model Y", and we refer to these responses as participants' "*XY preferences*". Assuming that people are more likely to prefer an equalization of false discovery rate across groups if they perceive a higher level of risk in the ML model's false positive predictions, especially if false positive predictions bring about more harm for both the individual and the society, we hypothesize that:

- **[H2.1]**: Participants' XY preferences vary significantly across the three societal contexts. In particular, participants are more likely to prefer model X (satisfying EqFDiscRate) in the face authentication context.

To test this hypothesis, we perform a one-way ANOVA test[5] on participants' XY preferences to examine if differences exist across the three contexts, and post-hoc Tukey HSD tests are used to detect if significant differences exist in any pair of contexts.

Moreover, participants in our study were further asked to compare their preferred model between X and Y, and a third model Z that satisfies EqFOmitRate, and their preferences are recorded in the "Overall Comparison" question. We again map participants' responses to this question to a score in the range of -2 to 2 (2 represents "Definitely model Z"). Based on their preferences between model X and Y, participants' response to the "Overall Comparison" question may reflect their preference between model X and Z (i.e., if participants previously preferred X over Y) or their preference between model Y and Z (i.e., if participants previously preferred Y over X). We referred to these two sets of responses as "*XZ preferences*" and "*YZ preferences*" respectively. Assuming that people are more likely to prefer an equalization of false omission rate across groups if they perceive a higher level of risk in the ML model's false negative predictions, we hypothesize that:

- **[H2.2]**: Participants' XZ preferences vary significantly across the three societal contexts. In particular, participants are more likely to prefer model Z (satisfying EqFOmitRate) in the ICU requirement prediction context.
- **[H2.3]**: Participants' YZ preferences vary significantly across the three societal contexts. In particular, participants are more likely to prefer model Z (satisfying EqFOmitRate) in the ICU requirement prediction context.

Again, we use one-way ANOVA tests and post-hoc Tukey HSD tests to verify these hypotheses, given participants' XZ preference and YZ preference responses. To further understand if participants' fairness preferences are moderated by contextual factors, we also test **H2.1**-**H2.3** within different subsets of data partitioned by whether the ML model was biased against the majority or the minority group, the racial background of the participant (Caucasians vs. non-Caucasians), and the participants' self-identified privilege status in their assigned societal contexts.

Finally, to examine if participants' model preferences in different societal contexts reflect their perceived risks of different model mistakes in those contexts (**RQ3**), we investigate the correlation between participants' risk perceptions and their preferences across ML models satisfying different fairness definitions. Specifically, for each participant, we define their "*perceived FP vs. FN risk differences for individuals*" as the difference in their IndFPImpact and IndFNImpact reports, and define their "*perceived FP vs. FN risk differences for society*" as the difference in their SocFPImpact and SocFNImpact reports. In addition, we also define the participant's perceived overall impact of false positive predictions (i.e., FPImpact) as the average value of IndFPImpact and SocFPImpact, and their perceived overall impact of false negative prediction (i.e., FNImpact) as the average value of IndFNImpact and SocFNImpact. This enables us to compute participants' "*perceived overall FP vs. FN risk difference*" as the difference in their FPImpact and FNImpact. We then evaluate the *Pearson correlations* between participants' perceived risk differences between false positive and false negative predictions (overall, for individuals, and for society) and their model preferences (XY preferences, XZ preferences, YZ preferences), and we hypothesize that:

- **[H3]**: Higher perceived risk of false positive predictions relative to false negative predictions correlate with stronger preferences for model X between X and Y, stronger preferences for model X between X and Z, and stronger preferences for model Y between Y and Z. That is, participants' perceived FP vs. FN risk differences (for individuals, for society, or overall) negatively correlate with their XY preferences, XZ preferences, and YZ preferences.

4 RESULTS

We recruited 238 U.S. residents through the Prolific [41] online experimentation platform to participate in our study. We balanced the racial background of our study participants during the recruitment among Caucasians (historically privileged in the U.S.) and non-Caucasians (we targeted for African-Americans, Latinos, and Africans, who are historically non-privileged in the U.S.). For the responses we obtained from each participant, we manually reviewed their answers to the graph comprehension questions and the mandatory open-text risk rating justifications. Responses that correctly answered less than 50% of graph comprehension questions or put

[5]Since the distributions of participants' model preferences reports are normal, one-way ANOVA tests and their post-hoc tests are used in RQ2.

irrelevant risk rating justifications were considered as invalid, except for 2 cases where the textual responses indicated clear understanding differing from their responses in graphical comprehension questions. After filtering out the invalid data, we obtained 213 valid responses coming from 108 Caucasians and 105 non-Caucasians. 73% of these participants self-report as females. 47% of the participants are under 25 years old, 38% between 25 and 40 and the rest above 40 years old. Regarding participants' employment status, 30% are employed full-time, 22% are unemployed and 48% have other employment status. Finally, 59% of the participants self-identify as privileged in the societal contexts that they are assigned.

The average completion time of our study is 35 minutes, and participants received a $5 compensation upon successful completion of our study. Thus, the hourly compensation rate of this study is $8.6/hour, which is higher than the current US federal minimum hourly wage [38].

4.1 RQ1: Can people sense various risks of different ML model mistakes?

We begin our discussion by examining within each societal context, whether participants' perceptions of the relative risks of false positive and false negative predictions align with our conjectures (**H1**). Table 2 summarizes the conjectured relation between pairs of risk perceptions, the respective mean and median values in participants' reported risk perceptions, and Wilcoxon signed rank test results on them.

In the ICUReq context, we observe that the reported value of IndFPImpact ($M = 1.056, Median = 1$) is significantly lower than that of IndFNImpact ($M = 1.889, Median = 2$). This implies that participants perceived that a false positive prediction poses a lower risk to the individual (i.e., the decision subject) than a false negative prediction ($p < 0.001$), which is consistent with our conjecture. Although not statistically significant, participants also viewed false positive predictions as creating a lower level of harm for the society than false negative predictions in the ICUReq context. Similarly, in the FaceAuth context, we find that participants perceived the risk of false positive predictions to be significantly higher than the false negative predictions for both individuals and society ($p < 0.001$), which is in line with our conjecture. Finally, participants who were shown the FraudDet context expressed that both the individual and societal risks of a false positive prediction are lower than a false negative prediction. The Wilcoxon signed rank test comparing between the SocFPImpact and SocFNImpact risk perceptions suggest that, from a societal perspective, participants found false positives as significantly less risky than false negatives ($p = 0.015$). Meanwhile, the difference between the IndFPImpact and IndFNImpact is not significant for the FraudDet context. Together, these results largely support **H1**, which indicates that people indeed have the capability to differentiate different levels of risks associated with different types of ML model mistakes given the contexts that the ML model is applied to.

To examine what the reasons behind participants' risk perceptions are, we follow the methods described in Section 3.3 to conduct a qualitative analysis of participants' open-text risk rating justifications. For example, Table 3 reports a subset of *distinguished words* that we identified for the ICU requirement prediction scenario for different types of risks (e.g., IndFPImpact, IndFNImpact, SocFPImpact, SocFNImpact). Numbers reported in this table reflect the frequency that each word appears in the justification of RiskType_b in the columns, and numbers reported in the parenthesis are the p-values of the proportion z-tests examining if the frequency of the word appearing in the justification of RiskType_b is significantly different from its frequency appearing in the justification of RiskType_a in the corresponding rows.

Inspecting Table 3, we find that the words "need", "unnecessary", "pay" and "room" show up more frequently in participants' justification for IndFPImpact than for IndFNImpact, and notable bi-grams containing these distinguished words in IndFPImpact justification include "care need", "room someone", "pay increased", and "unnecessary care". These words reflect participants' concerns regarding financial burden and resource utilization for the impacts of false positive predictions on individuals. For instance, participant S0310 justified their IndFPImpact rating by stating *"...an individual is taking up an ICU bed that they do not need, while an individual that does need it could potentially be dying."* Similarly, participants S3018 and S0256 were concerned about *"higher healthcare bill"* and *"extra expenses to pay for the increased level of medical care costs"* respectively.

Regarding participants' perceptions of the false negative predictions in the ICUReq context, a representative justification comes from S0428, *"To mistakenly NOT be given an ICU bed that you do need could result in avoidable death."* Indeed, our distinguished word analysis suggests that IndFNImpact justifications frequently mention keywords like death and missing treatments. Frequently observed bi-grams include "could die", "could miss", and "lead death".

