Discrimination Exposed? On the Reliability of Explanations for Discrimination Detection

Julian Skirzynski
UCSD
La Jolla, USA
jskirzynski@ucsd.edu

David Danks
UCSD
La Jolla, USA
ddanks@ucsd.edu

Berk Ustun
UCSD
La Jolla, USA
berk@ucsd.edu

Abstract

Many rules and regulations in areas such as lending or hiring cast explanations as a safeguard against algorithmic discrimination. The underlying assumption is that, for a given model, individuals could inspect explanations of predictions to contest discriminatory outcomes or flag the model as biased. This is a common-sense assumption that is easy to comply with. However, it is also very difficult to corroborate because it relies on unverifiable causal assumptions about which variables constitute proxies, and how the proxies affect the outcome variable. In order to make accurate claims, individuals must assume what these relationships are and be able to detect the proxies and their influence on predictions from the explanations. In this work, we study whether explanations help users detect algorithmic discrimination. We formalize the problem of detecting discrimination and introduce a synthetic robot classification task with known discrimination labels, overcoming real-world limitations where ground truth is unknown. We then design a user study that validates participants' understanding of explanations, protected attributes, proxies, and their causal strength, and isolates the utility of explanations. Our results show that human experts cannot reliably use explanations to flag discriminatory predictions irrespective of how much information about the predictions they have. Because the reliability of detection is low even under idealized conditions, these findings underscore the need for alternative anti-discrimination safeguards in practical settings.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *HCI design and evaluation methods*; • **Social and professional topics** → **Governmental regulations**; • **Security and privacy** → **Social aspects of security and privacy**; • **Computing methodologies** → **Machine learning**.

Keywords

interpretable machine learning, explanations, algorithmic discrimination, anti-discrimination policy, fairness, proxies

ACM Reference Format:
Julian Skirzynski, David Danks, and Berk Ustun. 2025. Discrimination Exposed? On the Reliability of Explanations for Discrimination Detection. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3715275.3732167

This work is licensed under a Creative Commons Attribution 4.0 International License.
FAccT '25, Athens, Greece
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1482-5/25/06
https://doi.org/10.1145/3715275.3732167

1 Introduction

Machine learning models are routinely used to automate decisions that affect people – be it to approve a loan [74], an insurance claim [35], or a public service [72]. Over the past decade, it has become clear that models can lead to inadvertent discrimination, as their predictions or performance can change across *protected attributes* such as sex, age, or race [11, 64, 66]. In applications like lending and hiring, such effects arise from *indirect discrimination* [67] as models without protected attributes (e.g., sex) assign predictions through proxies (e.g., credit_history).

To protect consumers from algorithmic discrimination in "high-risk" applications, legislators increasingly propose and enact rules and regulations that mandate a "right to explanation"; see e.g., the European Union [68, 69], Brazil [13], Korea [36], and the United States [1, 2]. These mandates are commonly based on the belief that consumers can use information in an explanation to contest wrongful decisions. For example, the explanations might enable them to notice that an automated "*decision is affected by a (legally) protected attribute.*"[73] or that "*the model would not have produced a beneficial decision when we alter irrelevant behavioral or group attributes*" [44].

Despite their widespread use in consumer protection laws, there is little evidence about whether explanations could actually be used to protect against algorithmic discrimination. Simply put, we lack answers to questions such as:

- *"Can consumers use explanations to detect an unfair decision?"*
- *"Can auditors use explanations to detect unfair models?"*
- *"How does the reliability of bias detection in each context depend on the information available to consumers and auditors? (e.g., does it change if they know the protected class of each point, or if they know a proxy variable that could lead to discrimination?"*

The absence of evidence is surprising since the earliest laws with a right to an explanation in a major consumer application were enacted over fifty years ago [see e.g., the adverse action provision in ECOA 65]. The reason why there is no evidence yet is that explanations are extremely challenging to validate in applied settings. First, there is confounding. Any failure of explanations could stem from: (1) incorrect beliefs about what constitutes a proxy among users; (2) a lack of guidelines on how to detect discrimination with explanations; or (3) a misuse of explanations. To rule out (1) and (2), we need a degree of control that is typically impossible in real-life settings. Second, there is an issue of scope. A study with experts that clearly shows failures arising from (3) could be relevant only for a narrow scope; e.g., it could only happen for credit decisions in the US. Finally, there is an ambiguity that stems from unverifiable causal assumptions and chance (e.g., which variable is a proxy, whether it affected a given decision, etc.). In this way, any claim lacks a clear ground-truth.

In this paper, we formalize the *discrimination detection* task and test if explanations can support human experts. Our work seeks to distill the most basic assumptions behind non-direct discrimination and create a minimal formal setup that enacts them. We also aim to identify and control for confounding factors and explanation *failure modes* to attribute detection performance directly to the explanations. Our goal is to determine the conditions under which explanations meaningfully support detecting discrimination. This evidence will inform whether alternative mechanisms are needed or if additional constraints on explanations suffice. Our main contributions include:

1. We formalize the problem of *discrimination detection* with explanations. We list the assumptions needed to assess if explanations help users detect discrimination and map them to failure modes of explanations in supporting discrimination claims.
2. We design a user study to evaluate the reliability of discrimination detection with explanations. Our design provides a sandbox environment where we can rule out key failure modes related to human interaction. We also control factors that would otherwise confound the results (e.g., the identity of the proxy).
3. We show through controlled human-subject experiments that people fail to perform reliably, irrespective of which explanations they see and how much knowledge about the problem they have. By showing that explanations fail to deliver on a simple task, these results stress the need for alternative solutions.

Related Work. We study explanations as a safeguard for algorithmic discrimination in domains such as lending and hiring [5, 30, 47]. In such domains, treatment parity requires models to output similar predictions across protected groups. In practice, models may violate this principle as a result of indirect discrimination via proxy variables [see e.g., 67, for a review]. These issues have motivated extensive work on detecting and mitigating discrimination – e.g., methods to train models that do not discriminate [see e.g., 80], to identify proxies in a third-party audit [see e.g., 4], and to enable reporting group or individual discrimination [21]. One widely-proposed solution to counter discrimination is to use explanations of model predictions [6, 8, 24, 27, 42, 49, 76]. A popular proposal is to detect discrimination through counterfactual explanations [6, 8, 24, 27, 42, 49, 76] that convey the minimal set of feature changes that modify the prediction. Counterfactual explanations have already shown marginal improvements in decision-making [23, 43, 45, 71, 77] and debugging model behavior [3, 50, 61]. Here, we study them in the task of detecting discrimination.

Our work is related to a stream of research on how humans interact with explanations [see e.g., 10, 15–19, 39, 40, 75, 78]. There is little work on using explanations to assess discrimination, with most works focusing on issues that can arise when generating explanations [e.g., lack of fidelity 9, 22, 54]. One of the key challenges of this question is a mismatch in *scope*. Assessing discrimination involves questions about causality at a population level. In contrast, explanations usually provide answers about model behavior at the instance level. The few studies on using explanations to detect discrimination at the instance level focus on tasks where models use protected characteristics [see e.g., 26, 56] and suggest that explanations help people spot discriminatory predictions. We study whether explanations work in the tasks envisioned by regulators, where users need to detect discrimination of individual predictions based on *proxy variables*. Our work formalizes discrimination by adopting a causal notion of fairness [see e.g., 38, 57] – e.g., "would my prediction change if I belonged to a different protected group." Our results align with the emerging picture from studies such as Goyal et al. [33]. Those authors demonstrate that users cannot use explanations to make less discriminatory decisions when discrimination comes from proxy variables. We explicitly highlight that users cannot tell which predictions are fair and which are not based on explanations. In this way, our research adds to a stream of prior results that show explanations influence perceptions of fairness. These prior studies demonstrate the importance of factors such as the prediction task [7], explanation type [12, 46, 60, 79], and information content [7, 12, 52, 62, 63].

2 Framework

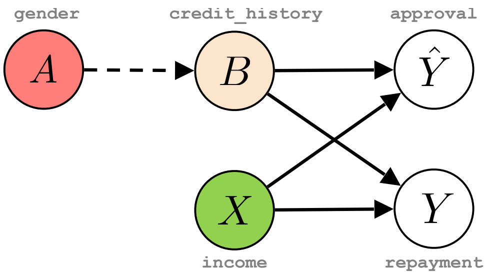


Figure 1: Causal diagram for discrimination detection. Model $f : B \times X \longrightarrow \hat{Y}$ returns prediction $\hat{Y}$ of an outcome variable Y given input proxy B and features X. We seek to determine if model predictions change with respect to protected attribute A through its proxy B, which is assumed to be related to the outcome Y. For example, in loan approval predictions ($\hat{Y}$), the model uses an individual's income (X) and credit history (B) as inputs. Gender (A) could affect credit history due to differences in credit scores or the intensity of credit usage found between men and women [see e.g, 48].

We consider a task where (un)fairness involves whether a model's predictions change based on a *protected attribute* A (e.g., `gender`). Specifically, we examine if altering the protected attribute would result in different model outputs for individual predictions. We assume that the features and outcomes in this task obey the causal relationships shown in Fig. 1. The model $f : X \times B \longrightarrow \hat{Y}$ is a deterministic function that predicts an outcome Y (e.g., `repayment`). Here, B denotes the proxy variable, and X denotes inputs that are independent of the protected attribute (e.g., X = `income`). The model satisfies two assumptions:

1. *Indirect Discrimination.* Model f does not use the protected attribute as input, but its predictions may change as a result of a *proxy* B (e.g., B = `credit_history`) for the protected attribute. [4, 67]
2. *Business Necessity.* The proxy B improves predictive accuracy, else it could be removed from the list of features [31]

These assumptions are met by the vast majority of models in applications where we care about discrimination. Models that use protected attributes as inputs would violate *treatment disparity* [11] by assigning different predictions to different groups. Therefore, they are typically omitted [11, 70]. In cases where the proxy did not improve accuracy, a model owner could avoid scrutiny by training a model without it.

Characterizing Discrimination. We characterize the fairness of predictions based on a (relaxed) notion of *counterfactual fairness* [38].

Definition 1. Given a fairness threshold $\delta \in [0, 1]$, we say that the prediction of model $f : X \times B \times A \to Y$ at point (x, b, a) is δ-counterfactually fair if changing the protected attribute can change the prediction by at most δ:

$$\Big| \underbrace{\Pr(\hat{Y}_a = f(x, b) \mid x, b, a)}_{\text{Current Prediction with } A = a} - \underbrace{\Pr(\hat{Y}_{a'} = f(x, b) \mid x, b, a)}_{\text{Counterfactual Prediction with } A = a'} \Big| \leq \delta$$

Here, $\hat{Y}_a$ is the current prediction of the model, $\hat{Y}_{a'}$ is the counterfactual prediction in a world where we set the protected attribute of the individual to $A = a'$, often written as $A \leftarrow a'$, and $\delta \in [0, 1]$ is a *fairness threshold* that represents the maximum degree to which a fair prediction can change as a result of this intervention.

