INTERVENTIONS AGAINST MACHINE-ASSISTED STATISTICAL DISCRIMINATION

JOHN Y. ZHU[1]

March 12, 2024

Abstract

This article studies how to intervene against statistical discrimination, when it is based on beliefs generated by machine learning, rather than by humans. Unlike beliefs formed by a human mind, machine learning-generated beliefs are verifiable. This allows interventions to move beyond simple, belief-free designs like affirmative action, to more sophisticated ones, that constrain decision makers in ways that depend on what they are thinking. Such *mind reading* interventions can perform well where affirmative action does not, even when the beliefs being conditioned on are possibly incorrect and biased.

JEL Codes: C53, D86, J71, L51
Keywords: discrimination, algorithm, fairness, human-AI, robust, bias, misspecified

[1]University of Kansas, johnzhuyiran@ku.edu.

1 Introduction

This article studies statistical discrimination by a decision maker (DM) who uses machine learning to compute beliefs. Unlike human beliefs, machine beliefs are verifiable. I show how conditioning on such beliefs can make interventions against statistical discrimination more effective – even if those beliefs are possibly incorrect and biased.

To fix ideas, consider the canonical statistical discrimination model of Coate and Loury (1993). A DM makes binary accept-reject decisions on a set of subjects, each with a group identity i, and a vector of other features x. Some x are more likely to occur if a subject is qualified. Any subject can invest to become qualified at a random cost that is iid across subjects. Qualification status, y, is unobserved by the DM. Instead, for each (i, x) subject, the DM forms a belief, $f(i, x)$, about the probability $y =$ qualified, and accepts them if and only if $f(i, x)$ is sufficiently high. It can be shown that, even though groups are ex-ante identical, an unfair equilibrium can still emerge due to self-fulfilling negative stereotypes: Relative to the favored group, subjects of the discriminated group face tougher acceptance standards (fewer x are accepted), which leads to fewer of them investing to become qualified, which rationalizes the tougher standards they face.

Coate and Loury (1993) then consider how affirmative action affects such statistical discrimination. Affirmative action requires the DM to accept subjects from different groups at equal rates. In all unfair equilibria, affirmative action *brings gains to the discriminated group*, by inducing the DM to lower the acceptance rate of the favored group while increasing the acceptance rate of the discriminated group, shrinking the acceptance rate gap. However, in the long run, these gains need not lead to a fair equilibrium, that would allow the intervention to be lifted. Indeed, under affirmative action, there exist steady states where the previously discriminated group is now patronized with extra low acceptance standards, and so continues to have less incentive to invest in becoming qualified. Without affirmative action, such a state is not in equilibrium. If the intervention is lifted, the patronizing lower standards revert to higher standards, and the original discrimination can re-emerge. This begs a natural question:

> Are there *ideal* interventions that bring gains to the discriminated group in all unfair equilibria *and* whose steady states are the set of fair equilibria?

I show that if the DM's beliefs, $f(i, x)$, are machine learning-generated, so that interventions can condition on them, then ideal interventions exist.

One example is what I call *mistaken identity*: Fix arbitrary nonnegative group identity weights $\{\alpha_i\}$, not all zero, and assign all subjects with the same x the same score, $\sum_i \alpha_i f(i,x)$. The DM is then required to choose a threshold and accept/reject a subject if their score is above/below that threshold. Loosely speaking, mistaken identity forces the DM to act as if it thinks all subjects have the same (possibly blended) identity. It is related to the proposals by Yang and Dobbie (2020) and Pope and Sydnor (2011) for eliminating proxy effects of group identity in linear regressions. Mistaken identity is color-blind, since all subjects with the same x are treated equally. However, it is more restrictive than just requiring the DM be color-blind. "Just be color-blind" is a belief-free intervention that still allows a clever DM to use proxies for group identity to discriminate. In contrast, mistaken identity uses the DM's own race-conscious beliefs against them to prevent such proxying.

Mistaken identity works effectively even when the DM holds incorrect, biased beliefs. Intuitively, biases causes the DM to perceive a qualification gap between groups that may differ from what actually exists. But this is undone by the intervention forcing the DM to think that everyone has the same identity. In addition to mistaken identity, I demonstrate the existence of other ideal interventions, such as what the algorithmic fairness literature calls *equal opportunity*. However, I show that equal opportunity lacks the robustness of mistaken identity to incorrect, biased beliefs.

Recently, the Supreme Court, in *SFFA v. Harvard* and *SFFA v. UNC*, ruled that admissions programs that factor in race violate the Equal Protection Clause of the 14th Amendment. Affected parties are now exploring how to preserve the gains made by minorities under decades of affirmative action. One promising option is to implement mistaken identity, since it yields attractive outcomes against statistical discrimination, is robust to incorrect, biased beliefs, and is color-blind. Indeed, many legal scholars have argued that color-blind interventions do "uphold the principles of the equal protection doctrine." (Yang and Dobbie, 2020) However, it is worth emphasizing, implementing the color-blind mistaken identity intervention still requires conditioning on underlying beliefs that are not color-blind. Based on its recent opinions, it is unclear what the Supreme Court thinks of such indirect race-consciousness. And yet, I show that such race-consciousness is necessary. Specifically, if the DM only has access to color-blind data, resulting in color-blind beliefs, then an identification problem arises that makes it impossible for any intervention to be ideal. Thus, my work justifies the need to create a clear space within the law for the use of color-blind interventions that, nevertheless, condition on not color-blind beliefs.

Related Literature. This article bridges two distinct but related literatures – statistical discrimination and algorithmic fairness. Both address the same problem: Unfair decision making driven by the statistical learning of some unobserved variable of interest, y, based on observed features (i, x). Typically, when this problem occurs, the joint distribution of (x, y) varies with i, which then causes the learning of y from x to be sensitive to i. The statistical discrimination literature has produced a rich theory of why the joint distribution of (x, y) can vary with i, and how sensitivity of statistical learning to i can cause unfair decisions to be made. See Fang and Moro (2011) and Onuchic (2022) for excellent surveys of the literature. Less progress has been made designing effective interventions. In particular, little attention has been paid to interventions that condition on what is statistically learned about y given (i, x) – that is, the beliefs about y. Conversely, the algorithmic fairness literature has provided numerous interventions to algorithms that achieve various fairness goals, including many designed to condition on machine learning-generated beliefs.[2] But much of this work is not guided by economic theory, leading to well known problems detailed below. My work attempts to synthesize the strengths of the two literatures, by evaluating a broad class of belief-contingent interventions within an economic model of statistical discrimination, in order to find robust, welfare enhancing interventions against unfair decision making driven by statistical learning.

Statistical Discrimination. Broadly speaking, theories of statistical discrimination fall into two groups, Phelpsian and Arrovian.

In Phelpsian models (Phelps, 1972; Aigner and Cain, 1977), the variation of the joint distribution of (x, y) with i is exogenous. As a result, it is difficult to build a theory of intervention design within the Phelpsian framework. Instead, recent research focuses on an important precursor to effective intervention – correct diagnosis of the problem. This involves characterizing information structures and outcome distributions consistent/inconsistent with statistical discrimination (Chambers and Echenique, 2021; Escudé et al., 2022; Deb et al., 2022; Martin and Marx, 2022), developing empirical strategies for distinguishing statistical discrimination from taste-

[2] Today, algorithmic decision-making is ubiquitous (Agrawal et al., 2018). Kleinberg et al. (2018a) demonstrates how algorithmic decision-making can substantially outperform human decision-making in a high stakes setting. Nevertheless, there is concern that algorithmic decision-making can be unfair, reinforcing pre-existing discriminatory behavior. For example, in a study on household credit markets, Fuster et al. (2022) shows how certain minority groups are adversely affected by the introduction of machine learning algorithms for predicting creditworthiness. The field of algorithmic fairness emerged in response to such concerns. See Barocas et al. (2019) for a textbook treatment.

based discrimination (Arnold et al., 2021; Marx, 2022), and identifying when statistical discrimination is being driven by inaccurate beliefs (Bohren et al., 2019, 2023). My result about how color-blinding data creates an identification problem, making certain fair and unfair equilibria statistically indistinguishable, is related to this line of research.

Arrovian models (Arrow, 1971) seek to explain the variation of (x, y)'s joint distribution with i as a possible equilibrium outcome of various things, including self-fulfilling expectations (Coate and Loury, 1993; Moro and Norman, 2004; Onuchic and Ray, 2023), search frictions (Rosén, 1997; Mailath et al., 2000; Jarosch and Pilossoph, 2019), rational inattention (Echenique and Li, 2022; Fosgerau et al., 2023), and learning traps (Bardhi et al., 2020; Che et al., 2020; Komiyama and Noda, 2020; Li et al., 2020). In these models, there is scope to consider interventions that can take us out of an unfair equilibrium, such as affirmative action, quotas, and various subsidies. However, the results have been mixed and the analysis has been somewhat ad hoc. In particular, belief-contingent interventions have been neglected.

In contrast, belief-contingent interventions are common in the algorithmic fairness literature. This distinction has been previously noted by Kleinberg et al. (2018b, 2020), who argue, more generally, that a fundamental difference between regulating human versus algorithmic decision-making is that, algorithms, unlike a human mind, can be audited.

Algorithmic Fairness. The myriad interventions emerging from algorithmic fairness can typically be categorized as targeting disparate treatment or disparate impact.

Interventions targeting disparate treatment are designed to not "factor in" group identity. Canonical examples include fairness through awareness (Dwork et al., 2012) and counterfactual fairness (Kusner et al., 2017; Chiappa, 2019). The regression modifications proposed by Yang and Dobbie (2020) and Pope and Sydnor (2011) can also be viewed as disparate treatment interventions.

Interventions targeting disparate impact are designed to ensure that the distribution of decisions experienced by different groups is fair in some statistical sense. Common statistical notions of fairness include statistical parity, equal opportunity, equal odds, group calibration, positive and negative class balance, predictive parity, and error rate balance. Zemel et al. (2013), Feldman et al. (2015), and Hardt et al. (2016b) introduce ways of constraining algorithms to guarantee that the selected decision policy satisfies statistical parity, equal opportunity, or equal odds.

Kleinberg et al. (2016) and Chouldechova (2017) show that some of the fairness goals disparate impact interventions are designed to achieve are mutually incompatible. Others have shown that these interventions can lead to Pareto-inferior outcomes and greater unfairness when measured in an application-appropriate way (Liu et al., 2018; Corbett-Davies et al., 2017; Hu and Chen, 2020; Corbett-Davies et al., 2023). Algorithmic fairness has also been slower to factor in agency problems many economists take for granted. These include conflicts of interest between the DM and whoever is imposing the intervention, and strategic responses of subjects to the DM being constrained by an intervention. All of this has led to calls to bring a more economic perspective to algorithmic fairness (Cowgill and Tucker, 2020).

The response has been a wave of new research aimed at better integrating algorithmic fairness and, more generally, algorithmic decision-making with economic theory. Innovations include explicit benefit functions (Heidari et al., 2018), limited ability to monitor or control the DM (Blattner et al., 2021; Liang et al., 2023), incentive effects on the DM (Fu et al., 2022), incentive effects on subjects (Hardt et al., 2016a; Eliaz and Spiegler, 2019; Hu et al., 2019; Kleinberg and Raghavan, 2020; Frankel and Kartik, 2022; Penn and Patty, 2023), retraining and equilibrium prediction (Perdomo et al., 2020), and long-term fairness considerations (Hu and Chen, 2018; Mouzannar et al., 2019; D'Amour et al., 2020; Liu et al., 2020; Puranik et al., 2022), including in the presence of biased beliefs (Williams and Kolter, 2019; Segal et al., 2023). Already, insights into interventions have emerged from this research. The equalized odds intervention has been shown to possess attractive equilibrium properties (Jung et al., 2020; Shimao et al., 2023), and algorithmic audits have been used to design interventions that can eliminate taste-based discrimination (Rambachan et al., 2020).

