Discrimination Exposed?
On the Reliability of Explanations for Discrimination Detection

ANONYMOUS AUTHOR(S)

Explanations are often cast as tools to uncover algorithmic discrimination. Given a model, we can explain its predictions to identify the rationale behind each outcome. We can present these explanations to decision subjects to let them contest potentially discriminatory outcomes. We can also present them to auditors to flag biased models. These beliefs – which have motivated rules and regulations surrounding explanation – are founded on inherently unverifiable assumptions. These include assumptions about the causal relationship between the inputs of a model and protected membership, the reliability of explanation to reveal salient information, and the ability of consumers or auditors to use information to make accurate claims about discrimination. In this work, we evaluate the viability of these beliefs under best-case assumptions. We consider a simple task where we can associate each prediction with a ground truth label. We design a user study where we can train participants to detect discrimination using explanations and evaluate the accuracy of claims surrounding explanations. We evaluate detection performance as we control the saliency of proxies of protected attributes, human knowledge about protected class, and their knowledge of causal mechanisms. Our results show that explanations fail to reliably flag unfair predictions and underscore the need for alternative safeguards to detect discrimination.

1 Introduction

Machine learning models are routinely used to automate decisions that affect people – be it to approve a loan [80], an insurance claim [37], or a public service [78]. Over the past decade, it has become clear that deploying models can lead to discrimination, as their predictions or performance can change across *protected attributes* such as sex, age, or race [10, 72]. In applications like lending and hiring, such effects arise from *indirect discrimination* [73] as models without protected attributes (e.g., sex) assign predictions through proxies (e.g., credit_history).

Many rules and regulations to protect consumers from discrimination in these sensitive domains revolve around explainability. In effect, multiple jurisdictions reference "discrimination" as a core reason for a "right to an explanation" in "high-risk" applications (e.g., EU [74, 75], Brazil [12], Korea [38] and proposed legislation in the United States [1, 2]). Our reliance on explainability stems from a widely-held belief that explanations can reveal that "*an algorithmic decision is affected by a (legally) protected attribute.*"[79]. In the event that this belief were true, post-hoc explanation methods provide a substantial benefit. Namely, they could safeguard against discrimination in ways that are easy to operationalize [6, 8, 23, 27, 48, 54, 82] – e.g., to audit black-box models without interfering in model development, or to provide decision subjects with information to contest adverse decisions.

Despite explanations being central to enforcing anti-discrimination laws, there is little evidence they can fulfill this function effectively. Simply put, we currently do not know the answers to questions such as "If we provide consumers with an explanation, can they effectively detect proxies?" or "If we ask auditors to check for proxies using explanations, can they retrieve such proxies?" or "How sensitive is this to causal assumptions or access to data?" This is surprising since the right to an explanation in a major consumer application was enacted over fifty years ago [see e.g., the adverse action provision in ECOA 71]. In this case, evidence is lacking because evaluating explanations requires technical validation and usability testing. The algorithms must produce faithful, relevant explanations. Users must be able to understand and utilize them effectively. In discrimination detection tasks, we face yet another barrier as any claim is subject to assumptions related to chance and causality (e.g., which variable is a proxy, whether it affected a given decision, etc.).

In this paper, we aim to test if explanations can assist humans in detecting discrimination, and characterize the conditions under which this assistance is meaningful. Our goal is to produce evidence to inform policy or compliance – either that we need to consider an alternative mechanism or that we need to impose additional conditions on explanations. Our approach seeks to distill the most basic assumptions behind non-direct discrimination and create a minimal 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 main contributions include:

1. We present a formal model for discrimination detection with explanations. Our model highlights the assumptions needed to assess if explanations help users detect discrimination. We use it to identify potential 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 for key failure modes related to human interaction and provides full control over our task – a machine-learning model, causal assumptions, and explanations.
3. We conduct controlled human-subject experiments. Our results show that participants 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, 31, 52]. In these domains, fair treatment requires models to output similar predictions across protected groups (i.e., treatment parity). In practice, models may violate this principle as a result of indirect discrimination via proxy variables [see e.g., 73, for a review]. These issues have motivated a extensive stream of work to detect and mitigate discrimination – e.g., methods to train models that do not discriminate [see e.g., 86], to identify proxies in a third-party audit [see e.g., 4], and to enable reporting group or individual discrimination [21]. Our work formalizes discrimination by adopting a causal notion of fairness [see e.g., 43, 61] - e.g., "would my prediction change if I belonged to a different protected group." [39]

Our work is related to a stream of research on how humans interact with explanations [see e.g., 9, 14–18, 44, 45, 81, 84]. Many works study if and how explanations impact decision-making [9, 14–19, 34, 44, 45, 47, 77, 87, 88]. Studies on counterfactual explanations that we use in this work [see e.g., 22, 25, 30, 41, 42, 68–70, 83] show marginal improvements in decision-making [22, 49, 50, 76, 83] and debugging model behavior [3, 55, 65]. There is less work on using explanations to assess discrimination, with most works focusing on issues that can arise when computing explanations ??e.g., lack of fidelity or data-related issues]balagopalan2022road, dai2022fairness, mhasawade2024understanding. As we discuss, 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 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, 60] 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 results align with the emerging picture from studies such as by Goyal et al. [35]. The 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 [11, 51, 64, 85], and information content [7, 11, 57, 66, 67].