We can set $\Pr(\hat{Y}_a = f(x, b) \mid x, b, a) = 1$ since no intervention is required. We can compute $\Pr(\hat{Y}_{a'} = f(x, b) \mid x, b, a)$ by setting $A \leftarrow a'$ and propagating its effect on the proxy B. For the causal structure in Fig. 1, $\Pr(\hat{Y}_{a'} = f(x, b) \mid x, b, a)$ can be written as:

$$\sum_{b' \in B} \underbrace{\Pr(\hat{Y} = f(x, b) \mid x, b, a')}_{\text{Prediction for } b'} \cdot \underbrace{\Pr(B = b' \mid a')}_{\text{Proxy Strength}}$$

Thus, a prediction $\hat{Y} = f(x, b)$ is δ-counterfactually fair if

$$\Big| 1 - \sum_{b' \in B} \Pr(\hat{Y} = f(x, b) \mid x, b, a') \cdot \Pr(B = b' \mid a') \Big| \leq \delta$$

The left hand-side of this inequality is the probability that the prediction flips as we intervene on the protected attribute. In what follows, we denote it as $p^{\text{flip}}_{x,b,a}$ and refer to it as the *flip rate*. The maximum flip rate we tolerate is defined by the fairness threshold δ. This threshold can be set on a task-by-task basis. For example, if we work with a model to screen resumes, then we could set $\delta = 0.2$ to reflect the "4/5ths rule" in U.S. employment discrimination law [28].[1] We write $p^{\text{flip}}_i := p^{\text{flip}}_{x_i,b_i,a_i}$, whenever it is clear which x, b, a is being discussed.

Discrimination Detection with Explanations. Many rules and regulations mandate explanations as an anti-discrimination measure, based on the assumption that they help users detect and contest unfair predictions. Testing this assumption is challenging because it requires both a verifiable ground-truth for discrimination, despite unknown causal relationships between the proxies, protected attributes, and the predicted outcome, as well as capturing the claims individuals make about predictions. To deal with these unknowns, we formalize our problem as a detection task – and measure users' detection accuracy. Given a model f, we associate each instance with two labels:

- $g_{i|f,\delta} := \mathbb{I}[p^{\text{flip}}_{x_i,b_i,a_i} > \delta]$, i.e., a "ground-truth" label that encodes actual discrimination in the prediction; it is an indicator the prediction is not δ-counterfactually fair.
- $\hat{g}_{i|f,e_i}$, i.e. a "prediction" that encodes a user's claim a prediction is unfair after seeing the explanation e_i.

In what follows, we write $g_i := g_{i|f,\delta}$ and $\hat{g}_i := \hat{g}_{i|f,e_i}$ when their dependencies are clear from context.

Under this model, the flip rate $p^{\text{flip}}_{x,b,a}$ is fixed for individuals with identical features (x, b, a). However, the actual outcome of intervening on the protected attribute is random. We assume it follows a Bernoulli distribution $G_i \sim \text{Bern}(p^{\text{flip}}_{x,b,a})$. In this case, we can interpret g_i in terms of hypothetical proportions: if we were to change the protected attribute for N individuals with features $(x, \hat{b}, a)$, then a δ-counterfactually fair model would assign different predictions to at most δN individuals. Since users only see one prediction for instance i, we interpret $\hat{g}_{i|f,e_i}$ as their *personal probability* the prediction would change under an intervention on A [see e.g., 25, for more details]. [2] We write this as $\hat{g}_{i|f,e_i} \approx \mathbb{I}[p^{\text{flip}}_i > \delta]$.

Measures. Given a model h, and a set of n individuals $\{(x_i, b_i)\}^n_{i=0}$ and ground-truth labels $\{g_i\}^n_{i=0}$, we can evaluate the reliability of discrimination claims $\{\hat{g}_i\}^n_{i=0}$ using standard performance measures for binary classification:

- $\text{TPR}(\delta) = \frac{|\{i:\ \hat{g}_i = g_{i|\delta} = 1\}|}{|\{i:\ g_{i|\delta} = 1\}|}$, which measures how often users correctly identify discriminatory predictions;
- $\text{FPR}(\delta) = \frac{|\{i:\ \hat{g}_i \neq g_{i|\delta} = 0\}|}{|\{i:\ g_{i|\delta} = 0\}|}$, which measures how often users incorrectly label a fair prediction as discriminatory;
- $\text{PPV}(\delta) = \frac{|\{i:\ \hat{g}_i = g_{i|\delta} = 1\}|}{|\{i:\ \hat{g}_i = 1\}|}$, which indicates the internal reliability of discrimination claims. In other words, PPV is the proportion of all claims where the flagged predictions are really discriminatory.

We expect the following:

- *Contesting Discriminatory Predictions*: Explanations can support individual claims when the claims are aligned with ground-truth labels. In this case, we should have that $\hat{g}_{i|f,e_i} = g_{i|f,\delta}$ for any explanation e_i where δ may change across users. We would want to observe detection that is always correct, i.e., $\text{PPV}(\delta) = 100\%$, finds all cases of discrimination, i.e., $\text{TPR}(\delta) = 100\%$, and makes no false alarms, i.e., $\text{FPR}(\delta) = 0\%$. In practice, we may state that explanations could help detect discrimination if we observe a PPV of 90% meaning most of participants' claims are indeed warranted.
- *Identifying Discriminatory Models*: Explanations could also support claims that a model discriminates by checking if the proportion of unfair predictions over a set of instances exceeds a model-level threshold τ. This use case provides some room for incorrect claims at the instance level. It is sufficient to estimate if the model discriminates for over $\tau\%$ of predictions. A model that clearly discriminates can tolerate many false alarms while still being correctly identified as discriminatory. Conversely, a clearly fair model can withstand some missed discriminatory cases. The closer the true discrimination rate is to τ, the more reliable individual detection needs to be.

[1] In what follows, we remain agnostic about the value of δ and report findings for all possible thresholds $\delta \in [0, 1]$.

[2] In this case, $\hat{g}_i = 1$ could be seen as indicating a sufficiently large change in subjective strength of belief.

Failure Modes. Users may fail to detect discrimination with explanations due to flawed beliefs or flaws in explanations. Given model h and an explanation, the user may claim $\hat{g}_i \neq g_i$ because:

Remark 1 (Recovery)**.** Users may be given an explanation that does not reveal the prediction changes with the proxy even though $f(x_i, b_i) \neq f(x_i, b_i')$ for $b_i' \neq b_i$. This is because there exist many different explanations for the same prediction, e.g., e_i, e_i' such that e_i presents no information about the proxy but e_i' does [14, 37]. This could lead the user seeing e_i erroneously determine that the counterfactual prediction never changes, i.e., $\Pr(\hat{Y}_{a_i'} = f(x_i, b_i) \mid x_i, b_i, a_i) = 1$, and the prediction is always fair.

Remark 2 (Misinterpretation)**.** Users may not know how to use explanations to support claims about discrimination, i.e., to assess the flip rate p_i^{flip}. Even if they do, they might not know how to extract that information from explanation e_i (e.g., there is no principled way of doing that when e_i is a feature attribution explanation).

Remark 3 (Misspecified Beliefs about Causal Mechanism)**.** Users may have incorrect beliefs about the proxy strength $\Pr(B \mid A)$, and incorrectly estimate the flip rate p_i^{flip}. With a fixed δ, this may lead them to become too sensitive or too lenient on discrimination, making erroneous claims.

Remark 4 (Knowledge of Protected Class)**.** Users may not know the true value of the protected attribute $A = a_i$ and think it is $A = a_i' \neq a_i$. This may lead them to estimate the incorrect flip rate, e.g., $1 - p_i^{\text{flip}}$ instead of p_i^{flip}, and make inaccurate claims.

Remark 5 (Misspecified Causal Beliefs)**.** Users may assume causal relationships that differ from those in Fig. 1. As a result, they may fail to detect discrimination if they believe B is not a proxy, or misattribute discrimination when given an explanation that highlights $f(x_i, b_i) \neq f(x_i', b_i)$ and they believe X is a proxy.

These failure modes are barriers to reliable detection as well as attribution. Each time we may find that explanations fail, we could attribute the failure to one of the listed causes. We can remedy the failure modes 1, 2 and 4 by designing better algorithms and procedures (e.g., methods to find all explanations, and procedures to collect protected attributes). The latter two modes pertain to issues that are inherently human and will change across users and tasks.

3 Experimental Design

We describe an experimental design to evaluate the reliability of explanations as a tool for aiding discrimination detection. Our design is explanation-agnostic and may be adapted to any explanation method by changing the instructions and the visual materials. We consider a simple task where: (1) we can teach participants the skills that we expect from auditors and verify their understanding through comprehension checks; (2) we can manipulate and elicit participant's beliefs in the causal model from Fig. 1; (3) we can collect data to evaluate fairness under different assumptions and use cases (e.g., for all $\delta \in [0, 1]$, with or without access to protected attributes, etc.).

Robot Classification Task. We consider a task where participants are asked to audit a model that predicts the reliability of fictional robots for NASA. The model was created to inform NASA's purchasing decisions by identifying which robots are reliable versus defective. While robot reliability is determined by their body parts, the two manufacturers, `Company X` and `Company S`, design their robots with slightly different components. This difference could lead to discrimination in the model's predictions with respect to the manufacturing company. Since NASA is legally prohibited from making decisions based on the company, participants must determine if the model's predictions are discriminatory or not.

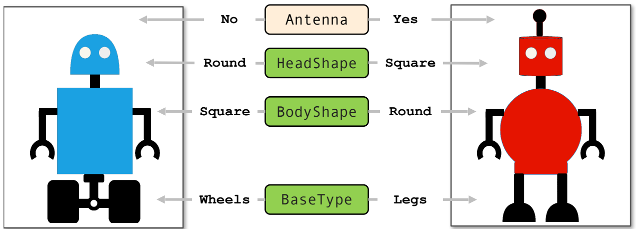


Figure 2: Overview of robot characteristics. We show two robots to cover all possible values of each characteristic. Our model predicts that each robot is reliable or defective using dummy variables $B = \mathbb{I}[\texttt{Antenna} = \texttt{Yes}]$, $X_1 = \mathbb{I}[\texttt{HeadShape} = \texttt{Round}]$, $X_2 = \mathbb{I}[\texttt{BodyShape} = \texttt{Round}]$ and $X_3 = \mathbb{I}[\texttt{BaseType} = \texttt{Wheels}]$.