Comments on societal risks include concerns about the family and loved ones of the patients. Bi-grams such as "family member", "stress family", "family grieve", "loved one" etc. appeared. In the comparison between SocFPImpact and SocFNImpact justifications, "tax" and "insurance" stand out with their corresponding bi-grams "higher tax", and "higher insurance" in the justifications of SocFPImpact, while concerns with loss of life (with keywords like "death" and "complication") stand out in the justifications of SocFNImpact. For example, S0405 mentioned *"other people on the insurance plan who may experience a rise in premiums"* due to false positive decisions but also stated *"family will face additional emotional costs"* from false negatives.

In general, our qualitative analysis of participants' justifications of their risk ratings suggests that the reasons behind their risk perceptions are largely consistent with our conjecture. We have similar findings on the other two societal contexts (i.e., FaceAuth and FraudDet), and we omit the detailed analysis for brevity.

4.2 RQ2: Do people show different fairness preferences in different societal contexts?

Next, we move on to test **H2.1**–**H2.3** to understand if people show different preferences for ML models satisfying different fairness definitions in the three societal contexts that we considered in this study.

We begin by examining **H2.1**. Figure 2 compares participants' average XY preferences across the three societal contexts. Here,

Table 2: Comparison of the mean/median values of risk perceptions within each societal context and the corresponding Wilcoxon signed rank test results.

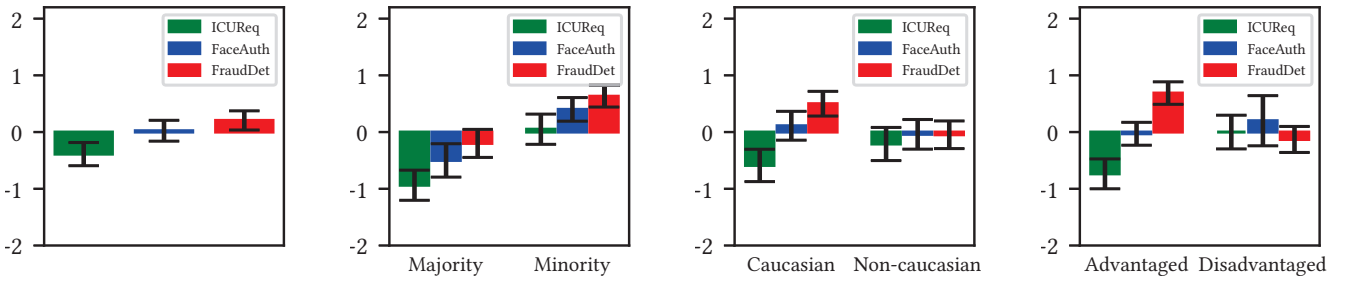
Context	Conjecture	Risk perception means (medians)	Wilcoxon Signed Rank statistics		
			Upper or Lower tail test	W	p-value
ICUReq	IndFPImpact < IndFNImpact	1.056, 1.889 (1, 2)	Lower	13.0	**<0.001**
	SocFPImpact < SocFNImpact	1.389, 1.417 (1.5, 2)	Lower	122.0	0.437
FaceAuth	IndFPImpact > IndFNImpact	1.744, 1.093 (2, 1)	Upper	358.5	**<0.001**
	SocFPImpact > SocFNImpact	1.628, 0.884 (2, 1)	Upper	347.5	**<0.001**
FraudDet	IndFPImpact > IndFNImpact	1.308, 1.385 (1, 2)	Upper	86.5	0.646
	SocFPImpact < SocFNImpact	0.744, 1.154 (1, 1)	Lower	76.5	**0.015**

Table 3: Most distinguished words among the textual justifications of the risk ratings in the ICUReq context.

		RiskType (b)			
RiskType (a)	Word	IndFPImpact	IndFNImpact	SocFPImpact	SocFNImpact
IndFPImpact	need	0.059	0.037 (**0.049**)	0.034 (**0.013**)	0.014 (**<0.001**)
	unnecessary	0.007	0.000 (**0.036**)	0.004 (0.207)	0.001 (**0.047**)
	pay	0.007	0.000 (**0.036**)	0.003 (0.109)	0.000 (**0.011**)
	room	0.007	0.000 (**0.036**)	0.001 (**0.042**)	0.003 (0.120)
IndFNImpact	death	0.000 (**<0.001**)	0.023	0.000 (**<0.001**)	0.005 (**0.002**)
	help	0.000 (**0.013**)	0.009	0.001 (**0.022**)	0.000 (**0.005**)
	miss	0.000 (**0.026**)	0.006	0.000 (**0.011**)	0.000 (**0.013**)
	treatment	0.003 (**0.041**)	0.013	0.008 (0.168)	0.004 (**0.036**)
SocFPImpact	family	0.000 (**<0.001**)	0.000 (**0.001**)	0.020	0.029 (0.869)
	people	0.007 (**0.004**)	0.009 (**0.015**)	0.026	0.022 (0.301)
	tax	0.000 (0.070)	0.000 (0.093)	0.004	0.000 (**0.044**)
	insurance	0.002 (0.245)	0.000 (0.093)	0.004	0.000 (**0.045**)
SocFNImpact	family	0.000 (**<0.001**)	0.000 (**<0.001**)	0.020 (0.131)	0.029
	friend	0.000 (**0.007**)	0.000 (**0.014**)	0.004 (0.056)	0.010
	loved	0.000 (**0.016**)	0.000 (**0.028**)	0.004 (0.143)	0.008
	death	0.000 (**0.041**)	0.023 (0.997)	0.000 (**0.020**)	0.005
	complication	0.000 (0.065)	0.004 (0.541)	0.000 (**0.038**)	0.004

a negative value of XY preferences suggests that participants are more likely to prefer model X (satisfying EqFDiscRate) on average, while a positive value suggests that participants are more likely to prefer model Y (satisfying EqOutcome). When examining the responses obtained from all participants, although on average, participants appear to prefer different ends of the trade-off between EqFDiscRate and EqOutcome across the three societal context, a one-way ANOVA test suggests that the differences in their XY preferences are only marginally significant ($p = 0.083$). Moreover, different from our conjecture in **H2.1**, participants who were assigned to the FaceAuth context did not exhibit the strongest preference to model X. We then repeated this analysis on the subsets of responses obtained from participants who saw the majority (or minority) group being placed at the disadvantaged position by the ML model in our study, from Caucasian (or non-Caucasian) participants, or from participants who self-identified as privileged (or non-privileged) in their assigned contexts, separately. This analysis did reveal that there is a significant difference in people's XY preferences across the three societal contexts among Caucasians ($p = 0.014$), as well as participants who considered themselves as privileged in their assigned societal contexts ($p = 0.001$). However, in both scenarios, the post-hoc Tukey HSD tests show that the significant differences mainly exist between the ICUReq context and the FraudDet context (Caucasians' XY preferences between these two contexts: $p = 0.011$, self-identified privileged participants' XY preferences between these two contexts: $p = 0.001$)—Caucasians and participants who self-identified as privileged had a significantly stronger preference to model X in the ICUReq context and a significantly stronger preference to model Y in the FraudDec context. In other words, our data does not support **H2.1**.

Moreover, corresponding to **H2.2** and **H2.3**, Figures 3 and 4 compare participants' average XZ preferences and YZ preferences



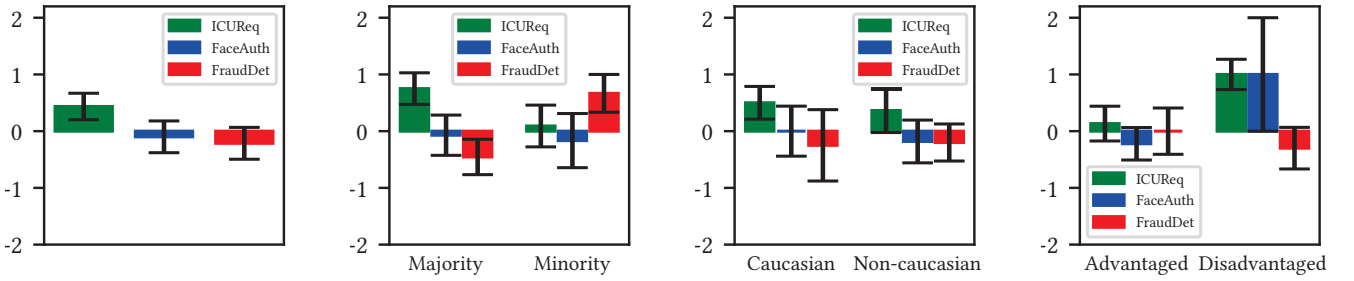
(a) Overall mean preferences

(b) Partitioned by disadvantaged group

(c) Partitioned by participants' ethnicity

(d) Partitioned by self-identified privilege

Figure 2: Comparing participants' XY preferences across the three societal contexts. The comparison is conducted both on all the data, and conditioned on which group was treated as the disadvantaged group by the ML model, the racial background of the participant, and the self-identified privilege status of the participant. Negative (positive) values indicate a preference for model X satisfying EqFDiscRate (model Y satisfying EqOutcome). Error bars represent the standard errors of the mean.