2 Framework

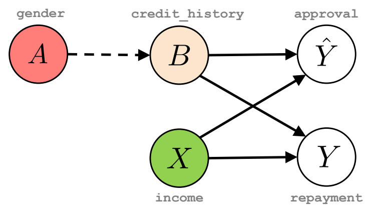


Fig. 1. Causal diagram for discrimination detection. Model $h : B \times X \longrightarrow \hat{Y}$ returns prediction $\hat{Y}$ of an outcome variable Y given input proxy B and features X. We seek to determine if model predictions change with respect protected attribute A through its proxy B, which 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, 53].

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 formalize this task through causal relationships between features and outcomes in a directed acyclic graph shown in Fig. 1. The model h is a deterministic function $h : X \times B \longrightarrow \hat{Y}$ that predicts an outcome Y (e.g., repayment). B denotes the proxy variable, and X denotes inputs that are independent of the protected attribute (e.g., X = income). The model satisfies two common assumptions:

1. *Indirect Discrimination.* The model does not use the protected attribute as input, but its predictions may change as a result of a variable B (e.g., B = credit_history) that is a *proxy* for the protected attribute. [4, 73]
2. *Business Necessity.* The proxy B can improve predictive accuracy, else the model owner could simply remove it from the list of features [32]

These assumptions are met by the vast majority of models in applications where we would care about discrimination. First, models directly using protected attributes would violate treatment disparity [10] by assigning different predictions to different groups, so they're typically omitted. Second, in cases where the proxy did not improve accuracy, a model owner could avoid scrutiny by training a model without it.

Characterizing Discrimination We determine the fairness of each feature vector based on a (relaxed) notion of *counterfactual fairness* [43].

Definition 1. Given a model h, we say that its prediction for a $(x, b, a) \in X \times B \times A$ is δ-counterfactually fair if changing the protected attribute can change the prediction by at most δ:

$$|\underbrace{\Pr(\hat{Y}_{A\leftarrow a} = h(x,b) \mid X = x,\ B = b,\ A = a)}_{\text{Current Prediction where } A = a} - \underbrace{\Pr(\hat{Y}_{A\leftarrow a'} = h(x,b) \mid X = x,\ B = b,\ A = a)}_{\text{Counterfactual Prediction when } A = a'}| \leq \delta$$

Here, $\hat{Y}_{A\leftarrow a}$ is the current prediction of the classifier, $\hat{Y}_{A\leftarrow a'}$ is the counterfactual prediction in a world where we set the protected attribute of the individual to $A = a'$, 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 $\Pr(\hat{Y}_{A\leftarrow a} = h(x,b) \mid X = x,\ B = b,\ A = a) = 1$ since there is no intervention required. We can compute $\Pr(\hat{Y}_{A\leftarrow a'} = h(x,b) \mid X = x,\ B = b,\ A = a)$ by setting the protected attribute to $A \leftarrow a'$ and propagating its effect on the proxy B. Given the causal structure in Fig. 1, we can express this term as:

$$\Pr(\hat{Y}_{A\leftarrow a'} = h(x,b) \mid X = x,\ B = b,\ A = a) = \sum_{b' \in B} \underbrace{\Pr(\hat{Y} = h(x,b) \mid X = x,\ B = b',\ A = a')}_{\text{Prediction for } b'} \cdot \underbrace{\Pr(B = b' \mid A = a')}_{\text{Proxy Strength}} \quad (1)$$

Taken together, a prediction $\hat{Y} = h(x,b)$ is δ-counterfactually fair if $|1 - \sum_{b' \in B} \Pr(\hat{Y} = h(x,b) \mid X = x,\ B = b',\ A = a') \cdot \Pr(B = b' \mid A = a')| \leq \delta$ The left hand side of this quantity is the probability 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 using a model to screening resumes in a job application, then we could set $\delta = 0.2$ to reflect the "4/5ths rule" in U.S. employment discrimination law [28]. In what follows, we remain agnostic about the value of δ and evaluate the potential to detect discrimination over all possible thresholds $\delta \in [0, 1]$.

Discrimination Detection with Explanations Many rules and regulations that mandate explanations as an anti-discrimination measure, based on the assumption that they help users identify and contest unfair predictions. We evaluate such claims by formalizing our problem as a detection task. Given a model h we associate each instance with:

- A "ground-truth" label $g_{i|h,\delta} := \mathbb{I}[p^{\text{flip}}_{x_i,b_i,a_i} > \delta]$ that reflects actual discrimination in the prediction; it is an indicator the prediction is not δ-counterfactually fair.
- A "prediction" label $\hat{g}_{i|h,e_i}$ that denotes user's claim a prediction is discriminatory; it is derived from analyzing the prediction alongside the explanation e_i.

In what follows, we write $g_i := g_{i|h,\delta}$, $\hat{g}_i := \hat{g}_{i|h,e_i}$, and $p^{\text{flip}}_i := p^{\text{flip}}_{x_i,b_i,a_i}$ when their dependencies are clear from context.

Although the probability that a prediction flips when intervening on the protected attribute is fixed for individuals with identical features (x, b, a), the actual outcome of this intervention is random. Assuming it follows a Bernoulli distribution $G_i \sim \text{Bernoulli}(p^{\text{flip}}_{x,b,a})$, we can interpret g_i in terms of hypothetical proportions: among N individuals $(x_i, \hat{b}_i, a_)$, where each $\hat{b}_i$ is drawn based on $A \leftarrow a'_i$, a δ-counterfactually fair model would yield different predictions for δN individuals. Since users only see one prediction for instance i, we interpret $\hat{g}_{i|h,e_i}$ as their *personal probability* the prediction would change under an intervention on A [see e.g., 24, for more details about this interpretation].[1] We write this as $\hat{g}_{i|h,e_i} \approx \mathbb{I}[p^{\text{flip}}_i > \delta]$.