We cast the identity of the company as our protected attribute A. We assume that the model predicts that a robot is reliable using a set of four salient characteristics shown in Fig. 2, namely: `Antenna`, `HeadShape`, `BodyShape`, `BaseType`. We represent the input variables as: $B := \mathbb{I}[\texttt{Antenna} = \texttt{Yes}]$, $X_1 := \mathbb{I}[\texttt{HeadShape} = \texttt{Round}]$, $X_2 := \mathbb{I}[\texttt{BodyShape} = \texttt{Round}]$, $X_3 := \mathbb{I}[\texttt{BaseType} = \texttt{Wheels}]$. In this setup, we have $2^4 = 16$ distinct combinations of input variables (B, X), and 32 distinct robots (A, B, X). We control all quantities that affect the discrimination by specifying the model's predictions for each robot and the prevalence of each robot (see Table 3 in Appendix B).

We can arbitrarily increase the number of distinct robots to show participants by introducing spurious features. In our case, we introduce $\texttt{Paint} \in (\texttt{Red}, \texttt{Blue})$. In this way, we can ensure that participants are shown new kinds of robots. This is crucial for three reasons: it prevents learning effects from seeing the same robot multiple times, which ensures that decisions are based on feature relationships rather than memorized patterns, and captures real-world tasks where each case presents unique characteristics.

We determine the ground-truth reliability for each robot Y by the random process:

$$
\begin{aligned}
A, X_1, X_2, X_3 &\sim \text{Bernoulli}(0.5) \\
B \mid A &\sim \text{Bernoulli}(p_{B|A}) \quad \text{where } p_{B|A} \text{ is set as per Table 1} \\
Y &\sim \text{Logistic}(B + X_1 + X_2 + X_3).
\end{aligned}
$$

We predict the reliability of each robot using a linear classifier that outputs "Reliable" for robots with an `Antenna` and one of the following: a `Round HeadShape`, a `Round BodyShape`, or `Wheels`:

$$
\begin{aligned}
f(B, X) = &\text{sign}(6B + 4X_1 + 4X_2 + 3X_3 - 8) = \\
&\mathbb{I}[B \text{ AND } (X_1 \text{ OR } X_2 \text{ OR } X_3)].
\end{aligned}
$$

Given our labels, this model has an accuracy of 88% over all possible robots.

	Proxy Strength		**Flip Rate**	
Regime	$A = 0$	$A = 1$	$A = 0$	$A = 1$
Weak	5%	10%	10%	5%
Medium	5%	55%	55%	45%
Strong	5%	95%	95%	90%

Table 1: Overview of parameters determining discrimination claims under each proxy regime. Proxy strength denotes $\Pr(B = 1 \mid A)$, whereas flip rate shows possible values of $p^{\text{flip}}_{x,b,a}$ when $f(x, b) \neq f(x, 1 - b)$. In other cases, the flip rate is 0.

Discrimination. Under the causal model and features we defined in our task, predictions have at most three flip rates $p^{\text{flip}}_{x,b,a}$. These flip rates are either 0 (if changing the proxy does not flip the prediction) or equal to $1 - \Pr(B = b \mid a')$, otherwise. This shows that the flip rate depends solely on $\Pr(B \mid A)$. We vary the strength of this relationship across three regimes (see Table 1) to evaluate how proxy strength affects discrimination detection and claims $\hat{g}_{i|h,\delta}$. This variation is crucial because real-world proxies range from weak correlations (e.g., zip codes as proxies for race) to almost perfect proxies (e.g., height as a proxy for gender). By testing different proxy strengths, we can assess whether participants' performance varies with proxy obviousness. In what follows, we also remain agnostic about the value of δ and evaluate the potential to detect discrimination over all possible thresholds $\delta \in [0, 1]$.

Explanations. To provide a label $\hat{g}_i$ and decide if the prediction $f(x_i, b_i)$ is discriminatory, users must estimate the flip rate p^{flip}_i and compare it to their fairness threshold δ. When users have correct assumptions about the proxy strength and causal structure, this requires checking whether changing the proxy from b_i to $1 - b_i$ flips the prediction. We test whether explanations help with this by comparing two types of explanations: e_i that include information about the proxy variable b_i (potentially revealing if the prediction flips with the change of the proxy and $f(x_i, b_i) = f(x_i, 1 - b_i)$), and explanations e'_i that do not use b (providing no insight about the flip). In this way, we address Remark 1.

Procedure. We implemented our task into an online user study that is fully controllable and addresses all failure modes from Section 2. Our study consists of four phases, shown in Fig. 3. The Training and Anchoring phases address Remark 2 and endow participants with the knowledge we would expect from auditors. The Elicitation phase directly measures participants' beliefs about proxy strength and protected attributes, addressing Remarks 3 and 4.

In this way, our setup allows us to evaluate $\hat{g}_i$ across different fairness thresholds δ and different proxy strengths under the causal structure from Fig. 1. This is because we can recompute the ground-truth labels $g_{i|h,\delta}$ using arbitrary values of a_i, δ and $\Pr(B = b \mid a_i)$. As a a result, we may also assess the impact of incorrect causal beliefs (and address Remark 5) by comparing claims $\hat{g}_i$ to g_i in the most beneficial scenario, where we assume participants have both the correct knowledge about protected attributes and the causal mechanism.

4 Experimental Evaluation

In this section, we present a user study where we evaluate whether explanations can effectively support discrimination detection in algorithmic decision-making. In particular, our experiment sought to determine if individuals could use explanations to make reliable discrimination claims across use cases in consumer protection. Our specific research questions include:

RQ1 Can participants reliably detect when a model discriminates for an individual? If so, this would suggest that explanations are an effective mechanism to exercise individual rights (e.g., to contest predictions that are unfair).

RQ2 Can participants reliably detect when a model discriminates after seeing explanations for its predictions over a representative population? If so, this would suggest that explanations could serve as an effective mechanism to audit models.

RQ3 How does the reliability of discrimination claims depend on the information that is available to participants? In particular, explanations may be an effective safeguard only in settings where participants know the protected class of each point (e.g., a third-party audit or an individual contesting).

RQ4 How does the reliability of claims depend on the correctness of causal assumptions (e.g., when the strength of the proxy match their beliefs)? In particular, explanations may be a viable mechanism only in settings where participants have correct beliefs about the strength of the proxy variable .

RQ5 How does the reliability of detection change if we provide participants with multiple explanations for each prediction? If it benefits correct detection, this would speak to the importance of diverse explanations [see, e.g., 55]

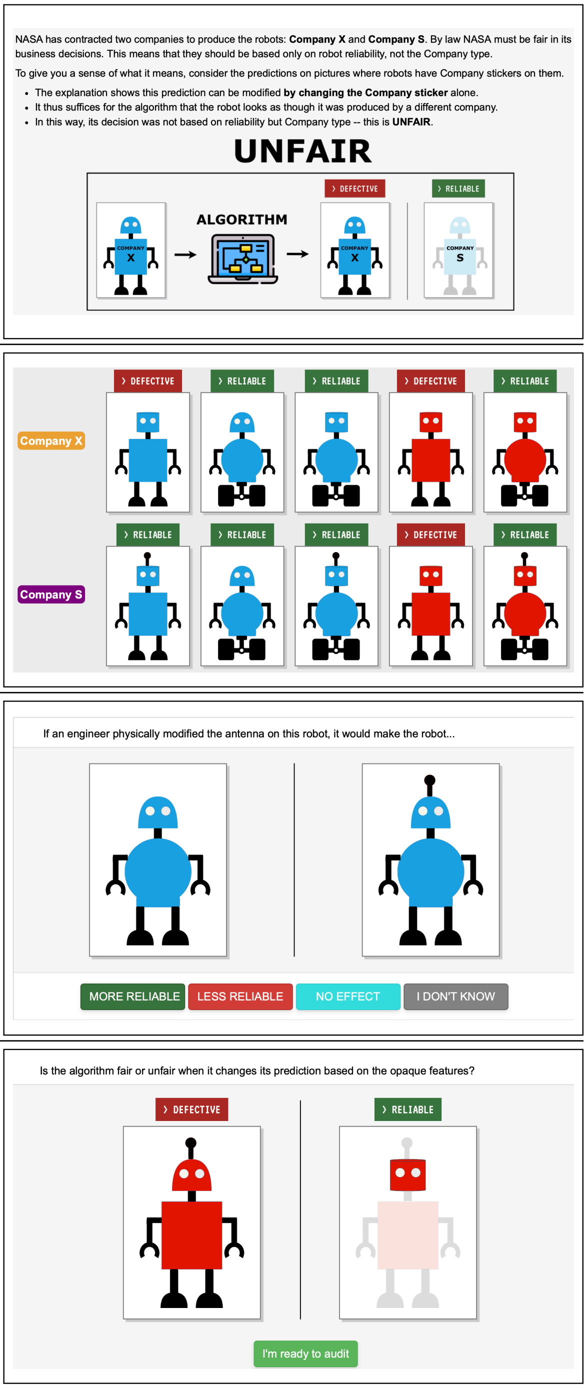
RQ6 Do participants behave in ways that are consistent and predictable? For example, will participants in each experiment make identical claims? In this case, inconsistency would highlight a need for standardization.

4.1 Setup

We used a study design with 6 conditions to see how discrimination detection would change under different assumptions on the strength of the proxy (Weak Proxy, Medium Proxy, Strong Proxy) and the number of explanations shown to participants (Single or Multiple).

1. Single: Participants were shown a single explanation for each prediction. This mimics real-world scenarios where users are given a single explanation to evaluate discrimination claims. In such tasks, a model may assign a prediction that is potentially unfair because it depends heavily on the proxy. However, an explanation may not reveal it.
2. Multiple: Participants were shown two competing explanations for each prediction, with one explanation always containing the proxy variable when it existed. This setup represents a scenario with maximum insight into the model's decision-making process. In this setup, the participants know exactly which predictions depend on the proxy and are potentially discriminatory.

Participants in each condition were shown a different set of robots to anchor their beliefs on the strength of the proxy. The sets differed by the number of robots in `Company S` with antennas: 1 robot for the Weak Proxy conditions, 3 robots for the Medium Proxy



1. Training Participants were introduced to four elements of the study: robots, their components, the prediction model, and the concept of discrimination. We used counterfactual explanations as the explanation method and presented them visually by highlighting modifiable robot parts. To explain discrimination, we used examples of robots with company stickers, establishing that predictions based on manufacturer identity were illegal. Participants completed a screening test where predictions were either discriminatory because they could be changed with company stickers or fair because they depended on robot parts. Participants then had three attempts to pass a comprehension quiz or were otherwise dropped from the study.