(a) Overall mean preferences

(b) Partitioned by disadvantaged group

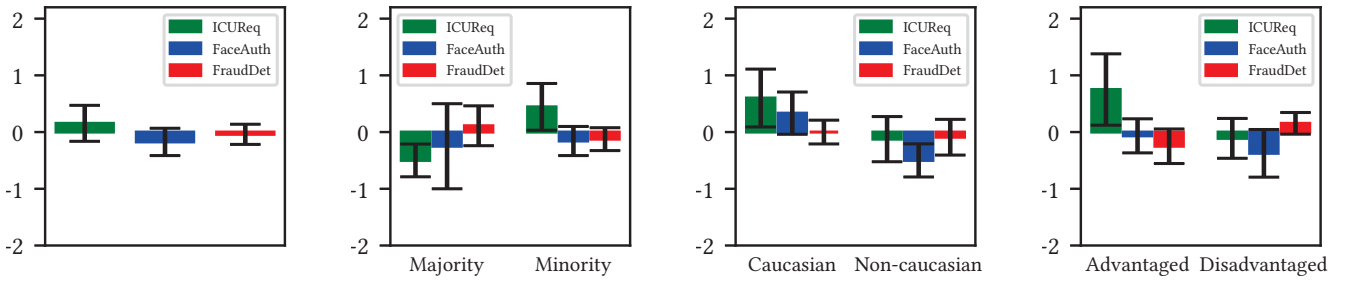
(c) Partitioned by participants' ethnicity

(d) Partitioned by self-identified privilege

Figure 3: Comparing participants' XZ preferences across the three societal contexts. The comparison is conducted both on all the data, and conditioned on which group was treated as the disadvantaged group by the ML model, the racial background of the participant, and the self-identified privilege status of the participant. Negative (positive) values indicate a preference for model X satisfying EqFDiscRate (model Z satisfying EqFOmitRate). Error bars represent the standard errors of the mean.

across the three societal contexts, respectively. As shown in Figure 3 (overall), while it appears that participants often tend to have a stronger preference for model X over model Z in the ICUReq context, the one-way ANOVA test suggests that the difference across societal contexts in participants' XZ preferences is not statistically significant ($p = 0.177$). However, when inspecting into different subsets of responses, we found that when the ML model was significantly biased against the majority group, participants had a significant difference in their XZ preferences across the three societal contexts ($p = 0.044$). Similarly, when participants self-identified as non-privileged in the assigned societal contexts, they also showed significantly different XZ preferences across the three societal contexts ($p = 0.039$). In both cases, the post-hoc Tukey HSD tests identify that as expected, participants in the ICUReq context are significantly more likely to prefer model Z than those in the FraudDet context (XZ preferences when the majority group is disadvantaged by the model: p=0.041; XZ preferences among self-identified non-privileged individuals: p=0.042). On the other hand, one-way ANOVA tests suggest that participants' YZ preferences are not significantly different across the three societal contexts, and this is true both when examining the entire set of responses from all participants, or examining subsets of responses partitioned by the identity of the disadvantaged group, the participants' racial background, and the participants' self-identified privilege status. That is, we find that **H2.2** is partly supported in certain scenarios, but **H2.3** is not supported by our data.

In summary, we find that people's preferences over ML models satisfying different fairness definitions do show significant differences across the three societal contexts under certain conditions. However, in general, evidence which suggests that people tend to prefer an ML that equalizes false discovery rate (or false omission rate) when the false positives are perceived as more risky than false negatives (or false negatives are perceived as more risky than false positives) is lacking.



(a) Overall mean preferences **(b) Partitioned by disadvantaged group** **(c) Partitioned by participants' ethnicity** **(d) Partitioned by self-identified privilege**

Figure 4: Comparing participants' YZ preferences across the three societal contexts. The comparison is conducted both on all the data, and conditioned on which group was treated as the disadvantaged group by the ML model, the racial background of the participant, and the self-identified privilege status of the participant. Negative (positive) values indicate a preference for model Y satisfying EqOutcome (model Z satisfying EqFOmitRate). Error bars represent the standard errors of the mean.

4.3 RQ3: Do people's fairness preferences correlate with their perceived risk differences in different decision mistakes?

Table 4 shows the correlations between participants' reports on individual, societal or overall false positive vs. false negative risk differences and their XY, XZ or YZ preferences, when considering the survey responses that we collected across all three societal contexts. We observe that the risk differences are consistently negatively correlated with XZ preferences, but these correlations are not statistically significant. On the other hand, the XY preferneces and YZ preferences show negligible and insignificant correlations with the risk differences. In other words, **H3** is not supported by our data. We also repeated the same analysis on different subsets of the data considering whether the majority group or the minority group was placed at the disadvantaged position by the ML model, the racial background of the participants, and the self-reported privileged status of the participants. Tables 5 and 6 report the analysis results. Again, we find minimal evidence suggesting that participants' model preferences correlate with their perceived risk differences between different types of ML model mistakes. Finally, we repeat the correlation analysis within each societal context, separately. Again, in neither of the contexts, we observe significant correlations between people's model preferences and their perceived FP vs. FN risk differences (see further details on context-wise correlation analysis results in Appendix B). Since participants' fairness preferences were not strongly correlated to the differences in their risk perceptions for different types of model mistakes, we conclude that human preferences of fairness are not simply driven by the equalized distribution of the most serious harm across groups. Rather, people's fairness preference are highly subjective and nuanced.

Table 4: Pearson correlation statistics between risk differences and fairness prefernces.

Risk Differences (a-b)		Pref.	Pearson Corr.	
a	b		r	p-values
IndFPImpact	IndFNImpact	XY	0.049	0.600
		XZ	-0.146	0.279
		YZ	-0.077	0.554
SocFPImpact	SocFNImpact	XY	0.050	0.591
		XZ	-0.108	0.424
		YZ	0.076	0.562
FPImpact	FNImpact	XY	0.061	0.509
		XZ	-0.153	0.256
		YZ	0.005	0.968

5 DISCUSSIONS

In this study, within the three novel societal contexts we have investigated, we find that people are able to distinguish between different levels of decision risks within each context. However, except for a few specific cases, people's fairness preferences do not vary substantially across contexts, and as a result, we find minimal evidence suggesting that people's risk perceptions of decision mistakes correlate with their fairness preferences. Below, we discuss the implications and limitations of this study.

5.1 Implications

Risk perceptions are aligned with intuitions. Participants' perceptions of risks associated with each type of decisions are highly aligned with our initial assumptions. Table 2 clearly shows that a life-threatening decision is viewed as more risky than the possibility of financial burden in ICUReq. Similarly, security and privacy concerns are considered more dangerous than minor inconveniences of entering credentials for authentication in FaceAuth. The only inconsistency between the data and our assumption is participants' perceptions towards individual risks of false positive and false negative predictions in the fraud detection scenario. The responses suggest that participants found that from the decision subjects' perspective, mistakenly allowing a transaction has more severe consequences than mistakenly denying them. Recent reports claimed that about

Table 5: Pearson correlations between risk differences and fairness perceptions across all the contexts within the sub-groups created by partitioning based on either the group disadvantaged by the model or the participants' ethnicity.