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

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

We expect the following:

- *Instance-Level Detection*: Explanations can support individual claims when the claims are aligned with ground-truth labels. In this case, we should have that $\hat{g}_{i|h,e_i} = g_{i|h,\delta}$ 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% which would mean most of the selected predictions are indeed discriminatory.
- *Model-level Detection*: 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 τ% 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] If one prefers a different interpretation of probability statements, then $\hat{g}_{i|h,e_i}$ can be reinterpreted; for example, $\hat{g}_i = 1$ could be understood 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 and that $h(x_i, b_i) \neq h(x_i, b_i')$ 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 hides the proxy but e_i' shows it [13, 40]. This could lead the user seeing e_i erroneously determine that the counterfactual prediction never changes, i.e., $Pr(\hat{Y}_{A\leftarrow a_i'} = h(x_i, b_i) \mid x_i, b_i, a_i) = 1$, 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). *User may have incorrect beliefs about the strength of the proxy $Pr(B \mid A)$, 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 $1 - p_i^{flip}$ instead of p_i^{flip}, and make inaccurate discrimination 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. Conversely, they could misattribute discrimination if they are shown an explanation that highlights $h(x_i, b_i) \neq h(x_i', b_i)$ 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 first three failure modes 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 inadvertently discriminatory or not.

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: $B := \mathbb{I}[\text{Antenna} = \text{Yes}]$, $X_1 := \mathbb{I}[\text{HeadShape} = \text{Round}]$, $X_2 := \mathbb{I}[\text{BodyShape} = \text{Round}]$, $X_3 := \mathbb{I}[\text{BaseType} = \text{Wheels}]$. 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).

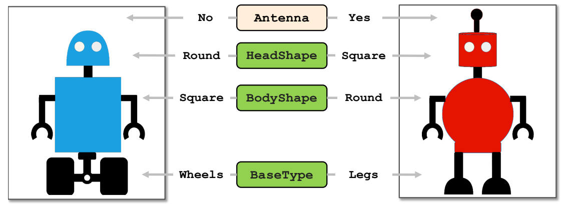


Fig. 2. Overview of robot characteristics. We show two robots to cover all possible values of each characteristic. Our model predicts that each robot is reliable or defective using dummy variables $B = \mathbb{I}[\text{Antenna} = \text{Yes}]$), $X_1 = \mathbb{I}[\text{HeadShape} = \text{Round}]$, $X_2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

We can arbitrarily increase the number of distinct robots to show participants by introducing spurious features, such as $\text{Paint} \in (\text{Red}, \text{Blue})$. 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 task where each case presents unique characteristics.

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

$$A, X_1, X_2, X_3 \sim \text{Bernoulli}(0.5)$$
$$B \mid A \sim \text{Bernoulli}(p_{B|A}) \quad \text{where } p_{B|A} \text{ is a set in Table 1}$$
$$Y \sim \text{Logistic}(B + X_1 + X_2 + X_3).$$

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

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

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

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 $1 - \Pr(B = b \mid A = a')$, 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 $h(x_i, b_i)$ 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 $h(x_i, b_i) = h(x_i, 1 - b_i)$), and explanations e_i' that do not use b (providing no insight about this relationship). 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.

	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 $h(x, b) \neq h(x, 1 - b)$. In other cases, the flip rate is 0.

Our setup allows 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 $g_{i|h,\delta}$. 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

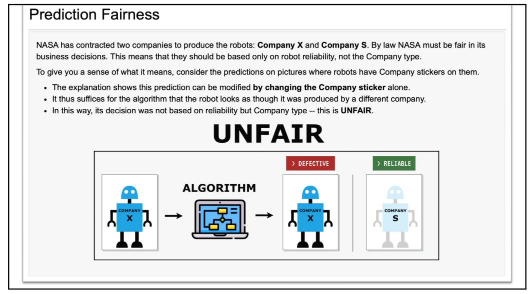
Our experiment sought to characterize the viability and effectiveness of explanations in detecting algorithmic discrimination. In particular, we sought to determine if individuals could use explanations to make reliable discrimination claims across use cases in consumer protection. Our specific research questions include:

- **RQ1** Can participants use explanations to make reliable claims for discrimination at an instance level? 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 who are shown explanations make reliable claims for discrimination at a model level? If so, this would suggest that explanations could serve as an effective mechanism to audit models.
- **RQ3** How does the reliability of claims depend on the information that is available to participants? In particular, explanations may be a viable mechanism only in use cases where participants have perfect information on the protected attributes of each instance (e.g., in a third-party audit).
- **RQ4** How does the reliability of claims depend on the correctness of causal assumptions (e.g., does 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 could provide participants with multiple explanations for each prediction? If so, this would speak to the importance of diverse explanations [see, e.g., 59]
- **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.

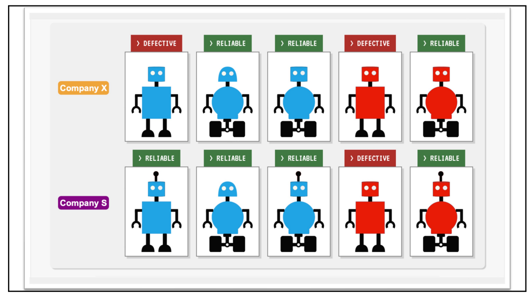
4.1 Setup

We used a study design with $2 \times 3 = 6$ conditions in which we varied the strength of the proxy variable $\in$ {Weak Proxy, Medium Proxy, Strong Proxy} and the format of explanations $\in$ {Single, Multiple}.