Human-AI Most work on human-AI interactions assume both machines and humans generate beliefs and make decisions. The focus is on if or when these two ostensibly substitutable parties can be combined to yield decisions that outperform any party by themselves (Athey et al., 2020; Gillis et al., 2021; Bastani et al., 2021; Donahue et al., 2022; McLaughlin and Spiess, 2022; Agarwal et al., 2023; Angelova et al., 2023). In contrast, machines and humans are complementary in my model: While the human provides the preferences, the machine provides all of the beliefs. This implies that interventions can fully condition on the beliefs of the DM, leading to sharp predictions. Future work should consider belief-contingent intervention design that can handle the human having private information about the distribution of y.

2 An Economic Model with Machine Learning

I take the statistical discrimination model of Coate and Loury (1993) and make one essential change: I replace their DM, who has rational expectations, with one who relies on machine learning to compute possibly incorrect but verifiable beliefs, that can then be used by a third-party.

There is a DM and a unit mass of subjects. Each subject possesses a vector of publicly observed features (i, x) and an unobserved binary class $y \in Y := \{q, u\}$. Here, $i \in I := \{w, b\}$ is the *group identity* (can generalize to more than two), and $x \in X := X_1 \times X_2 \times \ldots \times X_N$ are the *other features*. Assume X is finite.

The distribution of subjects over $I \times X \times Y$ is the result of actions taken by the subjects. Initially, each subject independently draws a group identity i and a cost $c \in (-\infty, \infty)$. Let λ_w and $\lambda_b = 1 - \lambda_w$ be the positive probabilities of drawing w and b, respectively, and let G and g denote the CDF and PDF of c, respectively. After a subject draws (i, c), they choose to join a class. Joining class q costs c, while joining class u costs 0. Assume g has full support. The possibility of negative costs, however remote, ensures that, in any equilibrium, there are subjects in both classes. Once a subject joins a class, their other features, x, are realized according to a statistical model of the following form:

Assumption 1. *There exists a nonempty* $\mathcal{Y} \subset \{1, 2, \ldots N\}$*, such that*

$$p(x|i, y) = p(x_{\mathcal{Y}}|y)p(x_{-\mathcal{Y}}|i, x_{\mathcal{Y}}) > 0 \quad \forall (i, x, y) \in I \times X \times Y.$$

Under Assumption 1, other features not indexed by $\mathcal{Y}$ can be viewed as proxies for group identity. At this point, the distribution of subjects over $I \times X \times Y$ is determined. It can be decomposed into a distribution, μ_{RE}, over the features space $I \times X$, and a conditional probability, f_{RE}, over the features space $I \times X$, that a subject is of class q. The DM obtains a possibly incorrect version, (μ, f), of (μ_{RE}, f_{RE}). From now on, I refer to μ as a *distribution* and f as a set of *beliefs*. Coate and Loury (1993) focus on the rational expectations case, $(\mu, f) = (\mu_{RE}, f_{RE})$, and consider the μ_{RE}-contingent intervention, affirmative action, but do not consider any intervention that conditions on f_{RE}. Since I interpret f as the verifiable output of machine learning, I study interventions that can condition on both μ and f.

After obtaining (μ, f), the DM makes a binary 1-0 decision (e.g., accept-reject, lend-deny, assign to high-low skill job) on each subject, by choosing a *decision policy*, defined to be a map $d : I \times X \to [0, 1]$. Given an (i, x) subject, d selects decision 1 with probability $d(i,x)$. If the selected decision is 1, the DM's payoff depends on the subject's class, $v_q > 0$ or $-v_u < 0$, while the subject's payoff is $\omega > 0$. If the selected decision is 0, both parties' payoffs are 0.

Clearly, given any d, if a group i subject with cost c is weakly better off joining class q, then any group i subject with strictly lower cost is strictly better off joining class q. Thus, for each group i, I imagine a representative group i subject choosing a cost threshold $\bar{c}(i)$ so that a group i subject with cost c joins class q if and only if $c \leq \bar{c}(i)$. This allows me to treat the model as a simultaneous-move game, $(X, \lambda_w, p, G, v_q, v_u, \omega)$, between three players: A pair of representative subjects choose a pair of cost thresholds $\bar{c} = (\bar{c}(w), \bar{c}(b))$ and the DM chooses a decision policy d.

Given $\bar{c}$, (μ_{RE}, f_{RE}) is

$$\begin{aligned} \mu_{RE}(i,x|\bar{c}(i)) &:= \lambda_i \left[G(\bar{c}(i))p(x|i,q) + (1 - G(\bar{c}(i)))p(x|i,u) \right], \\ f_{RE}(i,x|\bar{c}(i)) &:= \frac{G(\bar{c}(i))p(x|i,q)}{G(\bar{c}(i))p(x|i,q) + (1 - G(\bar{c}(i)))p(x|i,u)} \qquad \forall (i,x) \in I \times X. \end{aligned}$$

For simplicity, I will usually write $\mu_{RE}(i,x)$ and $f_{RE}(i,x)$, suppressing their dependence on $\bar{c}$. When working with a cost threshold pair, $\bar{c}^{\,\cdot}$, that is indexed with some superscript $\cdot$, I will write $\mu^{\cdot}_{RE}$ and $f^{\cdot}_{RE}$ to refer to μ_{RE} and f_{RE} given $\bar{c}^{\,\cdot}$. The same convention applies for subscripts.

A decision policy d can be viewed as a pair of group policies $(d(w), d(b))$ defined over X. Given $\bar{c}(i)$ and $d(i)$, the utility of the representative group i subject is

$$U_i(\bar{c}(i), d(i)) := \omega \sum_{x \in X} \frac{\mu_{RE}(i,x)}{\lambda_i} d(i,x) - \int_{-\infty}^{\bar{c}(i)} c g(c) dc.$$

To ensure U_i is well-defined, assume $\int_{-\infty}^{\infty} |c| g(c) dc < \infty$. Given d and (μ, f), the utility of the DM is

$$U_{DM}(d, \mu, f) := \sum_{(i,x) \in I \times X} d(i,x)\mu(i,x) \left[f(i,x)v_q - (1 - f(i,x))v_u \right].$$

2.1 Interventions

Given a set of other features X, the domain of distributions is the set of all maps $\mu : I \times X \to (0,1)$ satisfying $\sum_{(i,x) \in I \times X} \mu(i,x) = 1$, and the domain of beliefs is the set of all maps $f : I \times X \to (0,1)$.

Definition. *An intervention k specifies, for each (X, μ, f), a nonempty compact set, $k(X, \mu, f)$, of decision policies $d : I \times X \to [0, 1]$.*

Example. *Just be color-blind specifies, for each (X, μ, f), the set of all color-blind decision policies. A decision policy, d, is color-blind if $d(w, x) = d(b, x)$ $\forall x \in X$.*

Example. *Affirmative action specifies, for each (X, μ, f),*

$$k(X, \mu, f) = \left\{ d : I \times X \to [0, 1] \;\middle|\; AR(w, d, \mu) = AR(b, d, \mu) \right\},$$

where $AR(i, d, \mu) := \frac{\sum_{x \in X} d(i,x)\mu(i,x)}{\sum_{x \in X} \mu(i,x)}$ is the acceptance rate of group i.

Unlike the two previous interventions, the following interventions depend on the DM's machine learning-generated beliefs:

Example. *Equal opportunity specifies, for each (X, μ, f),*

$$k(X, \mu, f) = \left\{ d : I \times X \to [0, 1] \;\middle|\; TP(w, d, \mu, f) = TP(b, d, \mu, f) \right\},$$

where $TP(i, d, \mu, f) := \frac{\sum_{x \in X} d(i,x)\mu(i,x)f(i,x)}{\sum_{x \in X} \mu(i,x)f(i,x)}$ is the true positive rate of group i.

Example. *In a mistaken identity intervention, there exist group identity weights $\alpha_w, \alpha_b \geq 0$, not both zero, such that, for each (X, μ, f),*

$$k(X, \mu, f) = \left\{ d : I \times X \to [0, 1] \;\middle|\; d \text{ is a threshold function of } \alpha_w f(w, x) + \alpha_b f(b, x) \right\}.$$

A decision policy d is a threshold function of $\alpha_w f(w, x) + \alpha_b f(b, x)$ if d depends on (i, x) only up to $\alpha_w f(w, x) + \alpha_b f(b, x)$ – in particular, d is color-blind – and there exists an $\overline{f} \geq 0$ such that $\alpha_w f(w, x) + \alpha_b f(b, x) > \overline{f}$ $(< \overline{f}) \Rightarrow d(i, x) = 1$ $(= 0)$.

Definition. *Given an intervention k and a game $(X, \lambda_w, p, G, v_q, v_u, \omega)$, a strategy profile $(\overline{c}^*, d^*)$ is a k equilibrium supported by (μ, f) if*

1. *$\overline{c}^*(i) = \arg\max_{\overline{c}(i) \in \mathbb{R}} U_i(\overline{c}(i), d^*(i))$ for all groups i,*

2. *$d^* \in \mathcal{D}(k, \mu, f) := \arg\max_{d \in k(X, \mu, f)} U_{DM}(d, \mu, f)$.*

If $(\mu, f) = (\mu^*_{RE}, f^*_{RE})$, then drop the qualifier "supported by (μ, f)." If k is the trivial intervention that allows all decision policies, then drop the prefix k.

2.2 Characterizing Equilibria

This part reviews Coate and Loury (1993) Section I. Results are stated without proof.

Let $(\bar{c}^*, d^*)$ be an equilibrium. $(\bar{c}^*, d^*)$ can be decomposed into a pair of group equilibria, $\{(\bar{c}^*(i), d^*(i)) \mid i \in I\}$. Let us characterize group equilibria.

Define the *likelihood* function $l(x) := \frac{p(x_{\mathcal{Y}}|q)}{p(x_{\mathcal{Y}}|u)}$ for all $x \in X$. Since $p(x_{\mathcal{Y}}|y) > 0$ for all $(x, y) \in X \times Y$, l is a well-defined positive function taking finitely many values, $0 < l_1 < l_2 < \ldots < l_n$ for some n. Assume 1 is not a likelihood value. This is a generic property of games. It is not a crucial assumption, but it does simplify the analysis.

Given a group policy $d(i)$ and a likelihood value l_m, define $d(i|l_m)$ to be the probability a group i subject receives decision 1 under $d(i)$ conditional on having an x satisfying $l(x) = l_m$. $d(i|l_m)$ is independent of $\bar{c}(i)$. Given a *likelihood threshold* $l \in [0, l_n]$, define an l-threshold group policy to be a $d(i)$ that satisfies

$$d(i|l_m) = \begin{cases} 1 & \text{if } l_m > \lceil l \rceil \\ \frac{\lceil l \rceil - l}{\lceil l \rceil - \lceil l \rceil^-} & \text{if } l_m = \lceil l \rceil \\ 0 & \text{if } l_m < \lceil l \rceil, \end{cases}$$

where $\lceil l \rceil$ is the smallest likelihood value $\geq l$ and $\lceil l \rceil^-$ is the next smallest likelihood value (or 0, if $\lceil l \rceil = l_1$).