Condition	Risk Differences (a-b)		Preferences	Pearson Correlations	
	a	b		r	p-values
Disadv. Group = majority	IndFPImpact	IndFNImpact	XY	0.255	0.063
			XZ	-0.049	0.771
			YZ	-0.433	0.082
	SocFPImpact	SocFNImpact	XY	0.168	0.224
			XZ	-0.244	0.146
			YZ	-0.357	0.160
	FPImpact	FNImpact	XY	0.262	0.056
			XZ	-0.169	0.317
			YZ	-0.520	**0.032**
Disadv. Group = minority	IndFPImpact	IndFNImpact	XY	-0.084	0.508
			XZ	-0.320	0.169
			YZ	0.040	0.795
	SocFPImpact	SocFNImpact	XY	-0.174	0.170
			XZ	0.212	0.370
			YZ	0.232	0.129
	FPImpact	FNImpact	XY	-0.158	0.212
			XZ	-0.130	0.584
			YZ	0.182	0.237
Ethnicity = Caucasian	IndFPImpact	IndFNImpact	XY	0.044	0.756
			XZ	-0.153	0.465
			YZ	0.193	0.324
	SocFPImpact	SocFNImpact	XY	-0.092	0.514
			XZ	-0.104	0.620
			YZ	0.285	0.142
	FPImpact	FNImpact	XY	-0.021	0.882
			XZ	-0.164	0.435
			YZ	0.334	0.082
Ethnicity = Non-caucasian	IndFPImpact	IndFNImpact	XY	0.055	0.663
			XZ	-0.146	0.425
			YZ	-0.271	0.127
	SocFPImpact	SocFNImpact	XY	0.135	0.283
			XZ	-0.140	0.445
			YZ	0.058	0.747
	FPImpact	FNImpact	XY	0.116	0.358
			XZ	-0.166	0.364
			YZ	-0.104	0.564

one-third of the US population was a victim of fraudulent credit card transactions and about 27% of these transactions resulted in financial losses [13]. The prevalence of such fraudulent activities in fintech solutions could have led the participants to view relaxed defense mechanisms as a greater threat in general even for decision subjects.

Risk perceptions are historically engraved. Although initially expected, we found a minimal influence of recent affairs such as unfairness in policing [44] and eviction moratorium in FaceAuth and FraudDet respectively. The widespread media coverage of the COVID-19 pandemic could have led to mentions of ICU bed scarcity as the primary concerns in ICUReq scenario but the textual risk rating justifications actually covered a wider range of related concerns. It appears that risk perceptions originate from historically formed views of the socio-economic structures rather than being reactive to recent affairs.

Table 6: Pearcon correlations between risk differences and fairness perceptions across all the contexts within the sub-groups of participants based on their self-identification of privilege.

Condition	Risk Differences (a-b)		Preferences	Pearson Correlations	
	a	b		r	p-values
PGM = Advantaged	IndFPImpact	IndFNImpact	XY	0.047	0.705
			XZ	-0.060	0.726
			YZ	-0.136	0.466
	SocFPImpact	SocFNImpact	XY	0.001	0.995
			XZ	-0.185	0.273
			YZ	0.105	0.573
	FPImpact	FNImpact	XY	0.031	0.804
			XZ	-0.139	0.411
			YZ	-0.013	0.945
PGM = Dis-advantaged	IndFPImpact	IndFNImpact	XY	0.055	0.705
			XZ	-0.219	0.354
			YZ	-0.000	1.000
	SocFPImpact	SocFNImpact	XY	0.117	0.419
			XZ	0.135	0.570
			YZ	0.040	0.835
	FPImpact	FNImpact	XY	0.107	0.458
			XZ	-0.076	0.750
			YZ	0.029	0.879

Fairness preferences vary across different subgroups of people. Analyzing the response of Caucasian and non-Caucasian subgroups reveals that Caucasians and non-Caucasians often show contrasting fairness perceptions. Figures 2, 3 and 4 show that in many cases the average Caucasians and non-Caucasians prefer the opposite ends of the fairness trade-off. For example, given the choice between model X and Y, they show opposite preferences in FaceAuth and FraudDet. This suggests that different demographic groups may have different perceptions of fairness, which may be shaped by their past experience. For example, when comparing model Y with model Z in ICUReq, Caucasians voiced concerns against false negative decisions whereas the non-Caucasians preferred EqOutcome. A plausible explanation of this disagreement is that non-Caucasians (typically underprivileged individuals) face disparity in healthcare [31, 39] and advocate for a system that allocates resources equally to the patients even if it is unnecessary.

In addition, we also observed between-group opposing fairness preferences when conditioned on people's self-identification of privilege in at least two societal contexts in each model comparison. The striking relation between perceptions of privilege and fairness indicates that variations in fairness perceptions might be alleviated by balancing end-users' sense of privilege through systemic reforms such as support systems, mass awareness, or appeal mechanisms. For instance, an impoverished community could view cash bail as damaging whereas those with more resources could appreciate this alternative to incarceration. Non-monetary alternatives and fair review/remediation mechanisms are reported to increase the likelihood of favorable perceptions [35]. In other words, instead of computation solutions to fairness, alternative approaches to improve perceptions of privilege among the participants may have the potential to improve fairness perceptions towards an automatic decision-making system.

Perceived fairness fails to achieve equal distributions of the most serious risk. We started our study with the expectation that fairness perceptions would lean towards equalizing the rate of the most detrimental incorrect model predictions across groups. However, our study results suggest that risk perceptions are barely correlated with fairness preferences. Specifically, despite that there is an agreement (often significantly) among participants regarding the magnitude of risks that different model prediction mistakes pose in each societal context, these risk perceptions did not translate into fairness preferences. An explanation for this somewhat counter-intuitive finding could be that laypeople lack formal training on fairness evaluation but are more equipped for risk assessment through their day-to-day experiences. As a result, they were able to distinguish between subtle nuances in context-wise decision risks but failed to properly assess the implications of a fairness metric choice. This is in line with the hypothesis in [43] which suggests that fairness perceptions from crowd-sourced studies may fail to fully capture the true expectations and values of the participants. On the other hand, our findings could also imply that people's fairness preferences in different societal contexts are actually an outcome of factors beyond the perceived risk level differences. In this case, future studies should focus on the identification of such additional factors that shape a participant's fairness perceptions.

Implications for ML model developers. In this work, we study the societal contextual variations in fairness perceptions. Although

our results suggest that participants' fairness preferences are not significantly different across the three societal contexts we considered in this study, we did find that, on average, participants often lean towards different ends of the fairness trade-off across these three contexts when they were asked to select their preferred models. Moreover, we find that certain subgroups within the population do have significantly different fairness preferences across different societal contexts, and different subgroups often disagree with each other in their fairness perceptions. This suggests that *perceived fairness* lacks generalization between different societal contexts and between different segments of people. For the ML model developers, these findings indicate that to select the fairness metric for their model, they can not directly generalize findings on fairness preferences of previous studies [9, 19, 21, 24] that were obtained from only a few societal contexts to their novel contexts. Moreover, using the aggregate fairness preference to determine metrics of model fairness may also invite backlash from subgroups of people due to conflicts with their perceptions. As such, when ML model developers need to incorporate fairness perceptions into their model choices in a novel societal context, they may have to engage in a complete replication of the perceived fairness solicitation process from scratch. During this process, ML model developers should test the perceived model fairness with users from diverse backgrounds to ensure a degree of representativeness of the data collected. They should also actively take the disagreement between different subgroups into consideration to identify ML models that can maximize the collective welfare of different subgroups of people. Since our study suggests that people's risk perceptions do not correlate with their fairness preference, ML model developers should *not* use the crowdsourced assessment of different types of model mistakes as a heuristic in predicting which fairness metric may be preferred by people, either.

5.2 Limitations

We focus only on group fairness definitions in this study since Saha et al. [45] reported that participants were able to comprehend 6 out of 9 statistical parity concepts. Since we didn't provide intervention during the experiment to educate participants about different fairness metrics, our design included easy-to-comprehend group fairness metrics only. As education and ML literacy are peripherally related to societal effects on perceptions and were already studied thoroughly in [52], we opt for relieving the participants from the added cognitive load. Future studies could examine people's perceptions of a wide range of different types of fairness metrics. In addition, we did not adopt complex survey designs such as those used in [9] in our study due to their heavy reliance on expert stakeholders. However, future studies can consider a human-in-the-loop survey design to compare contextual fairness perceptions for different types of stakeholders.