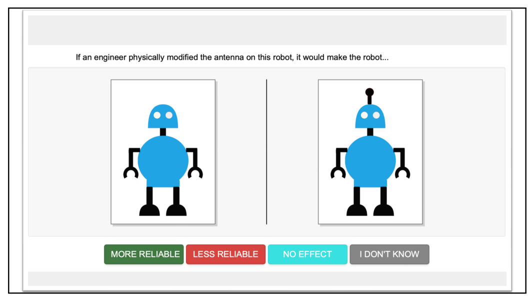
1. **Training**: Participants were introduced to four key elements of the study: robots, their components, a reliability prediction model, and the concept of discrimination. We used counterfactual explanations as the explanation method and presented them visually by highlighting modifiable robot parts. 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.



2. **Anchoring**: We presented participants with a set of robots to anchor their beliefs on the strength of the proxy and its impact on reliability. Each participant saw 10 robots from each company. We arranged the robots so that robots from each company shared the same features. We then assigned reliability labels and antenna to robots to anchor their beliefs on the impact of the proxy. The set contained two defective robots from $\mathrm{Company\ X}$ and one from $\mathrm{Company\ S}$. All robots from $\mathrm{Company\ X}$ had no antenna. $\mathrm{Company\ S}$ had 1/3/5 robots with antennas depending on the regime. Participants were explicitly told which feature distinguished the sets and were informed that proxy-based predictions *could* be discriminatory since the antenna can behave like the company sticker.



3. **Elicitation**: Participants were elicited for their beliefs on the protected class c_i and the effect of the proxy u_i on each possible robot. Participants saw a total of 16 robots for (X, B). We elicited beliefs regarding protected class by asking them to predict its manufacturer or state they don't know. We coded these as $\{0, 1, ?\}$. We elicited beliefs regarding the impact of the proxy by asking them how adding (or removing) an antenna from the robot would change reliability, allowing them to answer (more, less, no effect, or unknown), coded as $\{-1, 0, 1, ?\}$. Given the participant's beliefs in the protected attribute, we could recompute performance with participant-assumed attributes; we could also estimate $\Pr(B \mid A)$ to match their belief in the causal mechanism of the proxy. By storing reliability beliefs, we could analyze if these beliefs affect discrimination claims.



4. **Auditing**: Participants judged if predictions were discriminatory. They were shown an image of a robot, its prediction (always $\mathrm{Defective}$), and one or more of the closest counterfactuals. The participant aimed to select whether the 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). We collected discrimination claims $\hat{g}_i \in \{0, 1\}$.

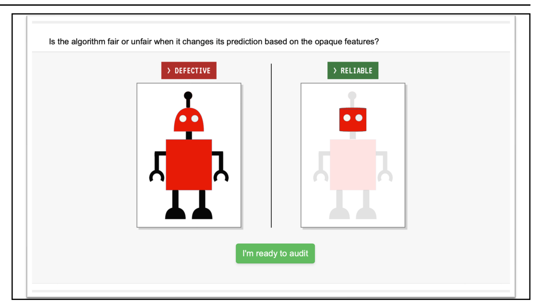


Fig. 3. All four phases of our experiment with their description.

1. Single: Participants were shown a single explanation for each prediction. This mimics real-world scenarios where participants might be given "the best explanation" or just *some* explanation and need to decide about discrimination. In this setup, an explanation might show no dependence on the proxy, but the prediction could still heavily rely on it, making it potentially discriminatory.

2. Multiple: Participants were presented with 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 proxy strength. The sets differed by the number of robots in Company S that had antennas: 1 robot for the Weak Proxy conditions, 3 robots for the Medium Proxy conditions and all five robots for the Strong Proxy conditions. Our evaluation also considered different levels of knowledge in the task:

1. Auditor Baseline: 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 assumption where the protected attributes are not readily available, and auditors have internal estimates of the true distributions.
2. Known Protected Attribute: Participants have perfect information about the protected attributes according to their elicited beliefs. This maps to an information regime where the auditor has access to the protected attributes (e.g., filing claims from consumers, or a third-party audit where the protected attributes are stored according to the law, such as audits (in New York) of employment decisions [36]).
3. Known Causal Mechanism: Participants have perfect information about the causal mechanism, i.e., the conditional distribution of the proxy matches their elicited beliefs. This is an idealized assumption and allows us to estimate best-case performance.

Counterfactual Explanations We let 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 $e(\boldsymbol{x}, f)$ that satisfies $f(\boldsymbol{x} + e(\boldsymbol{x}, f)) = 1$. When the set is minimal, we say that $e(\boldsymbol{x}, f)$ 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. First, they are easy to convey to participants because we can highlight the features that must change visually. Second, we can provide participants with clear guidelines on how to use them to correctly flag unfair predictions (i.e., via a comprehension quiz). Third, they directly relate to participant claims $\hat{g}_i$, and the fact 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 they reveal all ways in which a prediction can be flipped. We included a set of comprehension questions prior to the Auditing phase. Participants who failed this quiz three or more times were excluded from the study (10 excluded participants; exclusion rate of 8%). These quizzes ensured that participants understood how to apply each explanation and its guarantees with respect to discrimination claims.

4.2 Results

Overall, our results show that participants cannot reliably detect discrimination with explanations under any setup. The summary performance measurements of audits where participants were asked to flag discriminatory predictions based on a single explanation can be found in Fig. 6.