Definition. *A decision policy is a threshold decision policy if both of its group policies are threshold group policies. Two threshold decision policies are equivalent if, for each group, their threshold group policies share the same threshold.*

Given an l-threshold group policy, define $WW(l)$ to be the unique best-response cost threshold of a representative subject. For each likelihood value l_m, $WW(l_m) := \omega \sum_{x_{\mathcal{Y}} \in X_{\mathcal{Y}},\ \frac{p(x_{\mathcal{Y}}|q)}{p(x_{\mathcal{Y}}|u)} > l_m} (p(x_{\mathcal{Y}}|q) - p(x_{\mathcal{Y}}|u))$, and $WW(0) = 0$. WW is a single-peaked (due to 1 not being a likelihood value), piece-wise linear function. Figure 1a depicts an example WW with $n = 3$.

Next, for each likelihood threshold l, define $EE(l)$ to be the set of all cost thresholds for which an l-threshold group policy is a best-response. EE is a weakly decreas-

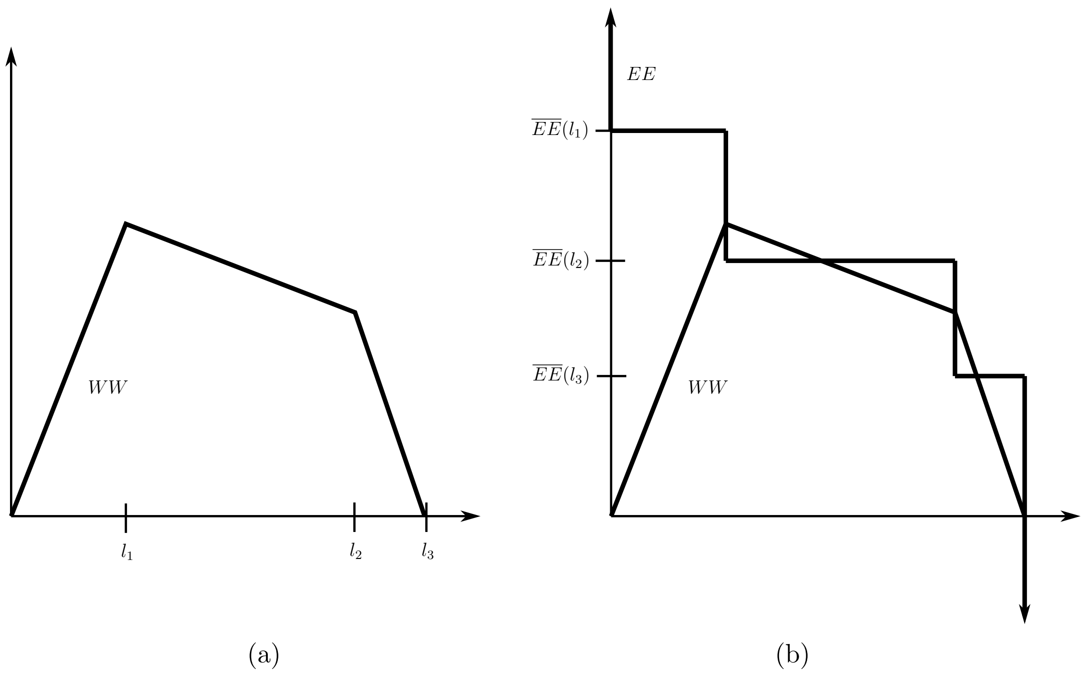

Figure 1

ing correspondence satisfying,

$$EE(l) = \begin{cases} [\overline{EE}(l_1), \infty) & \text{if } l = 0 \\ [\overline{EE}(l_{m+1}), \overline{EE}(l_m)] & \text{if } l = l_m \text{ for some } m \in \{1, 2, \dots n-1\} \\ \overline{EE}(l_m) & \text{if } l \in (l_{m-1}, l_m) \text{ for some } m \in \{1, 2, \dots n\} \\ (-\infty, \overline{EE}(l_n)] & \text{if } l = l_n, \end{cases}$$

where $\overline{EE}(l_m)$ is the unique solution to $G(\overline{EE}(l_m)) = \frac{1}{\frac{v_q}{v_u} \cdot l_m + 1}$. If the representative group i subject chooses $\overline{c}(i) = \overline{EE}(l_m)$, then the DM is indifferent between selecting decision 1 and 0 for any group i subject with other features x satisfying $l(x) = l_m$.

Figure 1b depicts an EE superimposed on a WW. Each intersection, $(l^*, WW(l^*))$, of EE and WW is associated with a set of group equilibria, $\{(\overline{c}^*(i) = WW(l^*), d^*(i))\}$, where $d^*(i)$ is any l^*-threshold group policy. Conversely, each group equilibrium is an element of one such set. The set of equilibria is the set of pairs of group equilibria, one for each group identity. Since EE and WW intersect, the set of equilibria is nonempty. Call two equilibria equivalent if, for each group, their group equilibria are associated with the same intersection of EE and WW. This partitions the set of equilibria into equivalence classes. Two equilibria are equivalent if and only if their threshold decision policies are equivalent.

Definition. *A threshold decision policy is fair if its group thresholds are the same. An equilibrium is fair if its decision policy, which is a threshold decision policy, is fair. Otherwise, it is unfair.*

In a fair equilibrium $(\overline{c}^*, d^*)$, $\overline{c}^*(w) = \overline{c}^*(b)$ and $AR(w, d^*, \mu^*_{RE}) = AR(b, d^*, \mu^*_{RE})$.

3 Ideal Interventions

In this section, I focus on interventions against a rational DM who obtains (μ_{RE}, f_{RE}).

Definition. *An intervention k is ideal if, for any game $(X, \lambda_w, p, G, v_q, v_u, \omega)$, the following two properties are satisfied:*

1. *(Gains to the Discriminated Group.) If $(\overline{c}^*, d^*)$ is an unfair equilibrium, then for all $d \in \mathcal{D}(k, \mu^*_{RE}, f^*_{RE})$,*

$$\left[\min_{i\in I} AR(i, d, \mu^*_{RE}), \max_{i\in I} AR(i, d, \mu^*_{RE})\right] \subsetneq \left[\min_{i\in I} AR(i, d^*, \mu^*_{RE}), \max_{i\in I} AR(i, d^*, \mu^*_{RE})\right].$$

2. *(Steady States Are Fair Equilibria.) Every k equilibrium is a fair equilibrium, and every equivalence class of fair equilibria contains a k equilibrium.*

The definition of an ideal intervention is motivated by Coate and Loury (1993), who write,

> A key question concerning affirmative action is whether the labor-market gains it brings to minorities can continue without it becoming a permanent fixture in the labor market.

One way to formalize their question is by asking, is affirmative action an ideal intervention? It turns out, the answer is no. The authors prove that affirmative action does not satisfy the second property of being ideal, by explicitly constructing affirmative action equilibria that are not fair equilibria. In these steady states, a less qualified group permanently requires the patronage of affirmative action.

Theorem 1. *Equal opportunity and mistaken identity are ideal interventions.*

Proof. See appendix. □

The key step is to show that both interventions, unlike affirmative action, induce the DM to choose a fair decision policy.

In the constrained optimization problem of maximizing the DM's utility subject to delivering fixed true positive rates to each group, the first-order conditions imply that the DM will choose, for each group, a threshold group policy. Equal opportunity then ensures that the thresholds are equal across groups, which means the chosen decision policy is fair.

Fix any group i, the beliefs $f_{RE}(i, x)$ viewed as a function of x are co-monotonic with the likelihood function $l(x)$. This implies any weighted sum of the $f_{RE}(i, x)$'s across groups is also co-monotonic with $l(x)$. Consequently, any group policy that is a threshold function of such a weighted sum is a threshold group policy. Mistaken identity, by requiring a common threshold function be applied across all groups, ensures that the chosen decision policy is fair.

Once we know the DM will select a fair decision policy, the remainder of the proof is the same for equal opportunity and mistaken identity.

In an unfair equilibrium, each group policy is still a threshold group policy, with the favored group having a lower threshold and the discriminated group having a higher threshold. Since the DM is now choosing a single threshold for both groups, utility maximization causes the DM to choose one somewhere in between the low and high thresholds. This implies gains to the discriminated group. Finally, fair decision policies provide equal incentives to invest in becoming qualified across both groups. This implies that steady states are fair equilibria. Theorem 1 follows.[3]

Let k' be any control that is weakly more restrictive that equal opportunity, but allows all fair decision policies when conditioned on a distribution and beliefs, (μ_{RE}, f_{RE}), that are rationalized by some $\bar{c}$. Then the arguments above imply that k' is ideal as well. For example, the equal odds intervention, that requires both true and false positive rates to be equalized, is ideal.

Despite both being ideal interventions, mistaken identity has some significant advantages over equal opportunity. Mistaken identity does not depend on the distribution of features. One obvious implication is that if, for whatever reason, a regulator is unable to obtain μ_{RE}, it can still implement mistaken identity. In fact, a much

[3]In contrast, affirmative action's requirement to equalize acceptance rates causes the DM to push past the equal incentives/thresholds point. The threshold for the previously discriminated group can become so low that the incentive it faces drops below that of the previously favored group again. This leads to the existence of unfair steady states under affirmative action where a less qualified group is permanently patronized.

stronger result is true: The ideal-ness of mistaken identity is robust to the DM obtaining any, possibly incorrect, μ.[4] In addition, the ideal-ness of mistaken identity is also robust to the DM having incorrect, biased beliefs f. In Section 4, I will formally state and prove these results and show that neither robustness property is satisfied by equal opportunity. Finally, unlike equal opportunity, mistaken identity is color-blind. This means that in certain important applications, mistaken identity might actually be legal while equal opportunity is not.

3.1 Color-Blind Data

That mistaken identity is both color-blind and ideal makes it an attractive option for preserving the gains to minorities, in the wake of the Supreme Court striking down affirmative action. However, mistaken identity still conditions on race-conscious beliefs, generated by a machine learning algorithm trained on data that has not been made color-blind. I now show that such race-consciousness is necessary, by proving that ideal interventions do not exist when data is made color-blind. This suggests a need to create a clear space in the law for the use of color-blind interventions that, nevertheless, condition on not color-blind beliefs.

Making data color-blind can be operationalized by modifying the original model in the following way: All decision policies, distributions, and sets of beliefs are now mappings from X instead of mappings from $I \times X$. As an abuse of notation, I continue to refer to these objects by d, μ, and f, respectively. The rational distribution and beliefs (μ_{RE}, f_{RE}) are defined as follows:

$$\begin{aligned} \mu_{RE}(x) &:= \mu_{RE}(w,x) + \mu_{RE}(b,x), \\ f_{RE}(x) &:= \frac{\mu_{RE}(w,x) f_{RE}(w,x) + \mu_{RE}(b,x) f_{RE}(b,x)}{\mu_{RE}(w,x) + \mu_{RE}(b,x)} \qquad \forall x \in X. \end{aligned}$$

Note, $\mu_{RE}(i,x)$ and $f_{RE}(i,x)$ are still well-defined when data is made color-blind. They are just unobserved by the DM. In particular, the acceptance rate, $AR(i, d, \mu_{RE})$, of the i group, defined to be $\frac{\sum_{x\in X} d(x)\mu_{RE}(i,x)}{\sum_{x\in X} \mu_{RE}(i,x)}$, is still computed using $\mu_{RE}(i,x)$, not

[4] Note, just because an intervention does not depend on μ does not necessarily mean that the DM's expected utility maximizing decision policy will not depend on μ. For example, if an intervention forces the DM to pick a single value $\overline{f}$, and only allows the DM to selection decision 0 if $f(i,x) = \overline{f}$, then, clearly, the DM's choice of decision policy will depend on μ. Thus, it is not immediate that any ideal intervention that does not depend on μ, such as mistaken identity, is robust to the DM obtaining any μ.