NASA has contracted two companies to produce the robots: **Company X** and **Company S**. By law NASA must be fair in its business decisions. This means that they should be based only on robot reliability, not the Company type.

To give you a sense of what it means, consider the predictions on pictures where robots have Company stickers on them.

- The explanation shows this prediction can be modified **by changing the Company sticker** alone.
- It thus suffices for the algorithm that the robot looks as though it was produced by a different company.
- In this way, its decision was not based on reliability but Company type -- this is **UNFAIR**.

UNFAIR

2. Anchoring We presented participants with a set of labeled robots to anchor their beliefs on the strength of the proxy and its effect on reliability. Each participant saw 5 robots from both companies, along with the ground-truth value of their reliability. The set contained 2 defective robots from `Company X` and 1 from `Company S`. All robots from `Company X` had no antenna while `Company S` had 1/3/5 robots with antennas in the `Weak Proxy`/`Medium Proxy`/`Strong Proxy` conditions, respectively. Participants were told that robots from both companies looked identical except for a specific feature (the antenna) that was more common in `Company S` robots. They were also informed that NASA suspected their algorithm might be making unfair predictions by using this feature as a proxy for company identity, since manipulating this feature would flip the prediction.

3. Elicitation We elicited participants' beliefs about the Company c_i and the effect of the proxy u_i on the reliability of each possible robot. Participants saw a total of 16 robots for (X, B), i.e., all possible robots. To elicit participants' beliefs in the protected class, we asked them to predict the robot's manufacturer or state they don't know. To elicit participants' beliefs about the ties between the proxy and robot reliability, we asked them how adding (or removing) the antenna from the robot would change reliability, allowing them to answer (more, less, no effect, or unknown). These responses allowed us to study the reliability of their discrimination claims under different assumptions on information access: knowing the protected attribute of each instance, and knowing the causal mechanism of the proxy.

If an engineer physically modified the antenna on this robot, it would make the robot...

4. Auditing We tasked participants with judging if predictions whose explanations they see were discriminatory. Participants were shown an image of a robot, its prediction (always `Defective`), and an explanation. A participant aimed to select whether a prediction was fair or unfair. This phase consisted of 16 rounds with all seven unique defective robots shown in different colors (2 robots appeared twice).

Is the algorithm fair or unfair when it changes its prediction based on the opaque features?

Figure 3: All four phases of our experiment with their description.

conditions, and all 5 robots for the Strong Proxy conditions. Our evaluation also considered different levels of knowledge in the task that participants could have. Specifically, since participants provided us with their beliefs about the predicted attribute of each robot instance, we could set them as ground-truth. Additionally, having their responses on all robots, we could estimate $\Pr(B \mid A)$ and match their belief in the causal mechanism of the proxy. We considered these additional conditions:

1. Unknown Protected Attribute: Participants have no information about the true protected attributes and estimate the distribution of the proxy based on the anchoring robot set. This is a realistic, baseline assumption where the protected attributes are not

readily available, and users have internal estimates of the true distributions.

2. Known Protected Attribute: We aligned protected attribute values with the beliefs we elicited from participants, giving them complete knowledge of these attributes. This reflects a setting where users have access to the protected attributes of each claim (e.g., when working with claims from consumers, or a third-party audit where the protected attributes are stored according to the law as in employment audits in [34]).
3. Known Causal Mechanism: Participants have perfect information about the the conditional distribution of the proxy and its causal mechanism. This is an idealized assumption and allows us to estimate best-case performance.

Counterfactual Explanations. We consider a setting where participants audit discrimination with counterfactual explanations. A *counterfactual explanation* (CE) describes how to change the inputs to a model to obtain a different prediction. Given a classifier $f : \mathcal{X} \rightarrow \{0, 1\}$ that assigns a prediction $f(\boldsymbol{x}) = 0$, a counterfactual explanation is a set of changes $e(\boldsymbol{x}, c_f)$ that satisfies $f(\boldsymbol{x} + e(\boldsymbol{x}, c_f)) = 1$. When the set is minimal, we say that $e(\boldsymbol{x}, c_f)$ is *a closest counterfactual.* Given our task, we can enumerate all possible explanations and select those that we choose to present.

Our interest in counterfactual explanations stems from three main benefits:

1. They are easy to convey to participants because we can highlight the features that must change visually.
2. We can provide participants with clear guidelines on how to use them to correctly flag unfair predictions (i.e., via a comprehension quiz).
3. Counterfactual explanations directly relate to participant claims $\hat{g}_i$, and the fact that they involve evaluating $p^{\text{flip}}_{x,b,a}$ because they list the exact changes needed to flip the prediction.

These benefits are far more difficult to achieve when, for example, we explain predictions with a feature attribution method because it is not clear how participants would use feature attribution scores to correctly flag unfair predictions [29].

Procedure. We recruited 126 participants through Prolific (20-23 per condition). All participants were fluent English speakers from the United States, comprising 74 females and 52 males, ages 19-74 (mean = 35). Each experiment lasted 32 minutes on average. We assigned each participant to 1 of the 6 conditions. Participants who saw a Single explanation were informed it may not be unique. Participants who saw Multiple explanations were informed that they reveal all ways in which a prediction can be flipped. We included a set of comprehension questions prior to the Auditing phase in the form of quizzes. In this way, we evaluated that participants understood the task, the proxy and how to establish discrimination. Participants who failed this quiz over 3 times were excluded from the study (10 excluded participants; exclusion rate of 8%).

4.2 Results

Overall, our results show that participants cannot reliably detect discrimination with explanations under any setup that could arise in practice. This result holds when users do not know the protected attributes, when they know them, and even when they know the probabilistic strength of the proxy. Representative performance measurements of audits where participants were asked to flag discriminatory predictions based on a single explanation can be found in Fig. 4 and in Fig. 5, whereas the summary measurements can be found in the Appendix in Fig. 9.

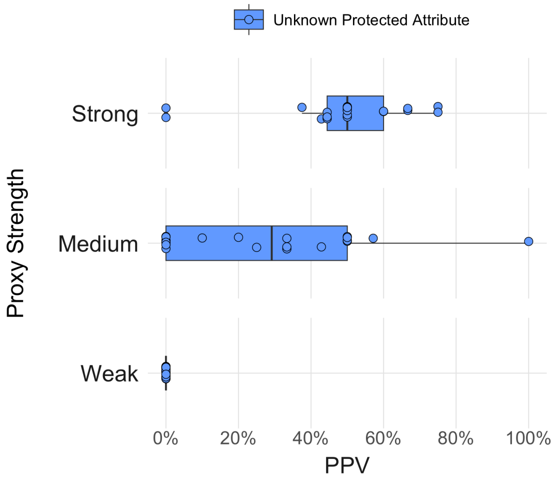


Figure 4: Distribution of PPV values for discrimination detection for participants with different beliefs on proxy strength. We show values for a threshold $\delta = 0.2$, i.e., where the prediction of a model is discriminatory if the probability it flips when changing the protected attribute is more than 20%.

On the Reliability of Discrimination Detection. We first consider a setting with threshold $\delta = 0.2$ – i.e., where we wish to flag predictions that change by over 20% after changing protected group membership. We choose this value because it is used in the U.S. employment law [34].

As seen in Fig. 4, PPV, a measure of the reliability of participant claims, indicates poor detection performance across all conditions. We would expect that participants can detect most discriminatory predicitons and avoid making false claims, which should lead to high values of PPV (e.g., ≈ 90%). In this case, we observe that in the Strong Proxy condition, where the proxy was the most prevalent and its presence in the explanation most often indicated discrimination, PPV was as low as 48% ± 4%. It was even lower, 28% ± 6% in the Medium Proxy condition to hit 0% in the Weak Proxy condition, where all predictions were fair at $\delta = 0.2$. This means that participants were correct in at most *half* of their discrimination claims. Further analysis revealed that this low reliability was affected by both missing most of the discriminatory predictions and flagging fair predictions. In the Strong Proxy condition where the results were the best, TPR reached only 44% ± 5% while maintaining substantial FPR (33% ± 5%). This means that participants incorrectly flagged 2 to 3 out of 10 fair predictions, and missed at least 3 out of 5 of all discriminatory predictions.

These findings challenge the idea of using explanations to detect discrimination in practice. Without additional assumptions or safeguards, humans both fail to detect most of the discriminatory

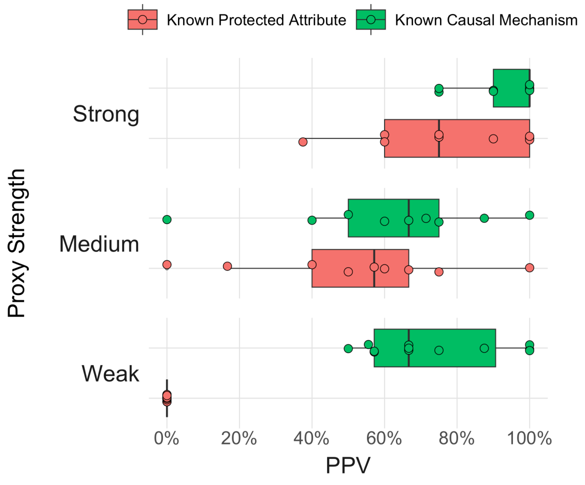


Figure 5: Distribution of PPV values of discrimination detection for participants with different beliefs about proxy strength. We show the PPV of discrimination claims when we assume that participants know the protected attributes of each robot and when they know the causal mechanism of the proxy, i.e., Pr$(B \mid A)$.

cases, and raise multiple false alarms. This combination risks letting discriminatory practices continue and triggering unnecessary investigations that waste resources and harm legitimate practices.

We report our findings across all thresholds in the Appendix in Fig. 9. As shown, unreliable detection is observed across almost all thresholds, increasing only at extremely low values. For these low thresholds ($\delta \leq 5\%$), nearly all predictions that depend on the value of the proxy are discriminatory, reflecting a conservative regime where minor flip rate changes indicate discrimination. Since participants tend to flag these predictions, they achieve high PPV ($\approx 75\%$). Still, as shown in Fig. 9, they maintain poor TPR and FPR of $\approx 30\%$. These values of TPR and FPR are also un upper bound for other δs.

On the Value of Information About Protected Attributes. A natural question is whether the poor detection performance stems from a lack of knowledge of protected attributes. To answer this question, we matched participants' attribute selections from the Elicitation phase with the corresponding predictions.