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 would change by over 20% given an intervention on protected group membership – given its importance in U.S. employment law [36].

As seen in Fig. 4, PPV, a measure of reliability of participant claims, indicates poor detection performance across all tested conditions. We would expect perfect, or at least very high PPV, say ≈ 90%, meaning that participants' detection is generally trustworthy. To the contrary, we observe that even in the Strong Proxy condition, where the proxy was the easiest to spot and its presence in the explanation most often indicated discrimination, PPV was as low as 48% ± 4% (see the blue boxes in Fig. 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 2-3 fair predictions. They also missed at least 3 out of 5 all discriminatory predictions.

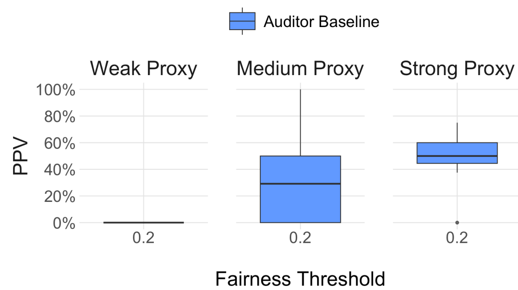


Fig. 4. Distribution of the Positive Predictive Value (PPV) at threshold $\delta = 0.2$ used in U.S. employment law [28] across all proxy strength conditions assuming the ground truth probabilities and causal mechanism of the proxy.

These results raise concerns about using explanations for discrimination auditing in practice. Without additional assumptions or safeguards, humans both fail to detect most of discriminatory cases, and raise multiple false alarms. This combination risks letting discriminatory practices continue and triggering unnecessary investigations that waste resources and potentially harm legitimate practices.

This poor performance is not due to the particular fairness threshold we selected. As seen in the blue line in Fig. 6, poor performance is observed systematically for all measures and almost all thresholds. This changes only at extreme values. For sufficiently high thresholds, all predictions become fair and since participants did claim discrimination, their performance drops. Conversely, at very low thresholds ($\delta \leq 5\%$ that exemplify a "better safe than sorry" approach), most proxy-dependent predictions are discriminatory. Since participants tend to flag these predictions, they achieve high PPV (≈ 75%) but still maintain poor TPR and FPR of ≈ 30%.

On the Sensitivity to Protected Attributes A natural question is whether the poor detection performance stems from a lack of knowledge of protected attributes. Perhaps participants reasoned about the hypothetical predictions under wrong assumptions. To answer this question, we matched participants' attribute selections from the Elicitation phase with the corresponding predictions.

Our results (see 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 ≈ 1 prediction), and TPR decreased by 6-7%, both across all conditions. This suggests that participants sometimes chose not to flag discrimination even when their own beliefs about protected attributes would warrant it. This 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 it has no effect – in 50% of the cases.

In total, knowledge of the protected attributes played a marginal role in detection performance. Even with access to these attributes, auditors still missed many discriminatory cases and raised multiple false alarms. As shown in Fig. 6, 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, leading to higher PPV (most claims were accurate), though their overall detection ability remained poor (low TPR and high FPR).

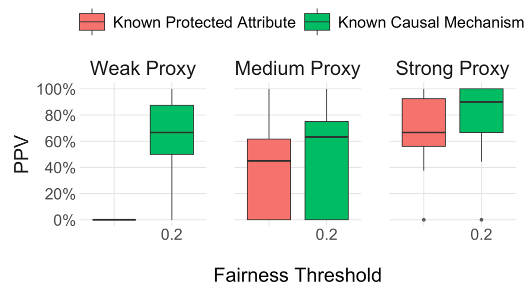


Fig. 5. Distribution of the Positive Predictive Value (PPV) at threshold $\delta = 0.2$ used in U.S. employment law [28] across all proxy strength conditions and under different assumptions on participant knowledge: known protected attributes (red), and known causal mechanism (green).

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. Still, neither PPV nor TPR/FPR ever reached a value we would consider satisfactory, as seen in Fig. 6. Overall, these results point to the fact that the lack of 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, they knew with certainty whether the prediction can 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 lead to good performance as we show in the Appendix in Fig. 11. On average, PPV was bounded by 40% across all conditions. TPR behaved irregularly but never exceeded 40%. FPR remained consistently at least 30%. The only exception occured 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, i.e., no knowledge (baseline), knowledge of protected attributes or knowledge about the causal mechanism of the proxy. Overall, people appear to be incapable of using explanations reliably even under idealized knowledge conditions.