$\mu_{RE}(x)$. The same is true for the utility of the representative i subject, which, along with the utility of the DM, is now defined as follows:

$$U_i(\overline{c}(i), d) := \omega \sum_{x \in X} \frac{\mu_{RE}(i,x)}{\lambda_i} d(x) - \int_{-\infty}^{\overline{c}(i)} cg(c)dc,$$
$$U_{DM}(d, \mu, f) := \sum_{x \in X} d(x)\mu(x) \left[f(x)v_q - (1 - f(x))v_u\right].$$

Given these changes, the definitions of intervention and k equilibrium supported by (μ, f) are unchanged. A decision policy d is naturally identified with a color-blind decision policy in the original model. Call d fair if the corresponding color-blind decision policy in the original model is a threshold decision policy that is fair. Given this identification, the set of equilibria corresponds to the set of just be color-blind equilibria of the original model. In particular, every fair equilibrium – that is, an equilibrium with a fair decision policy – corresponds to a fair equilibrium in the original model. So, it remains well-defined to talk about equivalence classes of fair equilibria. Given these definitions, the definition of an ideal intervention is unchanged.

Theorem 2. *When data is made color-blind, there do not exist ideal interventions.*

Proof. See appendix. □

To prove Theorem 2, it suffices to find two different games, sharing the same X, such that a fair equilibrium of the first game, $(\overline{c}_1^*, d^*)$, whose equivalence class contains only itself, and an unfair equilibrium of the second game, $(\overline{c}_2^*, d^*)$, share the same decision policy, d^*, the same distribution, $\mu^*_{RE,1} = \mu^*_{RE,2}$, and the same beliefs, $f^*_{RE,1} = f^*_{RE,2}$. Such a pair of fair/unfair equilibria presents an identification problem that puts the two properties of being ideal in conflict with each other: Let k be an ideal intervention. Since $(X, d^*, \mu^*_{RE,1}, f^*_{RE,1})$ arises from a fair equilibrium, whose equivalence class contains only itself, the second property implies $d^* \in k(X, \mu^*_{RE,1}, f^*_{RE,1})$. Since $(X, d^*, \mu^*_{RE,2}, f^*_{RE,2})$ arises from an unfair equilibrium, both properties imply $d^* \notin k(X, \mu^*_{RE,2}, f^*_{RE,2})$. But $k(X, \mu^*_{RE,2}, f^*_{RE,2}) = k(X, \mu^*_{RE,1}, f^*_{RE,1})$. Contradiction.

Many such pairs of fair/unfair equilibria exist. Here, I sketch out an example.

Suppose $X = X_1 \times X_2 = \{A, B, C\} \times \{W(\text{onderful}), B(\text{ad})\}$. Interpret X_1 as a letter grade and X_2 as a reference letter by an evaluator. In game 1, let us assume the statistical model captures an evaluator trying to describe intangible qualities of the subject that are not reflected by a letter grade. Both X_1 and X_2 depend only on a subject's class. Since there are no proxies for group identity and decision policies are color-blind, any equilibrium is fair. Consider the following beliefs f^*:

	W	B
A	3/4	1/4
B	2/3	1/5
C	1/3	1/6

f^* reflects the plausible idea that, although higher letter grades are better, the evaluator's letter is still the most important signal of qualification. Now suppose $v_q = v_u$, so that the DM strictly prefers to select decision 1 (0) on any x subject with $f^*(x) > \frac{1}{2}$ ($< \frac{1}{2}$). Then given f^*, the unique best-response of the DM is to choose the decision policy, d^*, that selects decision 1 if and only if $x_1 \in \{A, B\}$ and $x_2 = W$. One can easily imagine how game 1 could be parameterized so that, given the best-response $\overline{c}_1^*$ to d^*, the rational beliefs are precisely f^*: $f^*_{RE,1}(x) = f^*(x)$ for all $x \in X$. In such a game, $(\overline{c}_1^*, d^*)$ is a fair equilibrium, and it is obvious there are no other equivalent equilibria. Similarly, the rational beliefs for group i in the original model coincide with f^*, for both i: $f^*_{RE,1}(i, x) = f^*(x)$ for all $(i, x) \in I \times X$.

In game 2, let us assume the statistical model captures an evaluator who is biased against the b group, and uses "Wonderful" and "Bad" as code words for group identities w and b, respectively. In this case, X_2 is a proxy for group identity, and the decision policy, d^*, that was fair in game 1, is now unfair, since it uses the proxy X_2 to discriminate against b subjects. Now if the DM chooses d^*, the b group has low incentive to invest and best-responds with a low $\overline{c}_2^*(b)$, while the w group has high incentive to invest and best-responds with a high $\overline{c}_2^*(w)$. Given this unequal $\overline{c}_2^*$, the DM rationally assigns, for each letter grade, a lower belief to any subject described as bad by the evaluator, just like in f^*. Assuming the DM's rational beliefs are f^*, and $v_q = v_u$, then the best-response decision policy is d^*. Again, one can easily imagine how game 2 could be parameterized to make this story true. In such a game, $(\overline{c}_2^*, d^*)$ is an unfair equilibrium. Unlike in game 1, the rational beliefs for group i in the original model do not coincide with f^*, for either i. For the b group, $f^*_{RE,2}(b, x) < f^*(x)$ for all $x \in X$, while the opposite is true for the w group. Also, unlike f^*, $f^*_{RE,2}(i, \cdot)$ is constant over X_2 for both groups i.

Games 1 and 2 share the same X, and the pair of equilibria I just constructed share the same decision policy d^*, and the same beliefs f^*. A technical lemma shows that there are enough degrees of freedom to parameterize the two games so that the two equilibria's distributions, $\mu^*_{RE,1}$ and $\mu^*_{RE,2}$, are the same as well. Theorem 2 follows.

4 Incorrect Distributions and Beliefs

In the previous section, I focused on when the DM is rational and obtains (μ_{RE}, f_{RE}). In this section, I explore the performance of ideal interventions when the DM might have an incorrect (μ, f). To motivate this exploration, let us begin by constructing a new ideal intervention that highlights the pitfalls of assuming the DM has perfectly rational expectations.

Given X and beliefs f defined over $I \times X$, a subset, $\hat{\mathcal{Y}} \subset \{1, 2, \ldots N\}$, of the other features indices has the property that f depends on $I \times X$ only up to $I \times X_{\hat{\mathcal{Y}}}$ if

$$x_{\hat{\mathcal{Y}}} = x'_{\hat{\mathcal{Y}}} \Rightarrow f(i, x) = f(i, x') \ \ \forall x, x' \in X, i \in I.$$

Let $\hat{\mathcal{Y}}_1$ and $\hat{\mathcal{Y}}_2$ be two such subsets, then $\hat{\mathcal{Y}}_1 \cap \hat{\mathcal{Y}}_2$ also has the property: Let x, x' satisfy $x_{\hat{\mathcal{Y}}_1 \cap \hat{\mathcal{Y}}_2} = x'_{\hat{\mathcal{Y}}_1 \cap \hat{\mathcal{Y}}_2}$. Let x'' satisfy $x''_{\hat{\mathcal{Y}}_1} = x_{\hat{\mathcal{Y}}_1}$ and $x''_{\hat{\mathcal{Y}}_2} = x'_{\hat{\mathcal{Y}}_2}$. Then $f(i, x) = f(i, x'') = f(i, x')$. Thus, there exists a unique, possibly empty, minimal subset, $\hat{\mathcal{Y}}(f)$, with the property.

Example. *The no proxies intervention k specifies, for each (X, μ, f),*

$$k(X, \mu, f) = \left\{ d : I \times X \to [0, 1] \ \middle|\ d \text{ depends on } I \times X \text{ only up to } X_{\hat{\mathcal{Y}}(f)} \right\}.$$

No proxies tries to identify and exclude all other features that are proxies for group identity. Like mistaken identity, it is color-blind, does not depend on μ, and is ideal. However, it is intuitive that arbitrarily small deviations from rationality can cause no proxies to fail to exclude any other features for being a group identity proxy. I now make precise in what sense the ideal-ness of no proxies is not robust to small deviations from rationality.

Definition. *Given an intervention k, a game $(X, \lambda_w, p, G, v_q, v_u, \omega)$, and an $\varepsilon > 0$, a strategy profile $(\overline{c}_{k,\varepsilon}, d_{k,\varepsilon})$ is a k ε-equilibrium if it is a k equilibrium supported by some $(\mu_{k,\varepsilon}, f_{k,\varepsilon})$ satisfying $\|(\mu_{k,\varepsilon}, f_{k,\varepsilon}) - (\mu_{RE,k,\varepsilon}, f_{RE,k,\varepsilon})\|_2 \leq \varepsilon$.*

Definition. *An ideal intervention k is robust to small deviations from rationality if, for any game $(X, \lambda_w, p, G, v_q, v_u, \omega)$ and any $\delta > 0$, there exists an $\varepsilon > 0$ such that the following two properties are satisfied:*

1. *If* $(\bar{c}^*, d^*)$ *is an unfair equilibrium and* $\|(\mu^*, f^*) - (\mu^*_{RE}, f^*_{RE})\|_2 \leq \varepsilon$*, then for every* $d \in \mathcal{D}(k, \mu^*, f^*)$*, there exists a* $\hat{d} \in \mathcal{D}(k, \mu^*_{RE}, f^*_{RE})$ *satisfying* $\|d - \hat{d}\|_2 \leq \delta$*.*

2. *If* $(\bar{c}_{k,\varepsilon}, d_{k,\varepsilon})$ *is a* k ε*-equilibrium, then there exists a* k *equilibrium* $(\bar{c}_k, d_k)$ *satisfying* $\|(\bar{c}_{k,\varepsilon}, d_{k,\varepsilon}) - (\bar{c}_k, d_k)\|_2 \leq \delta$*.*

Let k be the no proxies intervention and let us reconsider game 2 described in Section 3.1. Recall, when data is made color-blind, $(\bar{c}^*_2, d^*)$ is an unfair equilibrium of the game. This implies $(\bar{c}^*_2, d^*)$, viewed as a strategy profile in the original model, is a just be color-blind equilibrium featuring a color-blind decision policy d^* that is not fair. Consequently, just be color-blind is not ideal. Note, $(\bar{c}^*_2, d^*)$ is not a no proxies equilibrium because $d^* \notin k(X, \mu^*_{RE,2}, f^*_{RE,2})$, which follows from the fact that d^* depends on X_2 but $f^*_{RE,2}(i, \cdot)$ is constant over X_2 for all i. However, for any $\varepsilon > 0$, it is possible to perturb $f^*_{RE,2}$ to some f^*_2 so that

1. $\|(\mu^*_{RE,2}, f^*_2) - (\mu^*_{RE,2}, f^*_{RE,2})\|_2 \leq \varepsilon$,

2. d^* remains the best-response color-blind decision policy,

3. f^*_2 is injective.