Our results, which can be seen in Fig. 5, show only marginal improvements: at $\delta = 0.2$, PPV increased to 39% ± 6% (Weak Proxy condition) and 37% ± 3% (Medium Proxy condition) from the baseline of 28%, with neither change reaching significance under Mann-Whitney U test ($p > 0.1$, $U \geq 156.5$). Only the Strong Proxy condition showed significant improvement, with PPV rising to 66% ± 7% from 48% ± 7% ($p < 0.05$, $U = 114.5$). We found similarly slight improvements for other measures: FPR dropped by approximately 10% (equivalent to $\approx$ 1 prediction), and TPR decreased by 6-7%, both across all conditions. These results suggest that participants sometimes chose not to flag discrimination even when their own beliefs about protected attributes would warrant it. This inconsistent behavior often occurred when participants believed changing the proxy has legitimate influence on reliability – e.g., on average, if participant believed the change in the CE affects robot reliability, they claimed the prediction is fair in 64% of the cases whereas if they thought the proxy has no effect – in 50% of the cases.

In total, we observe that whether participants know the protected attributes of the instances they inspect or not plays a marginal role in detection performance. Even with access to these attributes, participants still missed many discriminatory cases and raised multiple false alarms. As shown in Fig. 9 in the Appendix, this performance persisted across all δ values, except for very low thresholds where most proxy-dependent predictions were discriminatory. In these cases, participants correctly focused on such predictions, which lead to higher PPV (most claims were accurate), though their overall detection remained poor (participants' claims had low TPR and high FPR).

On the Sensitivity to Causal Assumptions. Our experiment also allows us to evaluate how performance would improve under best-case assumptions where humans have perfect information on the causal mechanism of the proxy. In this case, we assume Pr$(B \mid A)$ matches their beliefs. We found that this intervention significantly improved PPV at $\delta = 0.2$ across all conditions, as seen in green in Fig. 5. In the Strong Proxy condition, PPV went from 48% ± 4% to 77% ± 7% ($p < 0.001$, $U = 66.5$). In the Medium Proxy condition it went from 28% ± 6% to 49% ± 8% ($p \leq 0.05$, $U = 128.5$). In the Weak Proxy condition, PPV increased significantly above 0 to 61% ± 8%. This is because participants perceived a stronger proxy relationship than existed (over half of the participants assumed Pr$(B = 0 \mid A = 0) = 0$), and their discrimination claims were often warranted under these beliefs. As shown in Fig. 9, this effect persisted over other threshold values, consistently decreasing the higher the threshold, because fewer and fewer predictions were discriminatory and this did not match participants' claims. Still, neither PPV, TPR nor FPR ever reached a value we would consider satisfactory. TPR never exceeded 40% whereas FPR was consistently above 25%. Overall, these results point out that poor performance cannot readily be remedied by domain expertise.

On the Effect of Multiple Explanations. We next examined participants' performance when they were given full information about the prediction by being shown Multiple explanations. In this setup, participants knew with certainty whether the prediction could be flipped with the proxy or not. Such guarantees are rarely available in reality, but we make this assumption to test if explanations *could* work in idealized circumstances.

In short, this manipulation did not improve performance, as we show in Fig. 10. On average, PPV was bounded by 40% across all conditions. TPR behaved irregularly but never exceeded 40% as well. FPR remained consistently at least 30%. The only exception occurred in the Weak Proxy condition with extreme values of $\delta \leq 0.05$ with PPV reaching 77% ± 7% and TPR 63% ± 9% ($p < 0.01$, $U \geq 220$). However, this came at the cost of increased false positives (FPR as high as 55% ± 8% at $\delta = 0.2$). These results hold irrespective of the level of knowledge participants have about the predicted points and the environment. Overall, people seem incapable of using explanations reliably even under idealized conditions.

On Model Audits. Participants were unable to differentiate between cases when the model was fair or discriminatory. In a task

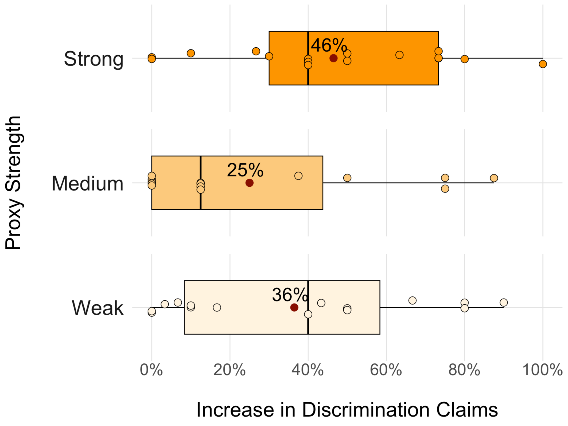


Figure 6: Increase in discrimination claims when explanations contained the proxy versus when they did not. We highlight mean increases, showing that participants consistently identified the proxy as a discrimination signal across all regimes.

where we would say that a model discriminated if over 20% of predictions were discriminatory, our model should be fair in the Weak Proxy condition and discriminatory in the Medium Proxy and Strong Proxy conditions. Nonetheless, participants were at most marginally affected by the proxy strength, and labeled the model discriminatory across all conditions (13/21, 10/20, and 16/21 participants across Weak Proxy, Medium Proxy, and Strong Proxy conditions, respectively). These proportions remained similar even when participants saw a comprehensive set of Multiple explanations (13/17 participants for the Weak Proxy condition, 13/19 participants for the Medium Proxy condition, 12/19 participants for the Strong Proxy condition claimed the model was discriminatory). This suggests people generally equate the presence of a proxy with discrimination, regardless of its strength. If we relied on explanations to judge models globally, this would unnecessarily block the deployment of multiple fair models.

On the Consistency of Auditors and Decision Subjects. Our results show that participants were consistent in their claims, primarily using the presence of the proxy variable as an indicator of discrimination. As shown in Fig. 6, participants claimed discrimination more frequently when shown explanations that highlight the proxy. Specifically, explanations that included the proxy increased discrimination claims by 25% in the Medium Proxy condition up to 46% in the Strong Proxy condition compared to explanations without the proxy. This increase was more pronounced when participants viewed Multiple explanations (i.e., 36-60% increase across conditions). The greater frequency of explanations that included the proxy in these conditions (14 predictions with such explanations in the Multiple explanation conditions versus 8 predictions in the Single explanation conditions) also contributed to a substantial increase in discrimination claims overall – from 30% in the Medium Proxy condition to 47% in the Strong Proxy condition.

Although on average, participants were responsive to the presence of the proxy variable, they often disagreed on which cases were discriminatory. For instance, in the Medium Proxy condition, predictions whose explanations contained the proxy were judged as discriminatory an average of 51% of the time. We believe this lack of agreement between participants is influenced by two main systematic factors. First, participants held different beliefs about robot reliability, and these beliefs affected their judgments. When participants believed the proxy indicated higher reliability, they were 20% more likely to label predictions whose explanations included the proxy as fair. While this pattern shows high variability ($p \approx 0.3$), we consistently observe it across proxy strength conditions and it aligns with participants' explicit statements (e.g., "*It is not unfair to say that robots with antennas work better*"). Second, participants held false beliefs about the causal *structure* of the problem as described in Remark 5. We observed a steady, low FPR of ≈ 30% even under perfect assumptions about participant knowledge. This effect can only be attributed to labeling predictions that do not depend on the proxy as discriminatory, falsely believing other features are proxies. This is because we observe roughly the same FPR for $\delta \approx 1$, meaning participants labeled predictions where $f(b, x) = f(b', x)$ as discriminatory. This sentiment can be found in participants' answers (e.g., saying "*I decided based on the body shape and the base type*"). In reality, we found that 36 out of 61 participants fell prey to these assumptions, including 8 participants who labeled predictions where the proxy was not present as discrimination.

4.3 Discussion

By using a controlled environment with clearly defined ground-truth discrimination label, we were able to precisely measure how explanations fail to support discrimination detection. This approach provided participants with optimal conditions. They received clear instructions, could identify the proxy and its relationship to the protected attribute, and received explicit explanations showing counterfactual outcomes. The fact that explanations failed as a safeguard against discrimination under these favorable conditions, or even when adapting the ground-truth to participant beliefs, suggests fundamental limitations of using explanations for this task. We discuss our results in more detail below.

Fundamental Detection Failure. Auditing with either a single explanation or a comprehensive set of multiple explanations does not allow humans to reliably detect discrimination. Neither does knowing the protected attribute of the audited predictions, nor correctly identifying the causal mechanism of the proxy. Participants detected more than 65% of the truly discriminatory cases (TPR), and had *at most* 77% correct detections (PPV), but only when their beliefs were treated as correct. Otherwise, the reliability of detection was approximately 50% with false alarms consistently around 30% (FPR). In practical terms, these findings indicate that every 1 in 4 individuals who file a discrimination claim fail in court due to insufficient evidence. These findings also indicate that nearly half of all genuinely discriminatory decisions would go completely undetected by affected individuals. This creates a problem of both wasted legal resources and unaddressed algorithmic harm.

Lack of Auditor Agreement. One could try looking at the auditing performance with respect to model discrimination as more of a success. After all, the model which was discriminatory for most thresholds (when the proxy was medium and strong) would be determined as such by an average auditor. However, when it comes to individual performance, the results look much worse. First, more than half of all the participants claimed the model with the weak proxy was discriminatory when it was not (26/38 participants). Second, barely over a half of the participants detected the model is discriminatory when it used a medium proxy (23/39 participants) and three-quarters of the participants when the model used a strong proxy (28/40 participants). We observed a lack of overall agreement between participants who essentially operated on their own beliefs about discrimination. This led to claims that were very rarely matching (Cohen's κ ranging from 0.05 to 0.14 across all conditions). This is also seen when we analyze predictions individually and find that every prediction was selected as discriminatory by at least 10% of the participants. Put together, if the same set of predictions were analyzed by two independent auditors, it could lead to two different results. Discriminatory models might escape detection while fair ones could face false accusations.

The fundamental reason why explanations are ineffective for discrimination detection is that they operate on individuals, whereas fairness must be evaluated over groups of (hypothetical) individuals. This tension is well-documented in formal definitions of fairness [59], and our experiments demonstrate how it impairs human performance. Our analysis revealed three specific challenges were the direct causes of people's failure:

Flawed Causal Assumptions. More than half of all participants (71 out of 118) fell prey to the beliefs that some features combined with the proxy are evidence of discrimination. 17 of the participants also thought that some combinations of features without the proxy can indicate discrimination. This led participants to incorrectly raise false alarms. This also led participants to not detect discrimination because they looked for "stronger proof" (e.g., one participant noted they looked for a combination of antenna and other features to claim discrimination).

Flawed Judgment of Proxy Strength. Over half of the participants overestimated proxy strength. This is best seen by the largely improved performance (PPV and TPR) under their own beliefs in the causal mechanism when the thresholds are low. This led to many false positives in claiming discrimination. We can expect people to misrepresent the proxy strength in reality, too, because it is rarely observable. This misrepresentation might lead to a claim that the whole model is discriminating, while it is perfectly valid (like in the Weak Proxy conditions).