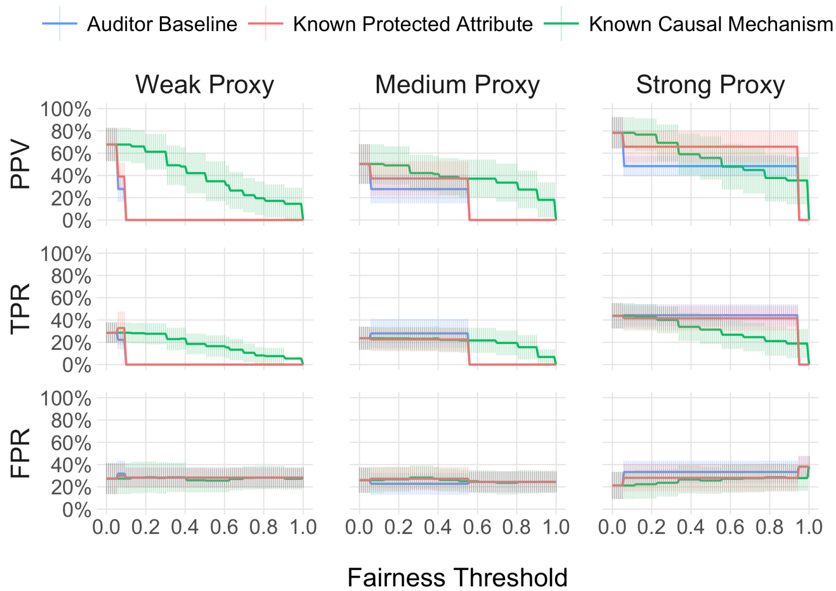


Fig. 6. Reliability of discrimination claims across all possible $\delta \in [0, 1]$ (right). We show the confidence intervals for PPV(δ), TPR(δ) and FPR(δ) across all proxy strength conditions and under different assumptions on participant knowledge: baseline performance (blue), known protected attributes (red), and known causal mechanism (green).

On Model Audits Participants were unable to differentiate between cases when the model was fair or discriminatory. In a task 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 for Weak Proxy, 13/19 for Medium Proxy, 12/19 for Strong Proxy participants 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 deployment of multiple fair ones.

On the Consistency of Auditors and Decision Subjects Our evidence shows that participants' claims were primarily driven by the presence of proxy variables in explanations. As expected, participants claimed discrimination 25-46% more frequently when explanations contained the proxy compared to when they did not (see Fig. 7). This effect was even more pronounced (36-60%) when participants viewed Multiple explanations. The increased exposure to explanations that contained the proxy in these conditions (14 instances versus 8 in the Single explanation conditions) led to a 30-47% increase in discrimination claims overall. These findings strongly suggest that proxy visibility directly impacts discrimination claims.

While participants were responsive to the presence of the proxy variable in the explanation, they often exercised nuance. In particular, we observed that participants consistently claimed that some predictions were "fair" even when the CE contained the proxy were judged as discriminatory. This behavior appears to be influenced by three systematic factors. First, their beliefs about robot reliability affected fairness judgments. Predictions were more likely to be labeled as fair by up to 20% when participants believed the proxy indicated higher reliability. While this pattern shows high variability ($p \approx 0.3$), it consistently appears across proxy conditions and aligns with participants' explicit statements (e.g., *It is not unfair to say that robots with antennas work better*). The other two key factors are that participants assumed different protected attributes, which led them to state no discrimination and misrepresented the true proxy strength.

Second, participants held false beliefs about the causal *structure* of the problem as described in Remark 3. We observed steady, low FPR of $\approx$ 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 $h(b, x) = h(b', x)$ 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. This belief makes sense but shows the danger of interpreting the presence of the proxy as a single indicator of discrimination.

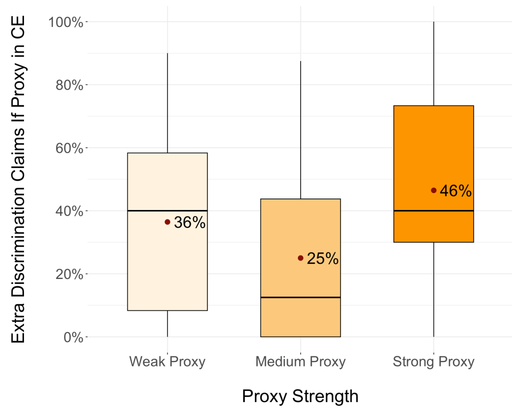


Fig. 7. Increase in discrimination claims when explanations contained the proxy versus when they did not. Mean values (red dots) show participants consistently identified the proxy as a discrimination signal across all regimes.

4.3 Discussion

By using a controlled environment with clearly defined ground truth, we were able to precisely measure how explanations fail to support discrimination detection. This approach provided participants with optimal conditions: clear information about proxy mechanisms, explicit explanations showing counterfactual outcomes, and detailed instructions. The fact that explanation-based discrimination detection failed under these favorable conditions, or even when adapting the ground truth to participant beliefs, suggests fundamental limitations of using explanations to detect discrimination. We discuss this 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, or 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, reliability of detection oscillated around 50% with false alarms consistently hovering around 30% (FPR). To put that into perspective while being lenient on the participants' performance, this means every fourth individual that files a discrimination claim fails in court. This also means almost half of individuals whose predictions were truly discriminatory miss this.

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 half spot 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 model or a set of predictions were analyzed by two independent auditors, it could lead to two different results. A discriminatory model could then be missed, and a fair model could be unfairly accused of discrimination.

The fundamental reason why explanations failed to aid 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 [63], and our experiments demonstrate how impairs human performance. Our analysis revealed three specific challenges that emerged from this mismatch and were the direct causes of people's failure:

Flawed Beliefs in the Causal Structure 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).

Proxy Strength Misrepresentation 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' responses, saying e.g., *It is not unfair to say 'robots with antennas work better'.*

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 for algorithmic fairness auditing. Through controlled experiments 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 the general causal mechanism of the proxy.

Our findings extend to real-world auditing scenarios. This is because real-world 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, incorrectly estimating 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 like the compounded 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 [62] or provisions for a "Right to an Explanation" in data regulation laws in the European Union [74], Brazil [12], and South Korea [38]. Mandatory fairness audits represent the second regulatory approach, e.g. in Slovenia mandates for algorithm pre-implementation [58], or in New York for third-party bias audits for automated employment decisions [36]. Similarly, the European Union's Digital Services Act requires algorithmic audits of "very large online platforms," including non-discrimination risk assessments [75]. Despite this momentum, there remains a lack of standardized practices for assessing algorithmic fairness as regulations provide limited guidance for how to conduct audits [46]. 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 [33] 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.