Property 3 implies that, under beliefs f^*_2, no feature except group identity is excluded by no proxies, in which case $k(X, \mu^*_{RE,2}, f^*_2)$ is the set of all color-blind decision policies and $d^* \in k(X, \mu^*_{RE,2}, f^*_2)$. Properties 1 and 2 then imply that $(\bar{c}^*_2, d^*)$ is a no proxies ε-equilibrium. Since $(\bar{c}^*_2, d^*)$ is not a fair equilibrium, this implies

Proposition 1. *No proxies is not robust to small deviations from rationality.*

One may try to remedy the fragility of no proxies by strengthening the definition of $\hat{\mathcal{Y}}(f)$ to exclude any other features over which f is "almost constant," rather than just "constant." The problem with such an approach is that, without knowing the parameters of the model, it is impossible to know what constitutes almost constant.

Unlike no proxies, equal opportunity – viewed as a correspondence from (μ, f) to decision policies holding X fixed – is continuous. Berge's Maximum Theorem implies

Theorem 3. *Continuous ideal interventions are robust to small deviations from rationality.*

Proof. See appendix. □

Mistaken identity is not continuous, an artifact of the discreteness of the space X. Recall, at a rational set of beliefs, any allowed decision policy under mistaken identity is fair, meaning both group policies are threshold group policies with the same threshold. However, at a small perturbation of rational beliefs, the most one can guarantee about any allowed decision policy is that, each group policy has a threshold l_i, and if $l_w \neq l_b$, then there is still some likelihood value l_m, such that $l_w, l_b \in (l_{m-1}, l_m)$. Thus, even though mistaken identity is not continuous, when rational beliefs are perturbed, the set of allowed decision policies increases by, at worst, a qualitatively small amount. Consequently, there is a slightly weaker sense in which mistaken identity is robust to small deviations from rationality:

Define two threshold group policies to be almost equal if there exists a likelihood value l_m such that both thresholds are elements of $[l_{m-1}, l_m]$. Define two threshold decision policies, d and $\hat{d}$, to be almost equal – denoted by $d \approx \hat{d}$ – if, for each group identity, their threshold group policies are almost equal. One can show that, under mistaken identity, for any game $(X, \lambda_w, p, G, v_q, v_u, \omega)$, there exists an $\varepsilon > 0$ such that

1. If $(\overline{c}^*, d^*)$ is an unfair equilibrium and $\|(\mu^*, f^*) - (\mu^*_{RE}, f^*_{RE})\|_2 \leq \varepsilon$, then for every $d \in \mathcal{D}(k, \mu^*, f^*)$, there exists a $\hat{d} \in \mathcal{D}(k, \mu^*_{RE}, f^*_{RE})$ satisfying $d \approx \hat{d}$.

2. If $(\overline{c}_{k,\varepsilon}, d_{k,\varepsilon})$ is a k ε-equilibrium, then there exists a k equilibrium $(\overline{c}_k, d_k)$ satisfying $d_{k,\varepsilon} \approx d_k$.

When two threshold decision policies d and $\hat{d}$ are almost equal, it means that, at worst, they differ, for each group i, only on subjects associated with a single likelihood value. Depending on the application, it can be reasonable to interpret this as a minor difference. For example, suppose X is the set of credit scores between 300 and 850, and higher scores mean higher likelihood values. Differing only on group i subjects associated with a single likelihood value – say, $l(750)$ – means the only difference between d and $\hat{d}$ is in the probability of decision 1 for the tiny fraction of group i subjects that score exactly 750. Moreover, even if we are in a game where each piece of data is binary (i.e. $|X_n| = 2$ for all $n \in \{1, 2, \dots N\}$), as long as there are many pieces of class related data (i.e. $|\mathcal{Y}|$ is large), there will typically be many likelihood values, and, again, the difference between d and $\hat{d}$ can be interpreted as minor.

4.1 Any Distribution μ and Biased Beliefs f

Recent work has highlighted the prevalence of statistical discrimination driven by incorrect, biased beliefs (Bohren et al., 2019, 2023). Motivated by these findings, I now move beyond small deviations from rationality in two ways. I impose no restriction on what kind of distribution, μ, the DM obtains, and I allow the DM's beliefs, f, to exhibit a wide array of biases toward the discriminated group and/or the favored group.

A *bias* $B = (B_w, B_b) : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous function satisfying $B(c, c) = (c, c)$ for all $c \in \mathbb{R}$. Given $\overline{c}$, the biased beliefs f_B are defined to be,

$$f_B(i, x|\overline{c}) := f_{RE}(i, x|B_i(\overline{c})) \quad \forall (i, x) \in I \times X.$$

For example, let β be some weakly increasing continuous function with $\beta(0) = 0$. Then under a bias of the form $B(\overline{c}) := (\overline{c}(w) + \beta(\overline{c}(w) - \overline{c}(b)), \overline{c}(b) + \beta(\overline{c}(b) - \overline{c}(w)))$ for all $\overline{c} \in \mathbb{R}^2$, the DM weakly overestimates the qualification of the favored group and weakly underestimates the qualification of the discriminated group. As β becomes steep, such an f_B represents significant potential for bias by the DM. From now on call such an f_B a polarizing bias with parameter β. Note, the condition $B(c, c) = (c, c)$ implies that when both groups are equally qualified, biased beliefs coincide with rational beliefs. This is needed to ensure that fair equilibria exist.

Let $(\overline{c}^*, d^*)$ be a fair equilibrium. Then it is also an equilibrium supported by (μ, f_B^*) for any μ and B. Let $(\overline{c}^*, d^*)$ be an equilibrium supported by (μ, f_B^*) for some μ and B. Call $(\overline{c}^*, d^*)$ unfair if it is not a fair equilibrium. See Figure 2 for two examples of unfair equilibria $(\overline{c}^*, d^*)$ supported by some (μ, f_B^*).

Definition. *An ideal intervention k is robust to any distribution and biased beliefs if, for any game $(X, \lambda_w, p, G, v_q, v_u, \omega)$, distribution μ, and bias B, the following two properties are satisfied:*

1. *(Gains to the Discriminated Group.) If $(\overline{c}^*, d^*)$ is an unfair equilibrium supported by (μ, f_B^*), then for all $d \in \mathcal{D}(k, \mu, f_B^*)$,*

$$\left[\min_{i \in I} AR(i, d, \mu_{RE}^*), \max_{i \in I} AR(i, d, \mu_{RE}^*)\right] \subsetneq \left[\min_{i \in I} AR(i, d^*, \mu_{RE}^*), \max_{i \in I} AR(i, d^*, \mu_{RE}^*)\right].$$

2. *(Steady States Are Fair Equilibria.) Every k equilibrium $(\overline{c}^*, d^*)$ supported by (μ, f_B^*) is a fair equilibrium, and every equivalence class of fair equilibria contains a k equilibrium $(\overline{c}^*, d^*)$ supported by (μ, f_B^*).*

Notice, the *Gains to the Discriminated Group* property is defined using μ_{RE}^*, not μ. This means when an ideal intervention that is robust to any distribution and

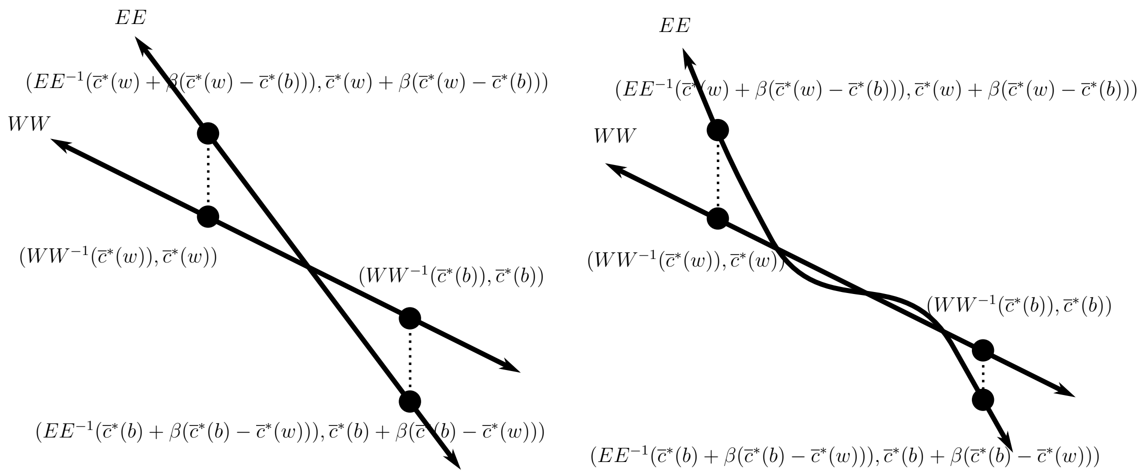

Figure 2: Two examples of unfair equilibria $(\overline{c}^*, d^*)$ supported by some (μ, f_B^*) where B is a polarizing bias with parameter β. For both i, we have $EE^{-1}(\overline{c}^*(i) + \beta(\overline{c}^*(i) - \overline{c}^*(-i))) = WW^{-1}(\overline{c}^*(i))$, and $d^*(i)$ is a $WW^{-1}(\overline{c}^*(i))$-threshold group policy. For conceptual clarity, EE has been drawn as a strictly decreasing function.

biased beliefs is imposed, even though the DM thinks the distribution is μ, one can still be confident that *real* gains will be made by the discriminated group.

If we weaken the definition above, by deleting the term ", distribution μ," and replacing all other instances of μ with μ_{RE}^*, then we have a natural definition of an ideal intervention being robust to biased beliefs, but not necessarily to any distribution. Similarly, there is a natural definition of an ideal intervention being robust to any distribution, but not necessarily to biased beliefs.

Theorem 4. *Mistaken identity is robust to any distribution and biased beliefs, while equal opportunity is neither robust to any distribution nor robust to biased beliefs.*

Proof. See appendix. □

Fix any group i. Even with a bias B, the DM's beliefs, $f_B(i, x)$, viewed as a function of x, are still co-monotonic with the likelihood function $l(x)$, just like $f_{RE}(i, x)$. As a result, the argument for why mistaken identity forces the DM to choose a fair decision policy under f_{RE} extends to the case of biased beliefs of the form f_B. The remainder of the proof that mistaken identity is ideal extends to a proof that mistaken identity is robust to any distribution and biased beliefs.

On the other hand, it is obvious equal opportunity is not robust to any distribution. The intuition for why equal opportunity is not robust to biased beliefs does not depend on the possibility of the DM having a strange bias. It suffices to restrict attention only to polarizing biases, where the DM weakly overestimates the qualification of the favored group and weakly underestimates the qualification of the discriminated group.

Consider an unfair equilibrium $(\overline{c}^*, d^*)$ with $\overline{c}^*(w) > \overline{c}^*(b)$. Then the *rational* true positive rate of the w group is higher: $TP(w, d^*, \mu^*_{RE}, f^*_{RE}) > TP(b, d^*, \mu^*_{RE}, f^*_{RE})$. Let us now examine the expression for the rational true positive rate of the w group,

$$\frac{\sum_{x \in X} d^*(w,x)\mu^*_{RE}(w,x)f^*_{RE}(w,x)}{\sum_{x \in X} \mu^*_{RE}(w,x)f^*_{RE}(w,x)},$$

and see how it is affected by the introduction of a polarizing bias B. Introducing such a bias weakly increases the beliefs about the w group: $f^*_B(w,x) \geq f^*_{RE}(w,x)$ for all $x \in X$. If all of these beliefs increase by the same percentage, then replacing $f^*_{RE}(w,x)$ with $f^*_B(w,x)$ in the fraction above causes the true positive rate to remain unchanged. However, in general, different beliefs will increase by different percentages, and, depending on which ones increase the most, the true positive rate of the w group can increase or decrease. The same is true for the b group. In particular, it is possible for a polarizing bias to simultaneously decrease the true positive rate of the w group and increase the true positive rate of the b group, reducing the true positive rate gap – possibly down to zero. Using this insight, it is possible to construct equal opportunity equilibria supported by polarizing biased beliefs, in which the decision policy is a threshold decision policy that is not fair, thereby violating the second property of being robust to biased beliefs.