Real Outcome Interference. Participants' judgments were sometimes influenced by their beliefs about the relationship between features and desirable outcomes. This led to errors. We observed this behavior across all conditions. For instance, in the Weak Proxy condition with Multiple explanations, participants claimed predictions as fair in 52% of the cases when they thought adding a proxy makes the robot reliable, and otherwise, only in 28%. Even though the median increase was about 20%, as many as 78 out of all 118 participants made a claim like this at least once. We could also see this sentiment in participants' written responses.

Limitations. Our results are limited by two main factors that were beyond our control. First, our participants had no prior training in statistics or probability. This might have affected their judgments, making them inconsistent with respect to, e.g., proxy strength and the causal mechanism. This is especially important since fairness audits depend on probabilistic claims. Second, every study run on paid-survey platforms such as Prolific has to deal with inattentiveness or lack of motivation. Despite our best efforts, the task we introduced was abstract and gave no immediate feedback. This could have made participants guess oftentimes and act inconsistently. They might have also had less incentive to perform thoughtfully, contrary to real auditors who may be bound by law.

5 Concluding Remarks

Our study demonstrates the fundamental limitations of using explanations as a safeguard against algorithmic discrimination. In our controlled experiment with human participants ($N = 126$), we found that explanations fail to reliably assist in discrimination detection, regardless of how much information they convey or if auditors know the protected attributes or proxy strength.

Our findings extend to real-world auditing scenarios. This is because these scenarios present far greater complexity, with more features, intricate relationships, and numerous plausible explanations to consider [20]. The failure modes that compromise human performance in our simple setup – flawed causal reasoning, incorrect estimate of the proxy strength, and real outcome interference – are likely to persist or worsen with increased complexity. Furthermore, these individual-level failures may compound in real-world settings where multiple stakeholders must coordinate their assessments, just as they did in our experiment. In total, this will lead to poor discrimination detection performance in applied settings.

This result is strongly related to a growing body of regulations on algorithmic discrimination and transparency. In recent years, jurisdictions worldwide have adopted two main approaches. The first approach emphasizes transparency and explanation rights – see e.g., ECOA's mandate for adverse action notices in lending [58] or provisions for a "Right to an Explanation" in data regulation laws in the European Union [68], Brazil [13], and South Korea [36]. Mandatory fairness audits represent the second regulatory approach, see e.g., pre-implementation mandates in Slovenia [53], bias audits for employment decisions in New York [34], or risk assessments of "very large online platforms," in the EU [69]. Despite this momentum, there remains a lack of standardized practices for assessing algorithmic fairness as regulations provide limited guidance for how to conduct audits [41]. Our results highlight two critical insights for policy. First, there is a need for standalone regulations specifically targeting algorithmic discrimination. Current policy relying on explanations is unreliable even under controlled conditions (see also [32] for a legal discussion). Second, while the "right to explanation" serves a valuable role in accessing other rights (as exemplified in EU regulations), it should not be considered sufficient for preventing discrimination. Rather, it must be deployed alongside robust anti-discrimination measures and systematic auditing procedures that do not solely rely on human interpretation of explanations.

Acknowledgments

This work was supported by funding from the National Science Foundation under awards IIS 2040880 and IIS 2313105, and the NIH Bridge2AI Center Grant U54HG012510.

References

[1] 116th Congress. 2019. Algorithmic Accountability Act of 2019. https://www.congress.gov/bill/117th-congress/house-bill/6580/text

[2] 117th Congress. 2022. Algorithmic Accountability Act of 2022. https://www.congress.gov/bill/117th-congress/house-bill/6580/text

[3] Abubakar Abid, Mert Yuksekgonul, and James Zou. 2022. Meaningfully debugging model mistakes using conceptual counterfactual explanations. In *International Conference on Machine Learning*. PMLR, 66–88.

[4] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54 (2018), 95–122.

[5] Ifeoma Ajunwa, Sorelle Friedler, Carlos E Scheidegger, and Suresh Venkatasubramanian. 2016. Hiring by algorithm: predicting and preventing disparate impact. *Available at SSRN* (2016).

[6] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion* 99 (2023), 101805.

[7] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 556–579.

[8] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[9] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1194–1206.

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[11] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *California law review* (2016), 671–732.

[12] 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*. 1–14.

[13] Brazil. 2020. Brazilian General Data Protection Law. https://iapp.org/media/pdf/resource_center/Brazilian_General_Data_Protection_Law.pdf

[14] Marc-Etienne Brunet, Ashton Anderson, and Richard Zemel. 2022. Implications of Model Indeterminacy for Explanations of Automated Decisions. *Advances in Neural Information Processing Systems* 35 (2022), 7810–7823.

[15] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[16] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[17] Zana Buçinca, Siddharth Swaroop, Amanda E Paluch, Finale Doshi-Velez, and Krzysztof Z Gajos. 2024. Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills. *arXiv preprint arXiv:2410.04253* (2024).

[18] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Effects of ai and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–42.

[19] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting high-uncertainty decisions through AI and logic-style explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 251–263.

[20] Sasha Costanza-Chock, Inioluwa Deborah Raji, and Joy Buolamwini. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1571–1583.

[21] Jessica Dai, Paula Gradu, Inioluwa Deborah Raji, and Benjamin Recht. 2025. From Individual Experience to Collective Evidence: A Reporting-Based Framework for Identifying Systemic Harms. *arXiv preprint arXiv:2502.08166* (2025).

[22] Jessica Dai, Sohini Upadhyay, Ulrich Aivodji, Stephen H Bach, and Himabindu Lakkaraju. 2022. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 203–214.

[23] Xinyue Dai, Mark T Keane, Laurence Shalloo, Elodie Ruelle, and Ruth MJ Byrne. 2022. Counterfactual explanations for prediction and diagnosis in XAI. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 215–226.

[24] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[25] Philip Dawid. 2017. On individual risk. *Synthese* 194, 9 (2017), 3445–3474.

[26] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[27] Mengnan Du, Fan Yang, Na Zou, and Xia Hu. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems* 36, 4 (2020), 25–34.

[28] Equal Employment Opportunity Commission. 1978. Uniform Guidelines on Employee Selection Procedures). Electronic Code of Federal Regulations. https://www.ecfr.gov/current/title-29/subtitle-B/chapter-XIV/part-1607/subject-group-ECFRdb347e844acdea6 29 CFR Part 1607.

[29] Carlos Fernández-Loría, Foster Provost, and Xintian Han. 2022. Explaining Data-driven Decisions Made by AI Systems: The Counterfctual Approach. *MIS Quarterly* 46, 3 (2022), 1635–1660.

[30] Ana Cristina Bicharra Garcia, Marcio Gomes Pinto Garcia, and Roberto Rigobon. 2024. Algorithmic discrimination in the credit domain: what do we know about it? *AI & SOCIETY* 39, 4 (2024), 2059–2098.

[31] Talia B Gillis, Vitaly Meursault, and Berk Ustun. 2024. Operationalizing the Search for Less Discriminatory Alternatives in Fair Lending. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 377–387.

[32] Talia B Gillis and Josh Simons. 2019. Explanation< Justification: GDPR and the Perils of Privacy. *JL & Innovation* 2 (2019), 71.

[33] Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III. 2024. The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 155–180.

[34] Lara Groves, Jacob Metcalf, Alayna Kennedy, Briana Vecchione, and Andrew Strait. 2024. Auditing work: Exploring the New York City algorithmic bias audit regime. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1107–1120.

[35] Kofi Immanuel Jones and Swati Sah. 2023. The Implementation of Machine Learning In The Insurance Industry With Big Data Analytics. *International Journal of Data Informatics and Intelligent Computing* 2, 2 (2023), 21–38.

[36] Dong Hyeon Kim and Do Hyun Park. 2024. Automated decision-making in South Korea: a critical review of the revised Personal Information Protection Act. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–11.

[37] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. *arXiv:2202.01602 [cs]* (Feb. 2022). http://arxiv.org/abs/2202.01602 arXiv: 2202.01602.

[38] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).

[39] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[40] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*. 29–38.

[41] Khoa Lam, Benjamin Lange, Borhane Blili-Hamelin, Jovana Davidovic, Shea Brown, and Ali Hasan. 2024. A framework for assurance audits of algorithmic systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1078–1092.

[42] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[43] Thao Le, Tim Miller, Ronal Singh, and Liz Sonenberg. 2022. Improving model understanding and trust with counterfactual explanations of model confidence. *arXiv preprint arXiv:2206.02790* (2022).

[44] Derek Leben. 2023. Explainable AI as evidence of fair decisions. *Frontiers in Psychology* 14 (2023), 1069426.

[45] Min Hun Lee and Chong Jun Chew. 2023. Understanding the effect of counterfactual explanations on trust and reliance on ai for human-ai collaborative

clinical decision making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–22.

[46] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[47] Michelle Seng Ah Lee and Luciano Floridi. 2021. Algorithmic fairness in mortgage lending: from absolute conditions to relational trade-offs. *Minds and Machines* 31, 1 (2021), 165–191.

[48] Geng Li. 2018. Gender-Related Differences in Credit Use and Credit Scores. *FEDS Notes* (22 June 2018). doi:10.17016/2380-7172.2188

[49] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301.

[50] Ana Lucic, Hinda Haned, and Maarten de Rijke. 2020. Why does my model fail? contrastive local explanations for retail forecasting. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 90–98.

[51] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[52] Marco Lünich and Birte Keller. 2024. Explainable Artificial Intelligence for Academic Performance Prediction. An Experimental Study on the Impact of Accuracy and Simplicity of Decision Trees on Causability and Fairness Perceptions. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1031–1042.

[53] Gianclaudio Malgieri. 2019. Automated decision-making in the EU Member States: The right to explanation and other "suitable safeguards" in the national legislations. *Computer law & security review* 35, 5 (2019), 105327.

[54] Vishwali Mhasawade, Salman Rahman, Zoé Haskell-Craig, and Rumi Chunara. 2024. Understanding disparities in post hoc machine learning explanation. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2374–2388.

[55] Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 607–617. doi:10.1145/3351095.3372850

[56] Chelsea M Myers, Evan Freed, Luis Fernando Laris Pardo, Anushay Furqan, Sebastian Risi, and Jichen Zhu. 2020. Revealing neural network bias to non-experts through interactive counterfactual examples. *arXiv preprint arXiv:2001.02271* (2020).

[57] Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. 2022. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*. PMLR, 16848–16887.