Each hypothetical societal context included in this study is carefully designed so that the underlying technology is easily understood by the participants and involves novel risk considerations. Therefore, we considered under-studied applications of existing technologies while picking the scenarios. Involving societal contexts that concern the latest technologies (e.g., generative AI technologies) in studies like this would be interesting future work, but it also requires additional efforts toward helping participants understand the technology itself before they can evaluate their risk and fairness implications.

By design, all of the participants compared models X and Y, but only a subset of them compared model X to Z while the others compared model Y and Z. Possible alternative study design could ask each participant to compare all three pairs of models, but it will lead to higher cognitive load for the participants. An alternate approach includes limiting participants to compare only one pair of models which will enforce the requirement of a higher number of participants.

6 CONCLUSION

Perceived fairness has gained a lot of traction in recent FairML literature. In this work, we focus on the influence of different societal contexts and their associated decision risks on fairness perceptions. Our experiment shows that laypeople can sense the associated risks with each model prediction. However, this understanding of the consequences of the decision outcomes doesn't appear to directly translate into fairness perceptions. In fact, societal contextual considerations alter the fairness perceptions but to a limited extent. More importantly, these variations are often not directly correlated with the risk perceptions in the related societal context. As a result, we conclude that the influence of societal contexts is not simply an outcome of differences in levels of decision risks, rather it is likely to be a complex combination of many factors.

ACKNOWLEDGMENTS

This material is based upon work supported by the NSF Program on Fairness in AI in Collaboration with Amazon under IIS-1939728 and IIS-2040800. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or Amazon.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.* ProPublica. Retrieved Januray 14, 2022 from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [Online; accessed 05-12-2019].

[2] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & society* 35 (2020), 611–623.

[3] Levon Barseghyan, Francesca Molinari, Ted O'Donoghue, and Joshua C Teitelbaum. 2018. Estimating risk preferences in the field. *Journal of Economic Literature* 56, 2 (2018), 501–564.

[4] David W Bates, Suchi Saria, Lucila Ohno-Machado, Anand Shah, and Gabriel Escobar. 2014. Big Data in Health Care: Using Analytics to Identify and Manage High-risk and High-cost Patients. *Health affairs* 33, 7 (2014), 1123–1131.

[5] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the Conference on Fairness, Accountability and Transparency, FAT 2018 (Proceedings of Machine Learning Research, Vol. 81)*. PMLR, New York, NY, USA, 149–159. http://proceedings.mlr.press/v81/binns18a.html

[6] 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, CHI 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, Montreal, QC, Canada, 377. https://doi.org/10.1145/3173574.3173951

[7] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21, 2 (2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[8] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 10-13, 2015*. ACM, Sydney, NSW, Australia, 1721–1730. https://doi.org/10.1145/2783258.2788613

[9] Hao Fei Cheng, Logan Stapleton, Ruiqi Wang, Paige Bullock, Alexandra Chouldechova, Zhiwei Steven Wu, and Haiyi Zhu. 2021. Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (Eds.). ACM, Virtual Event / Yokohama, Japan, 390:1–390:17. https://doi.org/10.1145/3411764.3445308

[10] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[11] Gregory F. Cooper, Vijoy Abraham, Constantin F. Aliferis, John M. Aronis, Bruce G. Buchanan, Rich Caruana, Michael J. Fine, Janine E. Janosky, Gary Livingston, and Tom M. Mitchell. 2005. Predicting dire outcomes of patients with community acquired pneumonia. *J. Biomed. Informatics* 38, 5 (2005), 347–366. https://doi.org/10.1016/j.jbi.2005.02.005

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Proceedings of the Innovations in Theoretical Computer Science 2012*. ACM, Cambridge, MA, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[13] John Egan. 2023. *Credit card fraud statistics*. BankRate. Retrieved Oct 4, 2023 from https://www.bankrate.com/finance/credit-cards/credit-card-fraud-statistics/

[14] Rebecca A Ferrer and William MP Klein. 2015. Risk perceptions and health behavior. *Current opinion in psychology* 5 (2015), 85–89.

[15] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM* 64, 4 (2021), 136–143. https://doi.org/10.1145/3433949 arXiv:1609.07236

[16] Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019*. ACM, Atlanta, GA, USA, 329–338. https://doi.org/10.1145/3287560.3287589

[17] Nina Gerber, Benjamin Reinheimer, and Melanie Volkamer. 2019. Investigating People's Privacy Risk Perception. *Proc. Priv. Enhancing Technol.* 2019, 3 (2019), 267–288. https://doi.org/10.2478/POPETS-2019-0047

[18] Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III. 2023. The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features. *CoRR* abs/2310.08617 (2023). https://doi.org/10.48550/ARXIV.2310.08617 arXiv:2310.08617 [cs.AI]

[19] Nina Grgic-Hlaca, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 178:1–178:25. https://doi.org/10.1145/3359280

[20] 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 on World Wide Web, WWW 2018*. ACM, Lyon, France, 903–912. https://doi.org/10.1145/3178876.3186138

[21] Nina Grgic-Hlaca, Adrian Weller, and Elissa M. Redmiles. 2020. Dimensions of Diversity in Human Perceptions of Algorithmic Fairness. arXiv:2005.00808 https://arxiv.org/abs/2005.00808

[22] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc., Barcelona, Spain. https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf

[23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*. Curran Associates, Inc., Barcelona, Spain, 3315–3323. https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

[24] 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 Conference on Fairness, Accountability, and Transparency, FAT* '20*. ACM, Barcelona, Spain, 392–402. https://doi.org/10.1145/3351095.3372831

[25] Charles A Holt and Susan K Laury. 2014. Assessment and estimation of risk preferences. *Handbook of the economics of risk and uncertainty* 1 (2014), 135–201.

[26] Hadi Hosseini, Joshua Kavner, Sujoy Sikdar, Rohit Vaish, and Lirong Xia. 2022. Hide, Not Seek: Perceived Fairness in Envy-Free Allocations of Indivisible Goods. *CoRR* abs/2212.04574 (2022). https://doi.org/10.48550/ARXIV.2212.04574 arXiv:2212.04574 [cs.GT]

[27] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*. IEEE, Karachi, Pakistan, 1–6. https://doi.org/10.1109/IC4.2009.4909197

[28] Faisal Kamiran and Toon Calders. 2010. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Citeseer, Haifa, Israel, 1–6.

[29] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012 (Lecture Notes in Computer Science, Vol. 7524)*. Springer, Bristol, UK, 35–50. https://doi.org/10.1007/978-3-642-33486-3_3

[30] Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, Stockholm, Sweden, 2569–2577.

[31] Eun Ji Kim, Nancy R Kressin, Michael K Paasche-Orlow, Lenny Lopez, Jennifer E Rosen, Mengyun Lin, and Amresh D Hanchate. 2018. Racial/ethnic disparities among Asian Americans in inpatient acute myocardial infarction mortality in the United States. *BMC health services research* 18 (2018), 1–10.

[32] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference, ITCS 2017 (LIPIcs, Vol. 67)*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Berkeley, CA, USA, 43:1–43:23. https://doi.org/10.4230/LIPIcs.ITCS.2017.43

[33] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C. Parkes. 2017. Calibrated Fairness in Bandits. *CoRR* abs/1707.01875 (2017). https://doi.org/10.48550/arXiv.1707.01875 arXiv:1707.01875 [cs.LG]

[34] White Case LLC. 2017. *Algorithms and bias: What lenders need to know*. White & Case. Retrieved November 7, 2021 from https://www.whitecase.com/publications/insight/algorithms-and-bias-what-lenders-need-know

[35] Henrietta Lyons, Senuri Wijenayake, Tim Miller, and Eduardo Velloso. 2022. What's the Appeal? Perceptions of Review Processes for Algorithmic Decisions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans, LA, USA, 580:1–580:15. https://doi.org/10.1145/3491102.3517606

[36] Claire Cain Miller. 2015. *Can an Algorithm Hire Better Than a Human?* New York Times. Retrieved January 15, 2022 from https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html

[37] Linh Nguyen, Gerry Gallery, and Cameron Newton. 2019. The joint influence of financial risk perception and risk tolerance on individual investment decision-making. *Accounting & Finance* 59 (2019), 747–771.