5 Conclusion

This article studies how to intervene against statistical discrimination by a DM, whose beliefs are generated by machine learning, rather than by humans. Unlike beliefs formed by a human mind, machine learning-generated beliefs are verifiable. This allows interventions to condition on the DM's beliefs. Such belief-contingent interventions can perform well where affirmative action does not. Specifically, I look for ideal interventions that bring gains to the discriminated group in all unfair equilibria and whose steady states are fair equilibria. The belief-free affirmative action intervention is not ideal because it has unfair steady states. However, I find two belief-contingent interventions – equal opportunity and mistaken identity – that are ideal.

In addition, mistaken identity is color-blind and robust to incorrect, biased beliefs. This makes mistaken identity an attractive option for preserving the gains to minorities, in the wake of the Supreme Court striking down affirmative action. However, mistaken identity does condition on race-conscious beliefs, generated by a machine learning algorithm trained on data that has not been made color-blind. I show that such race-consciousness is necessary, by proving that ideal interventions do not exist when data is made color-blind. This suggests a need to create a clear space in the law for the use of color-blind interventions that, nevertheless, condition on not color-blind beliefs.

6 Appendix

6.1 Proof of Theorem 1

Let $\mathcal{F}$ denote the set of fair decision policies.

Lemma 1. *Let k be equal opportunity. Given any cost threshold pair $\overline{c}$, $\mathcal{D}(k, \mu_{RE}, f_{RE}) \subset \mathcal{F} \subset k(X, \mu_{RE}, f_{RE})$.*

Proof. I first prove $\mathcal{D}(k, \mu_{RE}, f_{RE}) \subset \mathcal{F}$. Let $d \in \mathcal{D}(k, \mu_{RE}, f_{RE})$. Define

$$TP(d) := TP(w, d, \mu_{RE}, f_{RE}) = TP(b, d, \mu_{RE}, f_{RE}).$$

That $d \in \mathcal{D}(k, \mu_{RE}, f_{RE})$ means that the DM cannot be made strictly better off by swapping out $d(i)$ with a different $\hat{d}(i)$ with the same true positive rate. Thus, for each $i \in I$, $\{d(i,x)\}_{x \in X}$ is an element of

$$\begin{aligned} \underset{\{d_x\}_{x\in X}\ s.t.\ \frac{\sum_{x\in X} d_x \mu_{RE}(i,x) f_{RE}(i,x)}{\sum_{x\in X} \mu_{RE}(i,x) f_{RE}(i,x)} = TP(d)}{\arg\max} & \sum_{x \in X} d_x \mu_{RE}(i,x)\left[f_{RE}(i,x) v_q - (1 - f_{RE}(i,x)) v_u\right] \\ s.t. \quad & d_x \geq 0 \quad \forall x \in X \\ & d_x \leq 1 \quad \forall x \in X. \end{aligned}$$

The Kuhn-Tucker conditions imply there exists a $\gamma_i^* \geq 0$ such that, for each $x \in X$,

$$d(i,x) = \begin{cases} 0 & \text{if } \frac{\mu_{RE}(i,x)[f_{RE}(i,x)v_q - (1-f_{RE}(i,x))v_u]}{\frac{\mu_{RE}(i,x)f_{RE}(i,x)}{\sum_{x' \in X} \mu_{RE}(i,x')f_{RE}(i,x')}} < \gamma_i^* \\ 1 & \text{if } \frac{\mu_{RE}(i,x)[f_{RE}(i,x)v_q - (1-f_{RE}(i,x))v_u]}{\frac{\mu_{RE}(i,x)f_{RE}(i,x)}{\sum_{x' \in X} \mu_{RE}(i,x')f_{RE}(i,x')}} > \gamma_i^*. \end{cases}$$

Simplifying the fraction appearing in the first order condition yields

$$\begin{aligned} &\frac{\mu_{RE}(i,x)\left[f_{RE}(i,x)v_q - (1-f_{RE}(i,x))v_u\right]}{\frac{\mu_{RE}(i,x)f_{RE}(i,x)}{\sum_{x' \in X} \mu_{RE}(i,x')f_{RE}(i,x')}} \\ &= \sum_{x' \in X} \mu_{RE}(i,x')f_{RE}(i,x') \left[v_q - \frac{1-f_{RE}(i,x)}{f_{RE}(i,x)} v_u\right] \\ &= \sum_{x' \in X} \mu_{RE}(i,x')f_{RE}(i,x') \left[v_q - \frac{(1-G(\bar{c}(i)))}{G(\bar{c}(i))l(x)} v_u\right]. \end{aligned}$$

This means there exists a likelihood value l_{m_i} such that

$$d(i,x) = \begin{cases} 0 & \text{if } l(x) < l_{m_i} \\ 1 & \text{if } l(x) > l_{m_i}. \end{cases}$$

This implies that $d(i)$ is an l-threshold group policy where $l = d(i|l_{m_i})l_{m_i-1} + (1 - d(i|l_{m_i}))l_{m_i}$, where $l_0 := 0$. l_{m_i} can be chosen so that either $l = 0$ or $d(i|l_{m_i}) \in [0,1)$. With this characterization of $d(i)$, we can express the shared true positive rate as

$$\begin{aligned} TP(d) &= \frac{\sum_{x \in X} d(i,x)\mu_{RE}(i,x)f_{RE}(i,x)}{\sum_{x \in X} \mu_{RE}(i,x)f_{RE}(i,x)} \\ &= \frac{\sum_{x \in X, l(x)=l_{m_i}} \hat{d}(i,x)\mu_{RE}(i,x)f_{RE}(i,x)}{\sum_{x \in X} \mu_{RE}(i,x)f_{RE}(i,x)} + \frac{\sum_{x \in X, l(x)>l_{m_i}} \mu_{RE}(i,x)f_{RE}(i,x)}{\sum_{x \in X} \mu_{RE}(i,x)f_{RE}(i,x)} \\ &= \sum_{x \in X, l(x)=l_{m_i}} d(i,x)p(x|i,q) + \sum_{x \in X, l(x)>l_{m_i}} p(x|i,q) \\ &= d(i|l_{m_i}) \sum_{x \in X, l(x)=l_{m_i}} p(x|i,q) + \sum_{m \in \{m_i+1, m_i+2, \ldots n\}} \sum_{x \in X, l(x)=l_m} p(x|i,q) \end{aligned}$$

$$= d(i|l_{m_i}) \sum_{x\in X, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_{m_i}} p(x_\mathcal{Y}|q)p(x_{-\mathcal{Y}}|i,x_\mathcal{Y})+$$

$$\sum_{m\in\{m_i+1,m_i+2,\dots n\}} \sum_{x\in X, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_m} p(x_\mathcal{Y}|q)p(x_{-\mathcal{Y}}|i,x_\mathcal{Y})$$

$$= d(i|l_{m_i}) \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_{m_i}} \sum_{x_{-\mathcal{Y}}\in X_{-\mathcal{Y}}} p(x_\mathcal{Y}|q)p(x_{-\mathcal{Y}}|i,x_\mathcal{Y})+$$

$$\sum_{m\in\{m_i+1,m_i+2,\dots n\}} \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_m} \sum_{x_{-\mathcal{Y}}\in X_{-\mathcal{Y}}} p(x_\mathcal{Y}|q)p(x_{-\mathcal{Y}}|i,x_\mathcal{Y})$$

$$= d(i|l_{m_i}) \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_{m_i}} p(x_\mathcal{Y}|q) + \sum_{m\in\{m_i+1,m_i+2,\dots n\}} \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_m} p(x_\mathcal{Y}|q).$$

If $m_w \lesseqgtr m_b$, then clearly the true positive rate of group $w \gtreqless b$ the true positive rate of group b. So $m_w = m_b$. Now equality of true positive rates implies

$$\hat{d}(w|l_{m_w}) \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_{m_w}} p(x_\mathcal{Y}|q) = \hat{d}(b|l_{m_b}) \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=l_{m_b}} p(x_\mathcal{Y}|q),$$

which implies $d(w|l_{m_w}) = d(b|l_{m_b})$, which implies $d \in \mathcal{F}$.

Next, let $d \in \mathcal{F}$ with some threshold l. Previous calculations imply, for each i,

$$TP(i, d, \mu_{RE}, f_{RE}) = \frac{\lceil l\rceil - l}{\lceil l\rceil - \lceil l\rceil^-} \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}=\lceil l\rceil} p(x_\mathcal{Y}|q) + \sum_{x_\mathcal{Y}\in X_\mathcal{Y}, \frac{p(x_\mathcal{Y}|q)}{p(x_\mathcal{Y}|u)}>\lceil l\rceil} p(x_\mathcal{Y}|q),$$

which does not depend on i. So $\mathcal{F} \subset k(X, \mu_{RE}, f_{RE})$. □

Given $\bar{c}$, $f_{RE}(i,x)$ is co-monotonic with $l(x)$, meaning

$$f_{RE}(i,x) > f_{RE}(i,x') \Leftrightarrow l(x) > l(x') \qquad \forall x, x' \in X.$$

This means that $\alpha_w f(w,x) + \alpha_b f(b,x)$ is co-monotonic with $l(x)$ for arbitrary group identity weights $\alpha_w, \alpha_b \geq 0$, not both zero. Letting k be mistaken identity, we may now conclude that $k(X, \mu_{RE}, f_{RE}) \subset \mathcal{F}$, and, for any $d \in \mathcal{F}$, there is an equivalent $d' \in k(X, \mu_{RE}, f_{RE})$.

I have now shown that any best-response of a rational DM given any $\bar{c}$, subject to equal opportunity or mistaken identity, must be fair. I am now ready to prove, at once, that equal opportunity and mistaken identity are ideal. So, let k denote equal opportunity or mistaken identity.

First, consider an unfair equilibrium $(\overline{c}^*, d^*)$. Let l_i^* be the likelihood threshold associated with $d^*(i)$. Without loss of generality, assume $l_w^* < l_b^*$. Let $d \in \mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$. Since d is fair, both group policies are associated with the same likelihood threshold – call it l.

There exist integers $m_w \in \{0, 1, \dots n-1\}$ and $m_b \in \{1, 2, \dots n\}$ such that $l_w^* \in [l_{m_w}, l_{m_w+1})$, $l_b^* \in (l_{m_b-1}, l_{m_b}]$, and $m_w < m_b$. Recall the convention, $l_0 := 0$. If $l > l_{m_b}$, then the DM is strictly better off deviating to a fair decision policy with threshold l_{m_b}. By Lemma 1 and the characterization of the allowed decision policies under mistaken identity, such a deviation is feasible. It makes the DM strictly better off: All subjects that previously received decision 1 still receive decision 1, the additional w-subjects that receive decision 1 give the DM positive expected utility, and the additional b-subjects that receive decision 1 give the DM nonnegative expected utility. Thus, $l \leq l_{m_b}$.