[58] Bureau of Consumer Financial Protection. 2020. Equal Credit Opportunity (Regulation B); Revocations or Unfavorable Changes to the Terms of Existing Credit Arrangements. https://files.consumerfinance.gov/f/documents/cfpb_revoking-terms-of-existing-credit-arrangement_advisory-opinion_2022-05.pdf

[59] Drago Plečko, Elias Bareinboim, et al. 2024. Causal fairness analysis: a causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning* 17, 3 (2024), 304–589.

[60] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[61] Max Schemmer, Joshua Holstein, Niklas Bauer, Niklas Kühl, and Gerhard Satzger. 2023. Towards meaningful anomaly detection: The effect of counterfactual explanations on the investigation of anomalies in multivariate time series. *arXiv preprint arXiv:2302.03302* (2023).

[62] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1616–1628.

[63] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24, 1 (2022), 2.

[64] Vinith Menon Suriyakumar, Marzyeh Ghassemi, and Berk Ustun. 2023. When personalization harms performance: reconsidering the use of group attributes in prediction. In *International Conference on Machine Learning*. PMLR, 33209–33228.

[65] Winnie F Taylor. 1980. Meeting the Equal Credit Opportunity Act's Specificity Requirement: Judgmental and Statistical Scoring Systems. *Buff. L. Rev.* 29 (1980), 73.

[66] Taylor Telford. 2019. Apple Card algorithm sparks gender bias allegations against Goldman Sachs. *Washington Post* 11 (2019).

[67] Michael Carl Tschantz. 2022. What is proxy discrimination?. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1993–2003.

[68] European Union. 2018. General Data Protection Regulation, Art. 22. https://gdpr-info.eu/art-22-gdpr/

[69] European Union. 2024. The Digital Services Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32022R2065/

[70] Berk Ustun, Yang Liu, and David Parkes. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*. PMLR, 6373–6382.

[71] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial intelligence* 291 (2021), 103404.

[72] Michael Veale and Irina Brass. 2019. Administration by algorithm? Public management meets public sector machine learning. *Public management meets public sector machine learning* (2019).

[73] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.

[74] J Christina Wang and Charles B Perkins. 2019. How magic a bullet is machine learning for credit analysis? An exploration with FinTech lending data. *An Exploration with FinTech Lending Data (October 21, 2019)* (2019).

[75] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in AI-assisted decision-making. In *26th international conference on intelligent user interfaces*. 318–328.

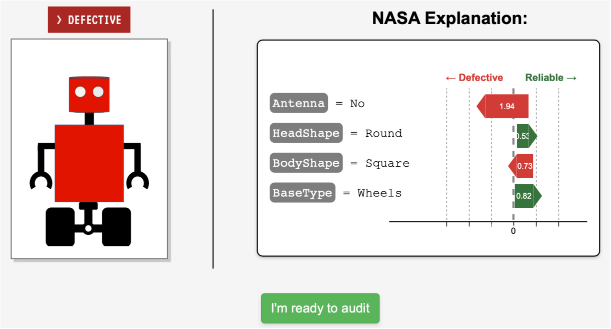
[76] Richard Warner and Robert H Sloan. 2021. Making artificial intelligence transparent: Fairness and the problem of proxy variables. *Criminal Justice Ethics* 40, 1 (2021), 23–39.

[77] Greta Warren, Ruth MJ Byrne, and Mark T Keane. 2023. Categorical and continuous features in counterfactual explanations of AI systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 171–187.

[78] Monika Westphal, Michael Vössing, Gerhard Satzger, Galit B Yom-Tov, and Anat Rafaeli. 2023. Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance. *Computers in Human Behavior* 144 (2023), 107714.

[79] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling fairness perceptions in algorithmic decision-making: the effects of explanations, human oversight, and contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

[80] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.



Antenna = No

HeadShape = Round

BodyShape = Square

BaseType = Wheels

Figure 7: Example of a SHAP explanation in our study.

A Table of Notation

B Supplementary Material on Experimental Design

In this Section, we provide supplementary materials on our experimental design. This includes the exact list of robots (points the model predicted on) with their closest counterfactual explanations in Table 3, and links to our GitHub repository with the code for the experiment and the experimental data.

B.1 Availability of data and material (data transparency)

Anonymized data from the experiments is available at https://github.com/juliannski/discrimination-detection/tree/main/results.

B.2 Code availability (software application or custom code)

The code for our Flask study is available at https://github.com/juliannski/discrimination-detection.

1. Run `pip3 install -r requirements.txt` to install the necessary requirements.
2. Then run `application.py` and open the link to the `localhost` to start the study.
3. Parameters listed at the top of the file can be used to run the study in different conditions.

C Supplementary Experimental Results

In this Section, we present the results of running our study with feature-attribution SHAP explanations [51]. We also provide additional figures for our experimental results from the main text.

C.1 Experiment with SHAP Explanations

We repeated our experiment with SHAP explanations and obtained results aligned with the results on counterfactual explanations. We recruited 23 participants in the Strong Proxy condition (13 female, English speaking, average age 40, 0% rejection rate, average completion time 40 minutes). The explanations were derived from the coefficients of the linear classifier we used in the paper. We added a small noise to each SHAP value to make them unique across the experimental trials. During the Training phase, participants saw 1 example where the model used the Company sticker only (unfair), 1 example where the Company sticker had a SHAP value of 0 (fair) and 2 examples where the Company sticker had a non-zero value. In these cases, participants were informed that the discrimination status was uncertain and were told to evaluate whether the proxy’s influence (as measured by its SHAP value) was substantial enough to make the prediction dependent on it. Participants could select either ’fair’ or ’unfair’ for these borderline scenarios. During the quiz, participants needed to order the robot parts based on their influence on the prediction shown in a sample SHAP explanation (see Fig. 7) to make sure they understand the relative influence on predictions that SHAP values communicate. As seen in Fig. 11, all of our metrics were roughly the same across all fairness thresholds δ with TPR and FPR of approximately 40% and PPV of 65%. This means that participant’s choices were almost like a coin flip. This should not be surprising since there is no reliable method of determining fairness using feature attribution explanations.

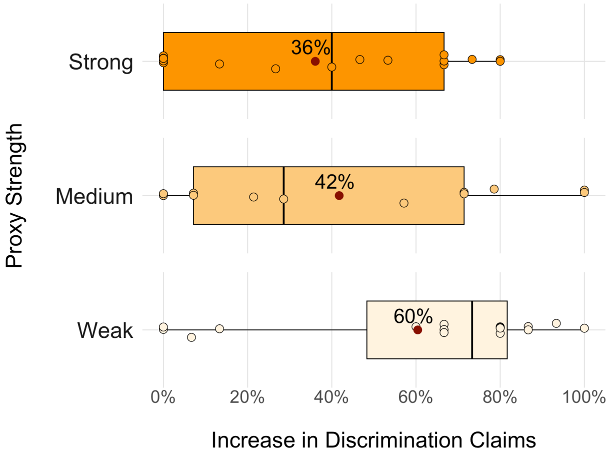


Figure 8: Increase in discrimination claims when explanations contained the proxy versus when they did not when participants saw Multiple explanations. We highlight mean increases, showing that participants consistently identified the proxy as a discrimination signal across all regimes.

C.2 Experiments in the Main Text

Fig. 9 presents the sensitivity analysis of all metrics we report in the main text. In addition to PPV, we also report TPR and FPR (the proportion of identified discriminatory predictions and the proportion of false claims, respectively).

Fig. 10 shows performance measures (PPV, TPR and FPR) across all thresholds $\delta \in [0, 1]$ in the conditions that used Multiple explanations. We detail the results of these studies in Section 4.2. Fig. 8 shows that participants’ claims depended on the presence of the proxy in the explanation also for Multiple explanations conditions. Finally, Fig. 12 shows the lack of agreement between the participants we discussed in Section 4.2, detailing how often each of the predictions used in the study was claimed to be discriminatory.

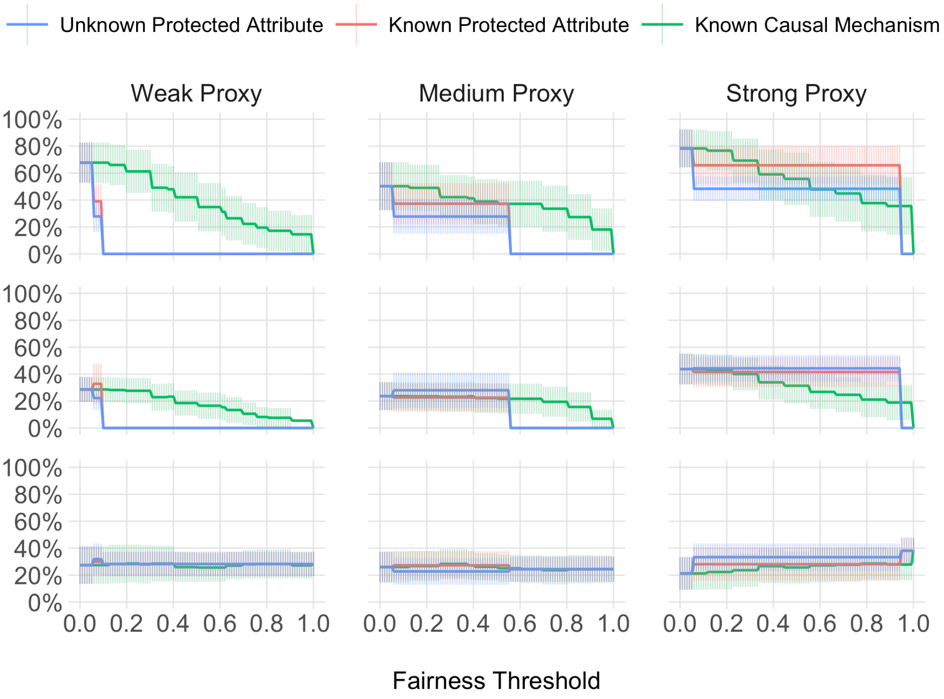


Figure 9: Reliability of discrimination claims across all possible fairness thresholds $\delta \in [0, 1]$ when participants saw s Single explanation. We show the confidence intervals for PPV(δ), TPR(δ), FPR(δ) for all proxy strength conditions and under different assumptions on participant knowledge.