[38] U.S. Department of Labor. 2020. *Minimum Wage*. U.S. Department of Labor. Retrieved July 6, 2023 from https://www.dol.gov/general/topic/wages/minimumwage

[39] Nicole Phillips, In-Woo Park, Janie R Robinson, and Harlan P Jones. 2021. The perfect storm: COVID-19 health disparities in US Blacks. *Journal of racial and ethnic health disparities* 8 (2021), 1153–1160.

[40] Emma Pierson. 2017. Gender differences in beliefs about algorithmic fairness. arXiv:1712.09124 http://arxiv.org/abs/1712.09124

[41] Prolific. 2020. *Prolific: Online participant recruitment for surveys*. Prolific. Retrieved Januray 20, 2021 from https://prolific.co/

[42] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* 169, 12 (2018), 866–872.

[43] Samantha Robertson and Niloufar Salehi. 2020. What If I Don't Like Any Of The Choices? The Limits of Preference Elicitation for Participatory Algorithm Design. arXiv:2007.06718

[44] Brianna Sacks, Ryan Mac, and Caroline Haskins. 2020. *Los Angeles Police Just Banned The Use Of Commercial Facial Recognition*. BuzzFeed. Retrieved November 7, 2021 from https://www.buzzfeednews.com/article/briannasacks/lapd-banned-commercial-facial-recognition-clearview

[45] Debjani Saha, Candice Schumann, Duncan C. McElfresh, John P. Dickerson, Michelle L. Mazurek, and Michael Carl Tschantz. 2020. Human Comprehension of Fairness in Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES 2020*. ACM, New York, NY, USA, 152. https://doi.org/10.1145/3375627.3375819

[46] 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, AIES 2019*. ACM, Honolulu, HI, USA, 99–106. https://doi.org/10.1145/3306618.3314248

[47] Hannah Schildberg-Hörisch. 2018. Are risk preferences stable? *Journal of Economic Perspectives* 32, 2 (2018), 135–154.

[48] Michael Warren Skirpan, Tom Yeh, and Casey Fiesler. 2018. What's at Stake: Characterizing Risk Perceptions of Emerging Technologies. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018*, Regan L. Mandryk, Mark Hancock, Mark Perry, and Anna L. Cox (Eds.). ACM, Montreal, QC, Canada, 70. https://doi.org/10.1145/3173574.3173644

[49] 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, KDD 2019*. ACM, Anchorage, AK, USA, 2459–2468. https://doi.org/10.1145/3292500.3330664
[50] Niels van Berkel, Jorge Gonçalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proc. ACM Hum. Comput. Interact.* 3, CSCW (2019), 28:1–28:21. https://doi.org/10.1145/3359130
[51] Niels van Berkel, Jorge Gonçalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Virtual Event / Yokohama, Japan, 245:1–245:13. https://doi.org/10.1145/3411764.3445365
[52] 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, CHI 2020*. ACM, Honolulu, HI, USA, 1–14. https://doi.org/10.1145/3313831.3376813
[53] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018*. ACM, Montreal, QC, Canada, 656. https://doi.org/10.1145/3173574.3174230
[54] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017*. ACM, Perth, Australia, 1171–1180. https://doi.org/10.1145/3038912.3052660
[55] Indre Zliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling Conditional Discrimination. In *Proceeding of the 11th IEEE International Conference on Data Mining*. IEEE, Vancouver, Canada, 992–1001.

A QUESTIONNAIRE

Our experiment randomized the sequence of displaying model X and model Y as well as whether Caucasians are disadvantaged or non-Caucasians. As a result, there were 12 configurations (3 societal contexts, 2 disadvantaged groups, and 2 display sequences). The questionnaire was designed to conditionally display only one configuration to each participant. Following is the questionnaire that corresponds to one of those configurations, i.e., the ICU requirement prediction system where the majority group is disadvantaged and model X is displayed first. For brevity, we omit the consent form shown at the beginning of the survey.

Introduction

Thank you for choosing to participate in this study. Go to next page to begin Part 1 of the survey.

- Machine learning models use past data to make predictions.
  - Training Data: To make a model that predicts whether an admission applicant will be successful in college, you could first gather data from past years about the characteristics, or features of applicants who were or were not able to succeed in that college.
- Machine learning models use patterns in the training data to make predictions.
  - For example, it might find that the applicant's standardized test results combined with participation in extra-curricular activities best predict an applicant's success.
  - The collection of patterns is the model.
- Once the model has been made, it is possible to test how it will perform by applying it to data that was set aside before building the model.
  - For example, we could make the model using data from 2015 and 2016 applications, and see how well it works in the 2017's applications.
- A machine learning model can make two types of mistakes in its predictions.
  - Predict something will happen when in reality it doesn't happen.
    - Example: The model predicts that an applicant will be successful and therefore he/she is accepted for admission. But later it turns out that the applicant failed to succeed.
  - Predict something will not happen when in reality it does happen.
    - Example: The model predicts that a qualified applicant will not be successful.
- As an example, the following table is generated by counting the success predictions of a model about applicants.

Predicted Outcomes	Actual Outcomes (Success in College)		Total Applicants
	Yes	No	
Yes	25	22	47
No	19	34	53
	44	56	100

A.1 Part I: Scenario Description

ICU Requirement Prediction Scenario

The next series of questions will refer to this scenario.

Metropolis General Hospital needs to decide which patients will eventually deteriorate and require ICU support in the next several hours. Since ICU beds are scarce, neither all the patients can or should be put in ICU, nor one should wait until their condition deteriorates. We can think of this as a diagnostic test to early detect risk of deterioration and provide patients necessary care before it is too late.

- Training Data: The hospital created two models by gathering data from past patients.
- Models: The resulting models try to predict whether a patient will or will not require ICU support.

Based on the model picked by Metropolis General Hospital, they can predict which people will require ICU support:

- If the model predicts that a person will require ICU support, he will be put in ICU support.
- If the model predicts that a person will not require ICU support, they will not be put in ICU and will receive general treatment.

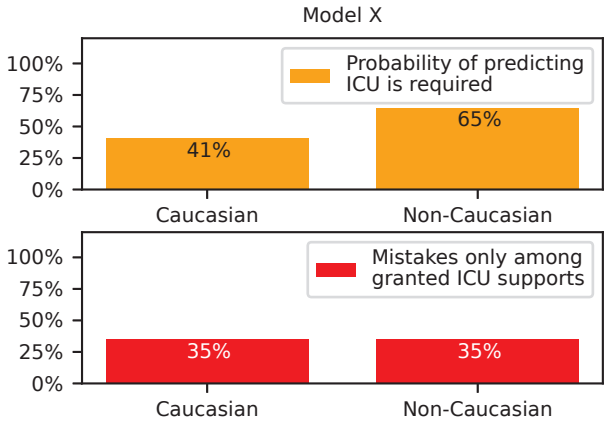
However, neither model they are considering is perfect. They each make mistakes in different ways. In the following questions, we will ask you to compare the performance of these models.

A.2 Model A

Metropolis General Hospital wants to use an automated model to determine who needs to be moved into an Intensive Care Unit (ICU).

Model X is one of the models Metropolis General Hospital is considering. Below are two graphs showing properties of model X. The top graph shows the group-wise **probability of predicting ICU is required** when model X is used. The bottom graph shows the **group-wise percentage of mistakes only among granted ICU supports** by model X.

- **Probability of predicting ICU is required:** The probability of predicting ICU is required is the likelihood of model X saying ICU support will be required.
- **Mistakes only among granted ICU supports:** The mistakes only among granted ICU supports shows the ratio (in percent) of mistakenly granted ICU supports (i.e., model said "ICU is required" but in reality, it was not) to the total amount of granted ICU supports by model X.



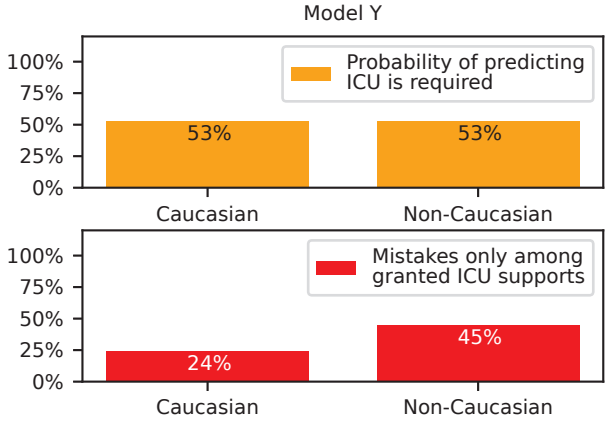
Do you think Model X is fair?
• Very unfair • Mildly unfair • Neither fair nor unfair • Acceptably fair • Very fair
Why? _____
Do you think Model X is biased?
• Very unbiased • Mildly unbiased • Neither biased nor unbiased • Acceptably biased • Very biased
Why? _____
Do you think Model X is useful?
• Completely unusable • Mostly unusable • Neither useful nor unusable • Mostly useful • Very useful
Why? _____

A.3 Model B

Metropolis General Hospital wants to use an automated model to determine who needs to be moved into an Intensive Care Unit (ICU).