If $l_b^* < l_{m_b}$, then $l_w^* \leq l_{m_b-1}$. This is because when 1 is not a likelihood value, it is impossible for $l_w^*, l_b^* \in (l_{m_b-1}, l_{m_b})$. It is the only place where the minor assumption that 1 is not a likelihood value is needed. In this case, if $l \in (l_{m_b-1}, l_{m_b}]$, then the DM is strictly better off deviating to the fair decision policy with threshold l_{m_b-1}. Under this deviation, all subjects that previously received decision 1 still receive decision 1, the additional w-subjects that receive decision 1 give the DM positive expected utility, and the additional b-subjects that receive decision 1 give the DM zero expected utility. Together, the results of this paragraph and the previous one imply $l \leq l_b^*$ and $AR(b, d^*, \mu_{RE}^*) \leq AR(b, d, \mu_{RE}^*)$.

By a symmetric argument $l \geq l_w^*$ and $AR(w, d, \mu_{RE}^*) \leq AR(w, d^*, \mu_{RE}^*)$. Since $l_w^* < l_b^*$, so $\overline{c}^*(w) \geq \overline{c}^*(b)$. And now, since d is fair, we have $AR(b, d, \mu_{RE}^*) \leq AR(w, d, \mu_{RE}^*)$. Finally, since $l_w^* < l_b^*$, it must be the case that $l < l_b^*$ – in which case, $AR(b, d^*, \mu_{RE}^*) < AR(b, d, \mu_{RE}^*)$, or $l > l_w^*$ – in which case, $AR(w, d, \mu_{RE}^*) < AR(w, d^*, \mu_{RE}^*)$. This proves the Gains to the Discriminated Group property.

Next, let $(\overline{c}^*, d^*)$ be a k equilibrium. Then d^* is fair, and so $\overline{c}^*(w) = \overline{c}^*(b)$. Since $\overline{c}^*(w) = \overline{c}^*(b)$, the set

$$\mathcal{D}(\mu_{RE}^*, f_{RE}^*) := \underset{d \in [0,1]^{|I \times X|}}{\arg\max}\, U_{DM}(d, \mu_{RE}^*, f_{RE}^*)$$

contains fair threshold decision policies. Let $d \in \mathcal{D}(\mu_{RE}^*, f_{RE}^*) \cap \mathcal{F}$. By Lemma

1 and the characterization of the allowed decision policies under mistaken identity, there exists a $\hat{d} \in \mathcal{D}(k, \mu^*_{RE}, f^*_{RE})$ equivalent to d. Since $d^* \in \mathcal{D}(k, \mu^*_{RE}, f^*_{RE})$, $U_{DM}(d^*, \mu^*_{RE}, f^*_{RE}) = U_{DM}(\hat{d}, \mu^*_{RE}, f^*_{RE})$. Since $\hat{d}$ and d are equivalent, $U_{DM}(\hat{d}, \mu^*_{RE}, f^*_{RE}) = U_{DM}(d, \mu^*_{RE}, f^*_{RE})$. This implies $d^* \in \mathcal{D}(\mu^*_{RE}, f^*_{RE})$, which then implies $(\bar{c}^*, d^*)$ is a fair equilibrium. Clearly, every fair equilibrium is equivalent to a k equilibrium. This proves the Steady States are Fair Equilibria property.

6.2 Proof of Theorem 2

The proof is by example. Following the argument outlined in Section 3.1, I will find two different games, sharing the same X, such that a fair equilibrium of the first game, $(\bar{c}^*_1, d^*)$, whose equivalence class contains only itself, and an unfair equilibrium of the second game, $(\bar{c}^*_2, d^*)$, share the same decision policy d^*, the same distribution $\mu^*_{RE,1} = \mu^*_{RE,2}$, and the same beliefs $f^*_{RE,1} = f^*_{RE,2}$.

Fix a cost distribution with G and g satisfying $G(0) = 1 - G(\frac{2}{3}) = \frac{1}{6}$ and $g(c) = 1$ for all $c \in [0, \frac{2}{3}]$, and consider the following family of games parameterized by constants $\gamma \in (0, \frac{1}{6})$ and $\delta \in (0, 1)$:

- $X = X_1 \times X_2 = \{A, B, C\} \times \{W, B\}$,
- $\lambda_w = \frac{1}{2}$,
- $p(x|i, y) = p(x_1|y)p(x_2|i)$ where
  - $p(x_1|q) = \frac{1}{3}$ for all $x_1 \in X_1$,
  - $p(C|u) = \frac{2}{3}$, $p(B|u) = \frac{1}{6} + \gamma$, $p(A|u) = \frac{1}{6} - \gamma$,
  - $p(W|w) = p(B|b) = 1 - \delta$,
- $v_q = v_u = \omega = 1$.

For any game in this family, EE and WW intersect five times as in Figure 3.

To begin, consider the limit "game" where $\gamma = \delta = 0$. Technically speaking, this is not a game because $\delta = 0$ implies the full support condition of Assumption 1 is violated.

Let $\bar{c}^*(w) = \frac{1}{3}$ and $\bar{c}^*(b) = 0$. The table for (μ^*_{RE}, f^*_{RE}) is:

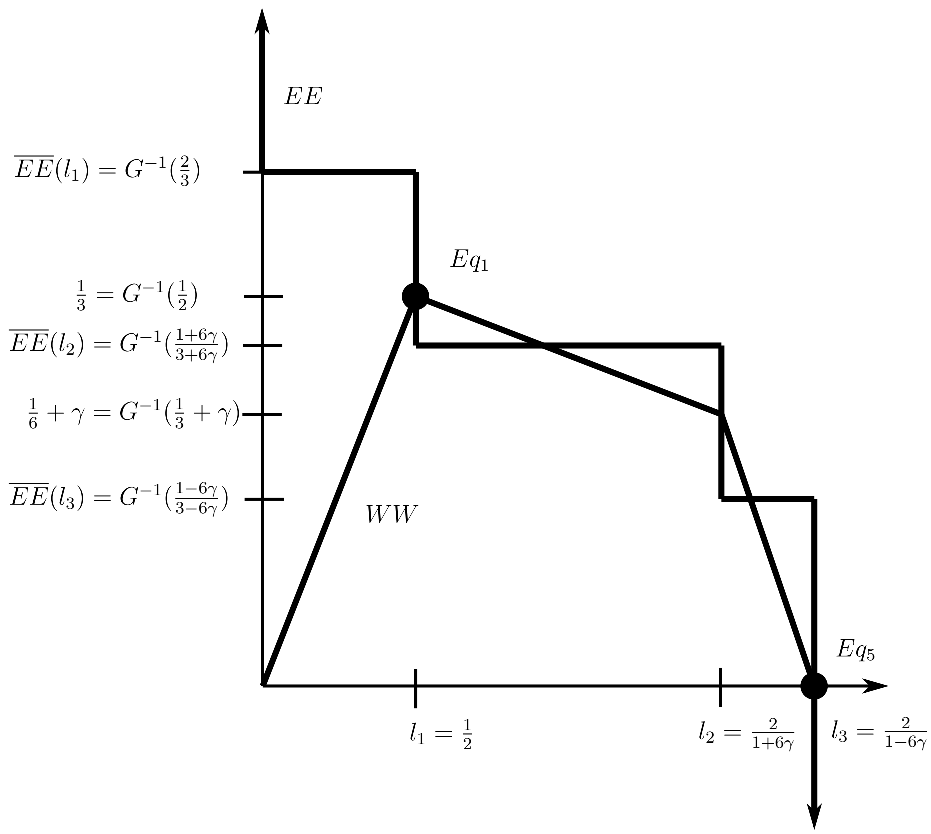

Figure 3

	W		B	
A	1/8	2/3	7/72	2/7
B	1/8	2/3	7/72	2/7
C	1/4	1/3	22/72	1/11

In particular, $f^*_{RE}(x) > \frac{v_u}{v_q+v_u} = \frac{1}{2}$ for $x \in \{(A,W),(B,W)\}$ and $f^*_{RE}(x) < \frac{1}{2}$ otherwise. Define the decision policy d^* where $d^*(x) = 1$ if $x \in \{(A,W),(B,W)\}$ and $d^*(x) = 0$ otherwise. Figure 3 implies $\overline{c}^*$ is the best response to d^*.

Continuity now implies, for all sufficiently small $\gamma, \delta > 0$, the best-response $\overline{c}^*_{\gamma,\delta}$ to d^* in the (γ,δ)-game is sufficiently close to $\overline{c}^*$ so that $f^*_{RE,\gamma,\delta}(x) > \frac{1}{2}$ for $x \in \{(A,W),(B,W)\}$ and $f^*_{RE,\gamma,\delta}(x) < \frac{1}{2}$ otherwise. This implies that in these (γ,δ)-games, $(\overline{c}^*_{\gamma,\delta}, d^*)$ is an equilibrium.

Clearly, $(\overline{c}^*_{\gamma,\delta}, d^*)$ is unfair. In fact, holding γ fixed, as δ goes to zero, $(\overline{c}^*_{\gamma,\delta}, d^*)$ converges in payoff to one of the two most unfair equilibria in the original model, where the w group equilibrium is associated with Eq_1 and the b group equilibrium is associated with Eq_5 in Figure 3.

The existence of a game with the same X, featuring a fair equilibrium, whose equivalence class contains only itself, with the same decision policy, distribution, and beliefs as $(\overline{c}^*_{\gamma,\delta}, d^*)$ follows from the following lemma.

Lemma 2. *Given an X, a distribution $\mu : X \to (0,1)$, a set of beliefs $f : X \to (0,1)$, and a non-random decision policy $d : X \to \{0,1\}$ that is a threshold function of f, there exists a game $(X, \lambda_w, p, G, v_q, v_u, \omega)$ with the property that, (μ, f) is the rational distribution and beliefs supporting a fair equilibrium with decision policy d, whose equivalence class contains only itself.*

Proof. The proof is constructive. Fix X and (d, μ, f) as in the lemma.

Let $\{\mathcal{G}\} \cup \{(q_x, u_x)\}_{x \in X}$ be a collection of variables satisfying the following system of equations:

$$\begin{aligned} \mu(x) &= \mathcal{G} q_x + (1-\mathcal{G}) u_x \qquad \forall x \in X, \\ f(x) &= \frac{\mathcal{G} q_x}{\mathcal{G} q_x + (1-\mathcal{G}) u_x} \qquad \forall x \in X, \\ \sum_{x \in X} q_x &= \sum_{x \in X} u_x = 1. \end{aligned}$$

There exists a unique solution:

$$\begin{aligned} \mathcal{G} &= \sum_{x \in X} \mu(x) f(x) \in (0,1), \\ q_x &= \frac{\mu(x) f(x)}{\mathcal{G}} \in (0,1) \qquad \forall x \in X, \\ u_x &= \frac{\mu(x)(1 - f(x))}{1 - \mathcal{G}} \in (0,1) \qquad \forall x \in X. \end{aligned}$$

Define the following game:

- λ_w can be any value,
- G satisfies $G(\overline{c}(w)) = G(\overline{c}(b)) = \mathcal{G}$, where $\overline{c}(w) = \overline{c}(b) = \sum_{x \in X} d(x)(q_x - u_x)$,
- $p(x|i, q) = q_x$ and $p(x|i, u) = u_x$ for all $(i, x) \in I \times X$,
- v_q, v_u satisfy $d(x) = 1 \Leftrightarrow f(x) > \frac{v_u}{v_q + v_u}$,
- $\omega = 1$.