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] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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).

[17] 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.

[18] 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.

[19] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *arXiv preprint arXiv:2301.07255* (2023).

[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] 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.

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

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

[25] Eoin Delaney, Arjun Pakrashi, Derek Greene, and Mark T Keane. 2023. Counterfactual explanations for misclassified images: How human and machine explanations differ. *Artificial Intelligence* 324 (2023), 103995.

[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] Maximilian Förster, Mathias Klier, Kilian Kluge, and Irina Sigler. 2020. Evaluating Explainable Artifical Intelligence - What Users Really Appreciate. In *28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020, Marrakech, Morocco, June 15-17, 2020*, Frantz Rowe, Redouane El Amrani, Moez Limayem, Sue Newell, Nancy Pouloudi, Eric van Heck, and Ali El Quammah (Eds.). https://aisel.aisnet.org/ecis2020_rp/195
[31] 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.
[32] 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.
[33] Talia B Gillis and Josh Simons. 2019. Explanation< Justification: GDPR and the Perils of Privacy. *JL & Innovation* 2 (2019), 71.
[34] Ana Valeria González, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivasan Iyer. 2021. Do explanations help users detect errors in open-domain QA? an evaluation of spoken vs. visual explanations. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 1103–1116.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] Issa Kohler-Hausmann. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.* 113 (2018), 1163.
[40] 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.
[41] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2022. Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2125–2137.
[42] Ulrike Kuhl, André Artelt, and Barbara Hammer. 2023. For Better or Worse: The Impact of Counterfactual Explanations' Directionality on User Behavior in xAI. In *World Conference on Explainable Artificial Intelligence*. Springer, 280–300.
[43] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. *Advances in neural information processing systems* 30 (2017).
[44] 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.
[45] 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.
[46] 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.
[47] Olesja Lammert, Birte Richter, Christian Schütze, Kirsten Thommes, and Britta Wrede. 2024. Humans in XAI: increased reliance in decision-making under uncertainty by using explanation strategies. *Frontiers in Behavioral Economics* 3 (2024), 1377075.
[48] 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.
[49] 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).
[50] 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.
[51] 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.
[52] 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.
[53] Geng Li. 2018. Gender-Related Differences in Credit Use and Credit Scores. *FEDS Notes* (22 June 2018). https://doi.org/10.17016/2380-7172.2188
[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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. https://doi.org/10.1145/3351095.3372850

[60] 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).

[61] 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.

[62] 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

[63] 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.

[64] 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.

[65] 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).

[66] 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.

[67] 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.

[68] Ronal Singh, Tim Miller, Henrietta Lyons, Liz Sonenberg, Eduardo Velloso, Frank Vetere, Piers Howe, and Paul Dourish. 2023. Directive explanations for actionable explainability in machine learning applications. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–26.

[69] Nina Spreitzer, Hinda Haned, and Ilse van der Linden. 2022. Evaluating the Practicality of Counterfactual Explanations.. In *XAI. it@ AI* IA*. 31–50.

[70] Ilia Stepin, Jose M Alonso-Moral, Alejandro Catala, and Martín Pereira-Far iña. 2022. An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information. *Information Sciences* 618 (2022), 379–399.

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

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

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

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

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

[76] 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.

[77] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

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

[79] 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.

[80] 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).

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

[87] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

[88] Zelun Tony Zhang, Felicitas Buchner, Yuanting Liu, and Andreas Butz. 2024. You Can Only Verify When You Know the Answer: Feature-Based Explanations Reduce Overreliance on AI for Easy Decisions, but Not for Hard Ones. In *Proceedings of Mensch und Computer 2024*. 156–170.

A Table of Notation

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
$g_{i\|h,\delta}$	Ground truth label indicating discrimination in prediction i
$\hat{g}_{i\|h,e_i}$	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.

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.

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.

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

Anonymized data from the experiments is available at https://anonymous.4open.science/r/cxai-93BB/results/results_closest_competing.

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

The code for our Flask study is available at https://anonymous.4open.science/r/cxai-93BB/.

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 [56]. 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 Proxycondition (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. Participants could select both fair and unfair answers in these cases and were told the discrimination label is uncertain and whether its influence (the SHAP value) is high enough to make the prediction depend on it. 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. 8) to make sure they understand the relative on influence on prediction that SHAP values communicate. As seen in Fig. 9, 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.

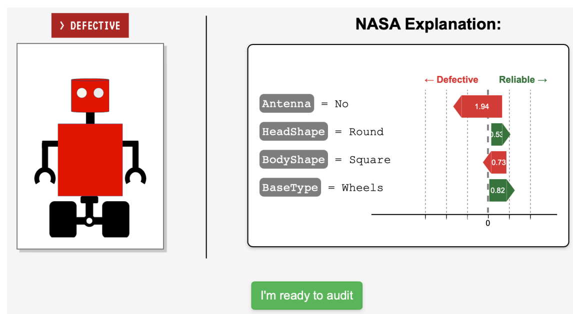


Fig. 8. Example of a SHAP explanation in our study.