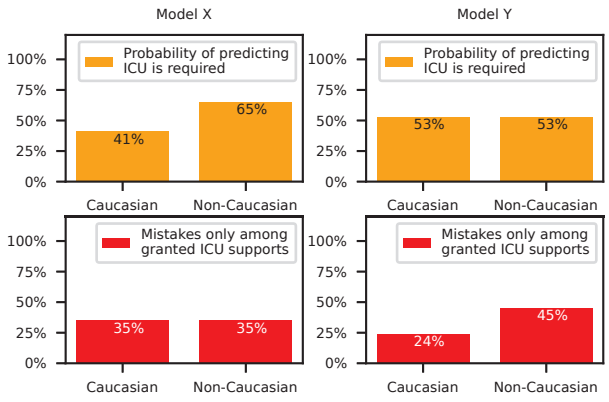
Model Y is one of the models Metropolis General Hospital is considering. Below are two graphs showing properties of model Y. The top graph shows the group-wise probability of predicting ICU is required when model Y is used. The bottom graph shows the group-wise percentage of mistakes only among granted ICU supports by model Y.

- Probability of predicting ICU is required: The probability of predicting ICU is required is the likelihood of model Y saying ICU support will be required.
- Mistakes only among granted ICU supports: The mistakes only among granted ICU supports shows the ratio (in percent) of mistakenly granted ICU supports (i.e., predicted "ICU is required" but in reality, it was not) to the total amount of granted ICU supports by model Y.



Do you think Model Y is fair?
• Very unfair • Mildly unfair • Neither fair nor unfair • Acceptably fair • Very fair
Why? _____
Do you think Model Y is biased?
• Very unbiased • Mildly unbiased • Neither biased nor unbiased • Acceptably biased • Very biased
Why? _____
Do you think Model Y is useful?
• Completely unusable • Mostly unusable • Neither useful nor unusable • Mostly useful • Very useful
Why? _____

Changed Answers



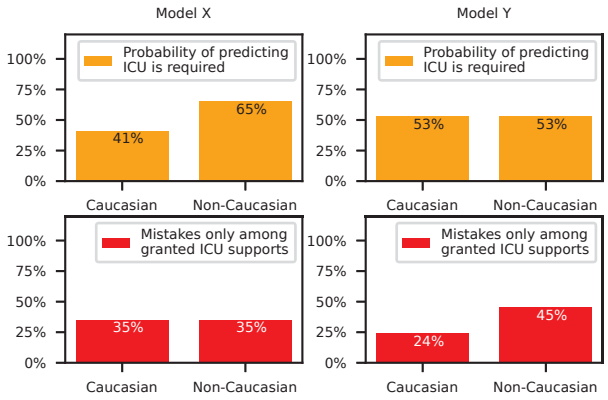
Did seeing the second model (right figure) change the answers you wished to give for the first model (left figure) you saw?

• Yes • No

What aspects of your answer would you wish to change and why?

A.4 Model X and Model Y

Now we're going to ask you to compare the two models you just saw. For reference, the figures are reproduced below.



Which model is more fair, model X or model Y?

• Definitely model X • Probably model X • Models X and Y are equally fair

• Probably model Y • Definitely model Y

Why? _____

Which model is more biased, model X or model Y?

• Definitely model X • Probably model X • Models X and Y are equally biased

• Probably model Y • Definitely model Y

Why? _____

Which model is more useful, model X or model Y?

• Definitely model X • Probably model X • Models X and Y are equally useful

• Probably model Y • Definitely model Y

Why? _____

Given a choice between model X and model Y, which would you choose?

• Definitely model X • Probably model X • Neither model X nor model Y

• Probably model Y • Definitely model Y

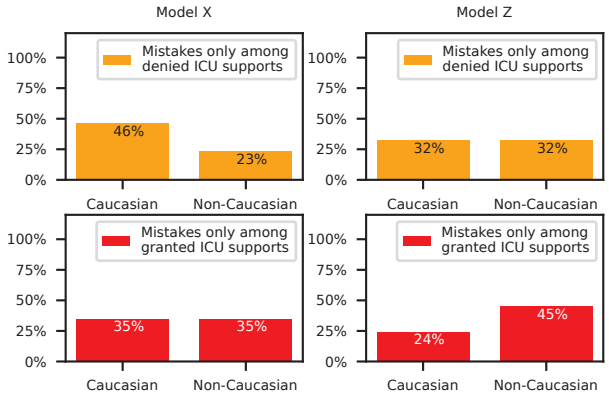
Why? _____

A.5 Model X/Y vs Model Z

You have chosen model X over model Y[6]. Now compare model X with model Z shown below.

The top graph for each model shows the group-wise percent of mistakes only among denied ICU supports.

- Mistakes only among denied ICU supports: The mistakes only among denied ICU supports shows the ratio (in percent) of mistakenly denied ICU supports (i.e., model said "ICU is not required" but in reality it was necessary) to the total amount of denied ICU supports



Which model is more fair, model X or model Z?

• Definitely model X • Probably model X • Both model X and Z are equally fair

• Probably model Z • Definitely model Z

Why? _____

Which model is more biased, model X or model Z?

• Definitely model X • Probably model X • Both model X and Z are equally biased

• Probably model Z • Definitely model Z

Why? _____

Given a choice between model X and model Z, which would you choose?

• Definitely model X • Probably model X • Neither model X nor model Z

• Probably model Z • Definitely model Z

Why? _____

A.6 Context Descriptor Questions

Now, we begin the 2nd part of this survey. Before beginning the questions we shall provide a few definitions and examples to help you understand the questions. Keeping in mind your answers in Part 1, complete these questions so that they best justify your answers in Part 1. For your convenience, your answers from Part 1 are reproduced below. [7]

- The person who directly receives the decision from the automatic decision making system is termed as an "individual".
  - E.g., if automatic decision making were used in deciding college admissions, the applicant will be an "individual".
- Anyone who is indirectly affected by the decision from the automatic decision making system is considered as part of the "society".
  - E.g., if an automatic decision making system is set up to decide whether to grant bail to a defendant, the decision also has an impact on the defendant's family, employer, other stakeholders, potential victims of future crime as well as the entire community. Therefore, all of them are categorized as part of the "society".

[6] Assuming model X was preferred over model Y.

[7] Reproduction of answers to Part 1 is omitted for the sake of brevity.

- In all the following discussions, we consider the harmful impacts of a decision.
- An incorrect decision can have immediate or long term consequences on both the individual and the society.
  - The immediate consequence of an improper bail decision on the individual is imprisonment.
  - The societal impact could be varied, e.g., the defendant's family may be subject to emotional and financial difficulties if the individual is needlessly imprisoned, but the family and others in the community may be harmed if a dangerous individual is released.
- The significance of the harmful impacts of a decision varies with context.
  - The decision to imprison an innocent person has severe impact on his/her life. At the same time, the decision to grant someone a bail who is likely to commit violent crimes later has high harmful impact on the society.
  - On the other hand, if college applicant's success predictions are used to recommend necessary additional mentoring then an incorrect prediction about an applicant's success may have low impact on his later success. But if such a prediction is used to decide college admission, then an incorrect prediction may have high impact on the applicant's future success.

In the following questions, we will ask you about level of severity of individual and social harmful consequences of each type of mistake a model can make in the given scenario.

From the perspective of an individual, how significant are the impacts of mistakenly predicting ICU support will be required?

• High • Moderate • Low

Why? _____

From the perspective of an individual, how significant are the impacts of mistakenly predicting ICU support will not be required?

• High • Moderate • Low

Why? _____

From the perspective of the society, how significant are the impacts of mistakenly predicting ICU support will be required?

• High • Moderate • Low

Why? Please also mention who you considered as part of the "society" for clarity of your answers.