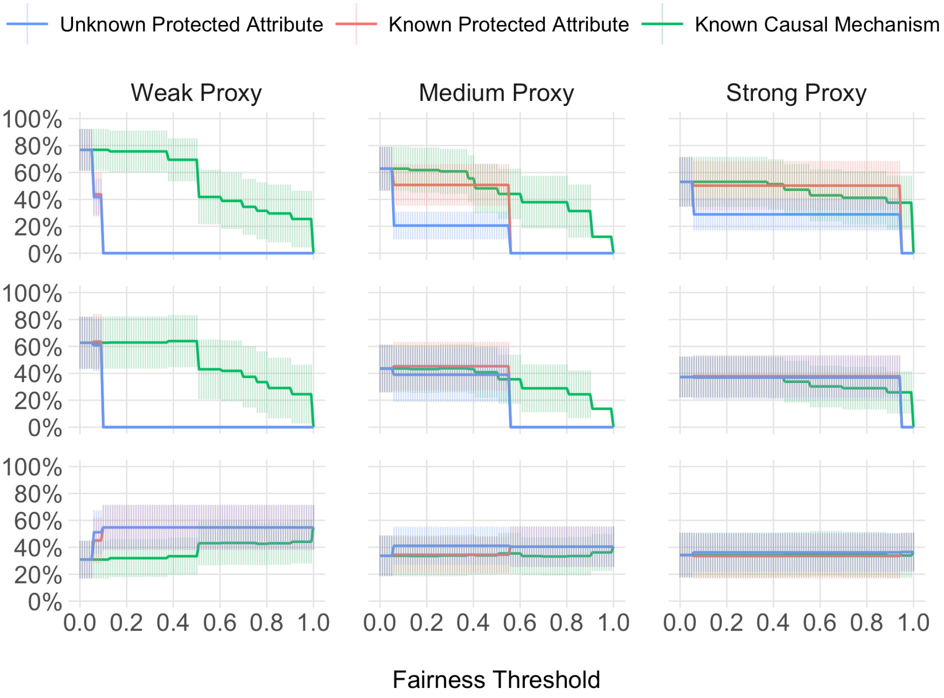


Figure 10: Reliability of discrimination claims across all possible fairness thresholds $\delta \in [0, 1]$ when participants saw Multiple explanations. We show the confidence intervals for PPV(δ), TPR(δ), FPR(δ) for all proxy strength conditions and under different assumptions on participant knowledge.

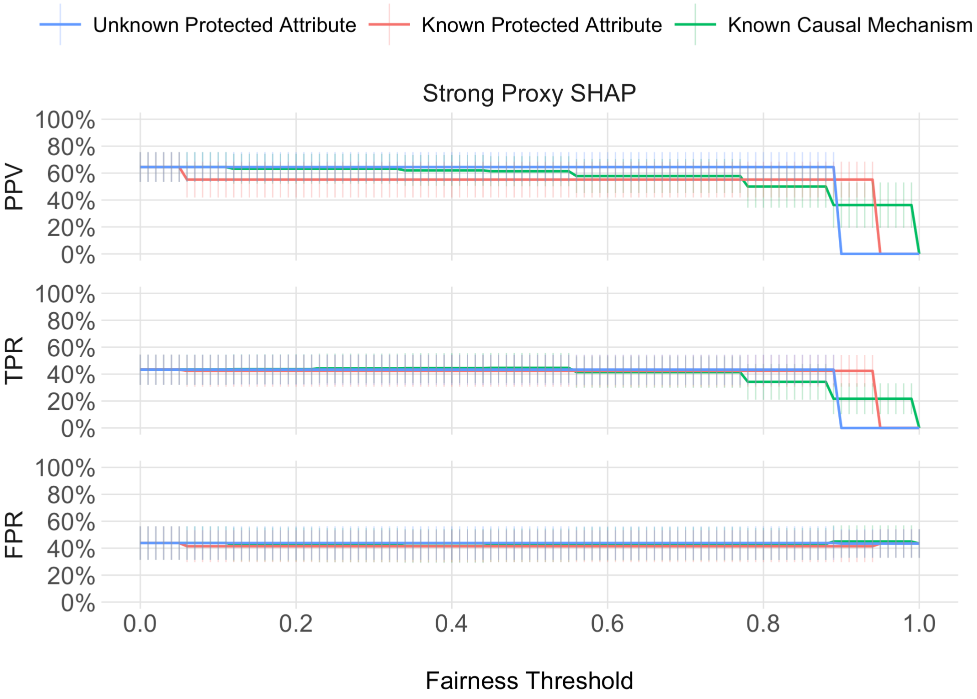
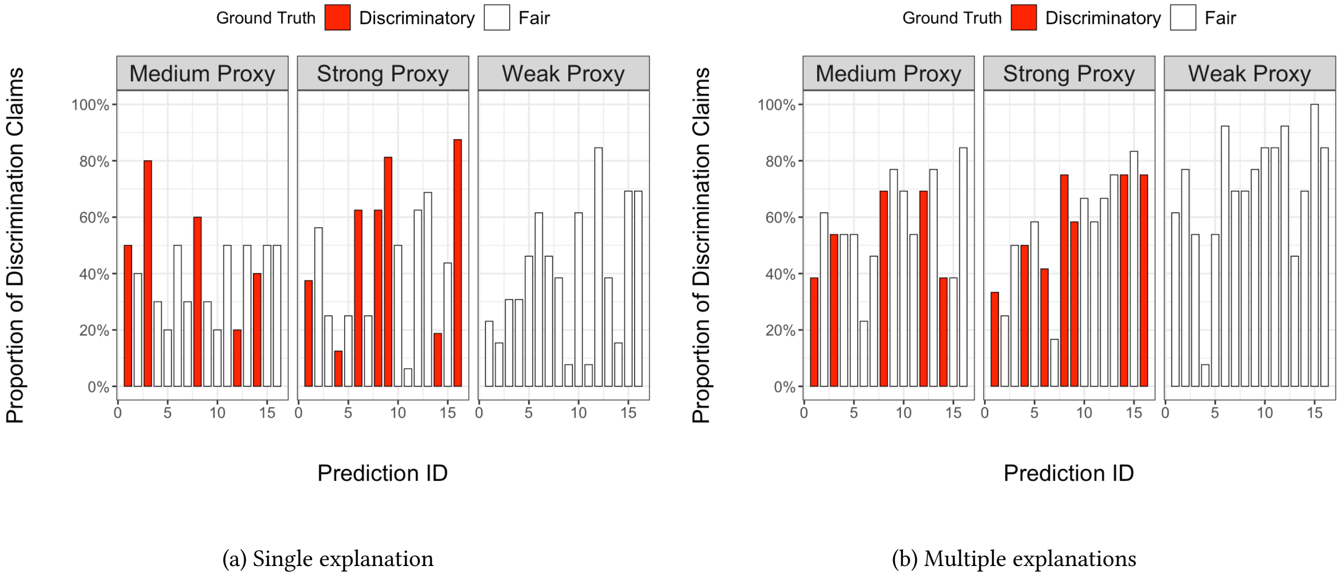


Figure 11: Performance metrics across all fairness threshold δ values when participants were assisted by SHAP explanations. Refer to Fig. 9 for the explanation of the plotted data. As seen, the detection is poor across all fairness thresholds with consistently high FPR, and consistently low TPR, both around 40%. This results in low reliability of claims as measured by PPV.



(a) Single explanation

(b) Multiple explanations

Figure 12: Discrimination claims per individual predictions in each of the proxy regimes when participants saw a single explanation (left) and multiple explanations (right). We can see that every prediction was judged as discriminatory by at least 10% of the participants. Participants were also not in full agreement with any of the predictions. On average, the agreement was roughly 50%.

Notation	Description
A	Protected attribute (e.g., company identity)
B	Proxy variable for the protected attribute (e.g., antenna)
X	Features independent of protected attribute (e.g., other robot parts)
Y	True outcome variable (e.g., reliability)
$\hat{Y}$	Predicted outcome from model h
$h(x, b)$	Model that predicts $\hat{Y}$ given inputs $X = x$ and $B = b$
$\phi_{x,b,a}$	Level of discrimination/probability prediction flips when intervening on A
δ	Fairness threshold representing maximum allowed discrimination
δ_{min}	Minimum fairness threshold for evaluation
$\delta^{internal}$	User's internal fairness threshold for making discrimination claims
$g_{i\mid h,\delta}$	Ground truth label indicating discrimination in prediction i
$\hat{g}_{i\mid h,e_i}$	User's claim about discrimination for prediction i given explanation $\mathcal{E}_i$
G_i	Random variable that determines if prediction i flips when intervening on A, following Bernoulli($\phi_{x,b,a}$)
$\mathcal{E}_i$	Explanation provided for prediction i
TPR(δ_{min})	True positive rate for discrimination detection at threshold δ_{min}
FPR(δ_{min})	False positive rate for discrimination detection at threshold δ_{min}
PPV(δ_{min})	Positive predictive value for discrimination claims at threshold δ_{min}

Table 2: Notation used in the paper.

Features				Prevalence		Counterfactual Explanations
Antenna	HeadShape	BodyShape	BaseType	Company X	Company S	
No	Square	Square	Legs	0.0071	0.0004	{Antenna, HeadShape}, {Antenna, BaseType}, {Antenna, HeadShape}, {BodyShape, BaseType}
No	Square	Square	Wheels	0.016	0.0008	{Antenna}
No	Square	Round	Legs	0.016	0.0008	{Antenna}, {BodyShape}
No	Square	Round	Wheels	0.0297	0.0016	{Antenna}, {BodyShape}
No	Round	Square	Legs	0.016	0.0008	{Antenna}, {BaseType}
No	Round	Square	Wheels	0.0297	0.0016	{Antenna}, {BaseType}
No	Round	Round	Legs	0.0297	0.0016	{BodyShape}, {BaseType}
No	Round	Round	Wheels	0.0434	0.0023	{BodyShape}, {BaseType}
Yes	Square	Square	Legs	0.0008	0.016	{HeadShape}, {BodyShape}, {BaseType}
Yes	Square	Square	Wheels	0.016	0.0297	{Antenna}, {HeadShape}
Yes	Square	Round	Legs	0.016	0.0297	{Antenna}, {BaseType}
Yes	Square	Round	Wheels	0.0023	0.0434	{Antenna}
Yes	Round	Square	Legs	0.016	0.0297	{Antenna}, {BodyShape}
Yes	Round	Square	Wheels	0.0023	0.0434	{Antenna}
Yes	Round	Round	Legs	0.0023	0.0434	{Antenna, BodyShape}, {Antenna, BaseType}, {BodyShape, BaseType}
Yes	Round	Round	Wheels	0.0028	0.0523	{Antenna, BodyShape}, {Antenna, BaseType}

Table 3: Overview of closest counterfactual explanations over all robot types. We consider 16 robots defined by four binary attributes: `Antenna`, `HeadShape`, `BodyShape`, `BaseType`. Each combination of characteristics (row) is predicted as predicted `Reliable` if it has an Antenna and one of the following conditions: a Round HeadShape, a Round BodyShape, or Wheels. Otherwise it is predicted `Defective`. Based on this specification, we obtain closest counterfactuals that allow flipping the prediction.