C.2 Experiments in the Main Text

Fig. 11 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. 12 shows that participants' claims depended on the presence of the proxy in the explanation also for Multiple explanations conditions. Finally, Fig. 10

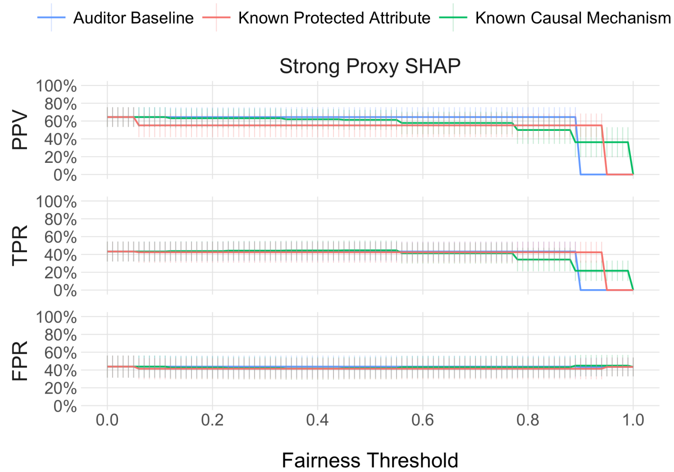
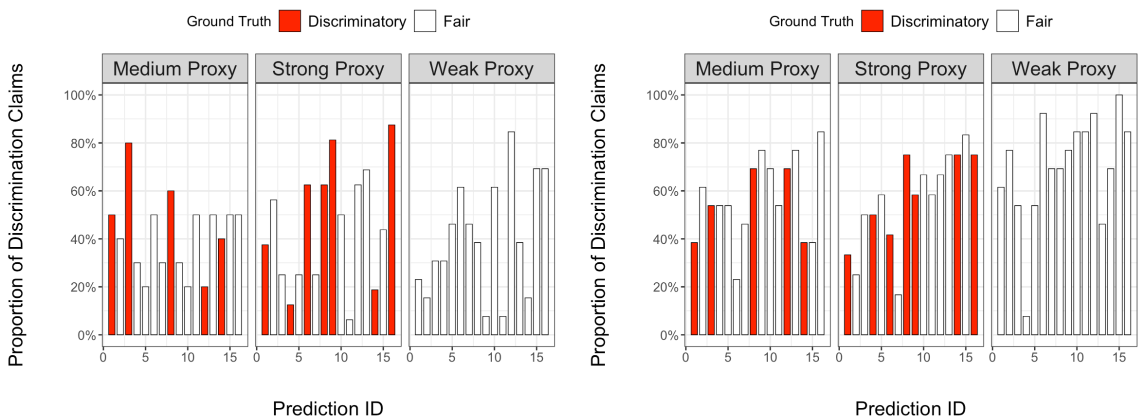


Fig. 9. Performance metrics across all fairness threshold δ values when participants were assisted by SHAP explanations. Refer to Fig. 6 for the explanation of the plotted data. As seen, the detection is poor across all fairness thresholds with consistently high FPR, and consistently low TPR, both around 40%. This results in low reliability of claims as measured by PPV.



(a) Single explanation

(b) Multiple explanations

Fig. 10. 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%.

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 claim as discriminatory.

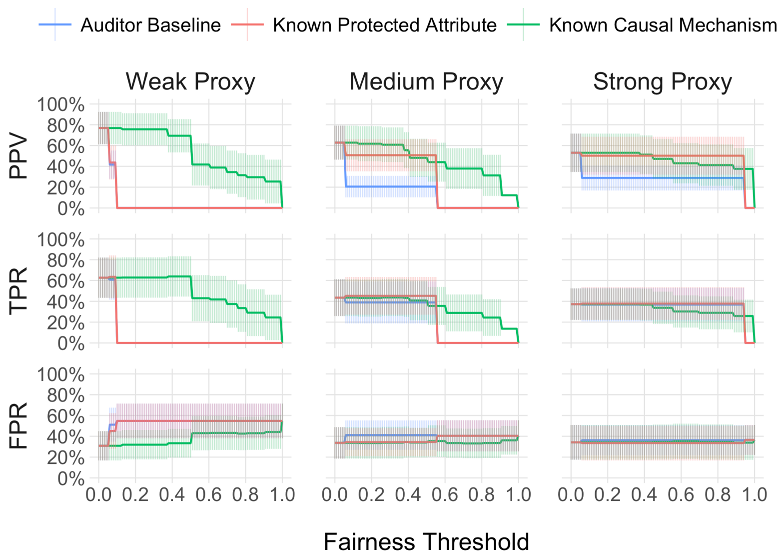


Fig. 11. Refer to Fig. 6 for the explanation of the plotted data. As shown, baseline performance (blue) is poor across all thresholds, with TPR not exceeding 50% and FPR around 30%, and sometimes exceeding this value. Knowledge of protected attributes (red) yields significant gains for PPV for medium and strong proxies but is otherwise unhelpful. Assuming auditors' beliefs about the causal mechanism (green) provides the biggest gains for performance, especially internal reliability in terms of PPV. It still leads to largely low TPR and moderate FPR. The latter metric remains problematic (around 30%) across all conditions, indicating persistent incorrect assumptions about feature-protected attribute relationships regardless of the level of insight.

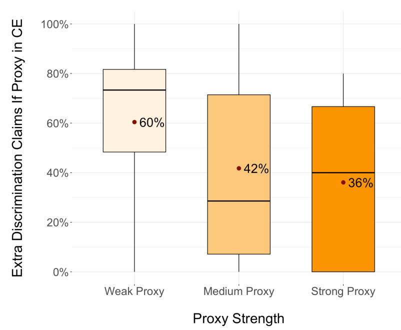


Fig. 12. Increase in discrimination claims when explanations contained the proxy versus when they did not. Mean values (red dots) show participants consistently identified the proxy as a discrimination signal across all regimes. The strongest effect occurs for the weak proxy because participants overestimated its strength.