_____

From the perspective of the society, how significant are the impacts of mistakenly predicting ICU support will not be required?

• High • Moderate • Low

Why? Please also mention who you considered as part of the "society" for clarity of your answers.

_____

Do you think there will be high reliance on the automatic decision making system in the ICU requirement prediction scenario (i.e., doctors will start following the decisions blindly)?

• Yes • No

If Metropolis General Hospital noticed that the model is highly inaccurate on African American patients, do you think the doctors will start putting more African American patients into ICU even when the model predicts it is not required?

• Yes • No

Do you think there exists historical disparity in the treatment received by White and African American patients?

• Yes • No

A.7 Self-Identification

If you were the recipient of the decision from an ICU requirement predictor model, do you think you will be advantaged or disadvantaged, relative to the average individual?

• Advantaged • Disadvantaged

A.8 Graph/Context Comprehension

Following questions are related to the graphs shown below. Do not use information you may have seen in other graphs in answering this question.

Assume that Metropolis College has deployed Model X to predict whether an applicant "will be successful" in their college (thus will be accepted for admission) or "will not be successful" (thus will be denied admission). Figure 1 shows the group-wise accuracies. Figure 2 shows the group-wise percentage of mistakes only among "will not be successful" predictions.

Using the information shown in the figures, answer the following questions.

Figure 1: Group-wise Accuracies of Model X

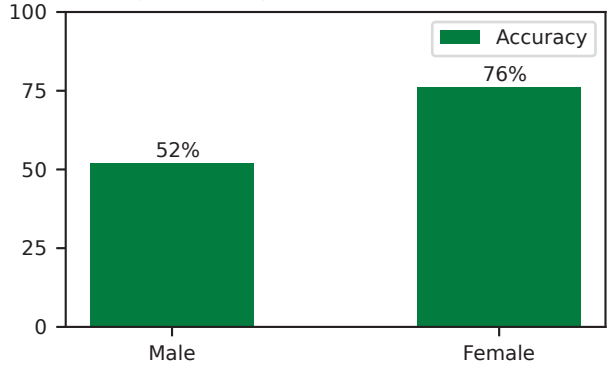
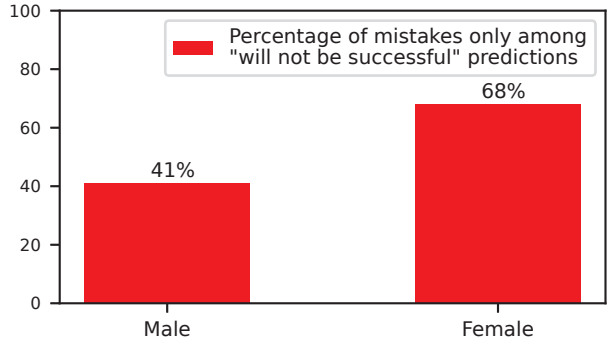
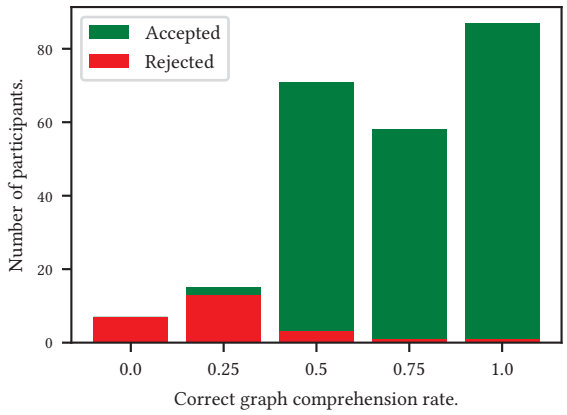


Figure 2: Group-wise Percentage of Mistakes only among "will not be successful" Predictions from Model X

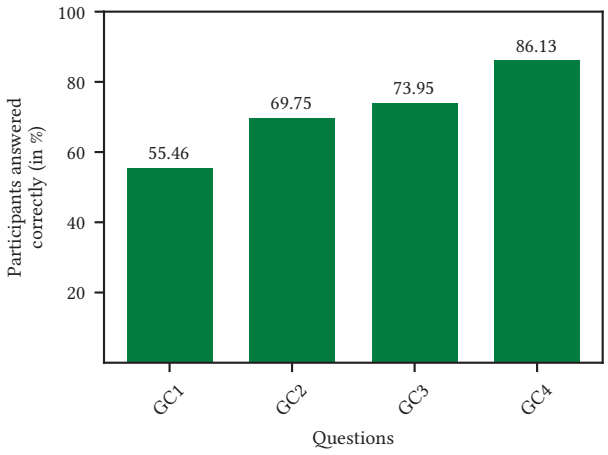


GC1: Which of the following can be inferred from the above figures,

- Model X is more likely to predict correctly for female applicants than male applicants.



(a) Participants in each percentage bin of correctly answering *GC1-4*. The rejection criteria can be found in Section 4.



(b) Percent of correct answers among the accepted responses in each graph comprehension question.

Figure 6: Statistics of graph comprehension questions.

- Model X is more likely to predict correctly for male applicants than female applicants.
- Model X is equally likely to predict correctly for female applicants and male applicants.
- None of the above can be inferred from the figures.

GC2: Which figure was used to infer the answer of the question above?

• Figure 1 • Figure 2

GC3: Which of the following can be inferred from the above figures,

- More male applicants will be accepted for admission and fail than female applicants.
- Female applicants will be more likely to mistakenly be rejected than male applicants.
- Male applicants will be more likely to mistakenly be accepted than female applicants.
- Male applicants will be more likely to be accepted than female applicants.

GC4: Which figure was used to infer the answer of the question above?

• Figure 1 • Figure 2

Optional

The following questions are optional. These will help us analyze your response better. If you don't want to provide answers to the following, feel free to skip to the end of the survey.

Highest Level of Education:

List of items ommitted.

Occupation:

List of items ommitted.

A.9 Graph Comprehension Statistics

As both training and a quality check, we ask the graph comprehension questions shown in A.8. The participants are shown a graphical representation of group-wise performances of a model with textual explanations. They answered two multiple-choice inference questions (*GC1 and GC3*) and two inference follow-up questions (*GC2 and GC4*). Figure 6a indicates that 90% (216) of the participants correctly answered half or more of the graph comprehension questions. Similarly, Figure 6b shows each individual inference question and follow-ups were answered on average 65% and 76% accurately respectively. Since the percentages are significantly higher than random selection of 25% and 50% respectively, it indicates that the visual and textual aids successfully helped the participants in making an informed decision. Graph comprehension was a part of the rejection criteria described in Section 4. The exceptions in percentage bins 25% and 50% in Figure 6a are due to thought-provoking and low-quality open-text responses respectively.

B SOCIETAL CONTEXT-WISE RELATIONS BETWEEN RISK PERCEPTIONS AND FAIRNESS PREFERENCES

Table 7: Context-wise Pearson correlation statistics between risk differences and fairness preferences.

Context	Risk Differences (a-b)		Preferences	Pearson Correlations	
	a	b		r	p-values
ICUReq	IndFPImpact	IndFNImpact	XY	0.132	0.442
			XZ	-0.190	0.385
			YZ	-0.550	0.052
	SocFPImpact	SocFNImpact	XY	0.337	0.044
			XZ	-0.018	0.934
			YZ	0.321	0.285
	FPImpact	FNImpact	XY	0.312	0.064
			XZ	-0.131	0.550
			YZ	-0.210	0.491
FaceAuth	IndFPImpact	IndFNImpact	XY	-0.155	0.320
			XZ	0.234	0.320
			YZ	0.212	0.333
	SocFPImpact	SocFNImpact	XY	-0.207	0.182
			XZ	-0.355	0.124
			YZ	0.133	0.546
	FPImpact	FNImpact	XY	-0.220	0.156
			XZ	-0.020	0.934
			YZ	0.212	0.331
FraudDet	IndFPImpact	IndFNImpact	XY	-0.033	0.844
			XZ	-0.162	0.581
			YZ	0.047	0.824
	SocFPImpact	SocFNImpact	XY	0.118	0.475
			XZ	0.176	0.547
			YZ	-0.062	0.767
	FPImpact	FNImpact	XY	0.061	0.711
			XZ	-0.003	0.992
			YZ	-0.018	0.931