It is straightforward to check that $(\overline{c}, d)$ is a fair equilibrium with $(\mu_{RE}, f_{RE}) = (\mu, f)$, whose equivalence class contains only itself. $\square$

6.3 Proof of Theorem 3

Lemma 3. *Define* $c_{max} := \max_{x \in X} WW(l(x))$. *If* $\overline{c}^*(i) = \arg\max_{\overline{c}(i) \in \mathbb{R}} U_i(\overline{c}(i), d(i))$ *for some* $d(i)$*, then* $\overline{c}^*(i) \in [-c_{max}, c_{max}]$.

Proof. By definition, $\overline{c}^*(i) = \omega \sum_{x \in X} [p(x|i,q) - p(x|i,u)] d(i,x)$. The right hand side of the equation achieves its highest value, c_{max}, when $d(i)$ is a $\lceil 1 \rceil^-$-threshold group policy. Likewise, it is achieves its lowest value, $-c_{max}$, when $1 - d(i)$ is a $\lceil 1 \rceil^-$-threshold group policy. $\square$

Let k be a continuous ideal intervention. Fix a game $(X, \lambda_w, p, G, v_q, v_u, \omega)$.

Let $(\overline{c}^*, d^*)$ be an unfair equilibrium. Suppose there exists a $\overline{\delta} > 0$ and a sequence $(d_t, (\mu_t^*, f_t^*), \varepsilon_t > 0)_{t \in \mathbb{Z}^+}$ such that $\lim_{t \to \infty} \varepsilon_t = 0$ and, for all t, $d_t \in \mathcal{D}(k, \mu_t^*, f_t^*)$, $\|d_t - \hat{d}\| \geq \overline{\delta}$ for all $\hat{d} \in \mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$, and $\|(\mu_t^*, f_t^*) - (\mu_{RE}^*, f_{RE}^*)\| \leq \varepsilon_t$. Since $[0,1]^{|I \times X|}$ is compact, $(d_t)_{t \in \mathbb{Z}^+}$ has a convergent subsequence. By picking such a subsequence and relabelling, it is without loss of generality to assume there exists a decision policy d such that $\lim_{t \to \infty} d_t = d$. Since $\|d - \hat{d}\| \geq \overline{\delta}$ for all $\hat{d} \in \mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$, so $d \notin \mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$.

Since $U_{DM}(d, \mu, f)$ is continuous and $k(X, \mu, f)$ is compact-valued and continuous in (μ, f), by Berge's Maximum Theorem, $\mathcal{D}(k, \mu, f)$ is upper hemicontinuous. So $d \in \mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$. Contradiction. This implies, for every $\delta > 0$, there exists an $\varepsilon > 0$ such that, if $\|(\mu^*, f^*) - (\mu_{RE}^*, f_{RE}^*)\|_2 \leq \varepsilon$, then for every $d \in \mathcal{D}(k, \mu^*, f^*)$, there exists a $\hat{d} \in \mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$ such that $\|d - \hat{d}\| \leq \delta$. Right now, the ε depends on $(\overline{c}^*, d^*)$. However, since the number of equivalence classes of equilibria is finite, ε can be chosen independently of $(\overline{c}^*, d^*)$. This proves that k satisfies the first property of being robust to small deviations from rationality.

Now, suppose there exists a $\overline{\delta} > 0$ and a sequence $\left((\overline{c}_{\varepsilon_t}^*, d_{\varepsilon_t}^*), (\mu_{\varepsilon_t}^*, f_{\varepsilon_t}^*), \varepsilon_t > 0\right)_{t \in \mathbb{Z}^+}$ such that $\lim_{t \to \infty} \varepsilon_t = 0$ and, for all t, $(\overline{c}_{\varepsilon_t}^*, d_{\varepsilon_t}^*)$ is a k ε_t-equilibrium supported by $(\mu_{\varepsilon_t}^*, f_{\varepsilon_t}^*)$ and $\|(\overline{c}_{\varepsilon_t}^*, d_{\varepsilon_t}^*) - (\overline{c}^{*'}, d^{*'})\| \geq \overline{\delta}$ for all k equilibria $(\overline{c}^{*'}, d^{*'})$. By Lemma 3, $\overline{c}_{\varepsilon_t}^* \in [-c_{max}, c_{max}]^2$ for all t. So, just like before, $\left((\overline{c}_{\varepsilon_t}^*, d_{\varepsilon_t}^*), (\mu_{\varepsilon_t}^*, f_{\varepsilon_t}^*)\right)_{t \in \mathbb{Z}^+}$ has a convergent subsequence, and it is without loss of generality to assume there exists a $(\overline{c}^*, d^*)$ such that $\lim_{t \to \infty} \left((\overline{c}_{\varepsilon_t}^*, d_{\varepsilon_t}^*), (\mu_{\varepsilon_t}^*, f_{\varepsilon_t}^*)\right) = ((\overline{c}^*, d^*), (\mu_{RE}^*, f_{RE}^*))$. Since $\|(\overline{c}^*, d^*) - (\overline{c}^{*'}, d^{*'})\| \geq \overline{\delta}$ for all k equilibria $(\overline{c}^{*'}, d^{*'})$, so $(\overline{c}^*, d^*)$ is not a k equilibrium. Berge's Maximum Theorem now implies $d^* \in \mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$ and $\overline{c}^*$ is a best response to d^*. Thus,

$(\overline{c}^*, d^*)$ is a k equilibrium. Contradiction. This proves that k satisfies the second property of being robust to small deviations from rationality.

6.4 Proof of Theorem 4

To prove that mistaken identity is robust to any distribution and biased beliefs, it suffices to follow along the proof that mistaken identity is ideal and verify that all the arguments remain valid when (μ_{RE}, f_{RE}) is generalized to (μ, f_B) for some μ and B. Here, I highlight a few waypoints.

Let k be mistaken identity. Given any $c \in \mathbb{R}$, it is the case that $f_{RE}(i, x|c)$ is co-monotonic with $l(x)$. In particular, given $\overline{c}$, $f_B(i, x|\overline{c}) = f_{RE}(i, x|B_i(\overline{c}(w), \overline{c}(b)))$ is co-monotonic with $l(x)$. So, for any μ and B, $k(X, \mu, f_B) \subset \mathcal{F}$, and, for any $d \in \mathcal{F}$ there is an equivalent $d' \in k(X, \mu, f_B)$.

Let $(\overline{c}^*, d^*)$ be an unfair equilibrium supported by (μ, f_B^*) for some μ and B. Then d^* is a threshold decision policy that is not fair. Without loss of generality, we can assume $l_w^* < l_b^*$. Let $d \in \mathcal{D}(k, \mu, f_B^*)$. Then d is fair with some likelihood threshold l. And now the exact same argument as before shows that k satisfies the Gains to the Discriminated Group property.

Next, let $(\overline{c}^*, d^*)$ be a k equilibrium supported by (μ, f_B^*) for some μ and B. Since d^* is fair, $\overline{c}^*(w) = \overline{c}^*(b)$. When the cost-thresholds of representative subjects are equal, the bias B has no affect on beliefs. So $f_B^* = f_{RE}^*$. This implies $\mathcal{D}(\mu, f_B^*) = \mathcal{D}(\mu, f_{RE}^*) = \mathcal{D}(\mu_{RE}^*, f_{RE}^*)$. Now the rest of the proof that k satisfies the Steady States are Fair Equilibria property goes through exactly like before, except replace $\mathcal{D}(\mu_{RE}^*, f_{RE}^*)$ with $\mathcal{D}(\mu, f_B^*)$, $\mathcal{D}(k, \mu_{RE}^*, f_{RE}^*)$ with $\mathcal{D}(k, \mu, f_{RE}^*)$, and k equilibrium with k equilibrium supported by (μ, f_B^*) for any μ and B.

Now, let k be equal opportunity. To show that k is not robust to biased beliefs, it suffices to find an unfair equilibrium $(\overline{c}^*, d^*)$ and a bias B, such that $(\overline{c}^*, d^*)$ is also an unfair equilibrium supported by (μ_{RE}^*, f_B^*), and, under (μ_{RE}^*, f_B^*), d^* equalizes true positive rates across groups. This makes $(\overline{c}^*, d^*)$ an equal opportunity equilibrium supported by (μ_{RE}^*, f_B^*) that is not a fair equilibrium. From now on, I restrict attention only to B of the form $B(\overline{c}) := (\overline{c}(w) + \beta(\overline{c}(w) - \overline{c}(b)), \overline{c}(b) + \beta(\overline{c}(b) - \overline{c}(w)))$ for all $\overline{c} \in \mathbb{R}^2$, where β is some weakly increasing continuous function with $\beta(0) = 0$.

To find such an equilibrium, consider the simplified version of the family of models in the proof of Theorem 2, where X_2 is removed. In the analysis below, when the letter B appears, it will be clear from context if B denotes a bias or a letter grade taken by x. The group equilibria are still represented by Figure 3. Let $(\overline{c}^*, d^*)$ be the unfair equilibrium where the w group equilibrium is associated with Eq_1 and the b group equilibrium is associated with the second intersection of WW and EE from the left. Based on Figure 3 and the characterization of true positive rates under threshold group policies in the proof of Theorem 1, we know the rational true positive rate of the w group is $\frac{2}{3}$, and the rational true positive rate of the b group is less than $\frac{2}{3}$ and converges to $\frac{2}{3}$ as γ converges to $\frac{1}{6}$.

Now, consider a bias $\hat{B}$ with $\hat{\beta}$ satisfying $\hat{\beta}(\overline{c}^*(w) - \overline{c}^*(b)) = \overline{EE}(l_1) - \frac{1}{3}$ and $\hat{\beta}(c) = 0$ for all $c \leq 0$. $(\overline{c}^*, d^*)$ is also an unfair equilibrium supported by $(\mu^*_{RE}, f^*_{\hat{B}})$. Given $\overline{c}^*(w)$, as γ converges to $\frac{1}{6}$, $(\mu^*_{RE}(w,A), \mu^*_{RE}(w,B), \mu^*_{RE}(w,C))$ converges to

$$\left(\frac{1}{2}\cdot\left(\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{2}\cdot 0\right), \frac{1}{2}\cdot\left(\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{1}{3}\right), \frac{1}{2}\cdot\left(\frac{1}{2}\cdot\frac{1}{3}+\frac{1}{2}\cdot\frac{2}{3}\right)\right) = \left(\frac{1}{12}, \frac{1}{6}, \frac{1}{4}\right),$$

and $(f^*_{\hat{B}}(w,A), f^*_{\hat{B}}(w,B), f^*_{\hat{B}}(w,C))$ converges to

$$\left(\frac{\frac{2}{3}\cdot\frac{1}{3}}{\frac{2}{3}\cdot\frac{1}{3}+\frac{1}{3}\cdot 0}, \frac{\frac{2}{3}\cdot\frac{1}{3}}{\frac{2}{3}\cdot\frac{1}{3}+\frac{1}{3}\cdot\frac{1}{3}}, \frac{\frac{2}{3}\cdot\frac{1}{3}}{\frac{2}{3}\cdot\frac{1}{3}+\frac{1}{3}\cdot\frac{2}{3}}\right) = \left(1, \frac{2}{3}, \frac{1}{2}\right).$$

Together, they imply that $TP(w, d^*, \mu^*_{RE}, f^*_{\hat{B}})$ converges to

$$\frac{\frac{1}{12}\cdot 1+\frac{1}{6}\cdot\frac{2}{3}}{\frac{1}{12}\cdot 1+\frac{1}{6}\cdot\frac{2}{3}+\frac{1}{4}\cdot\frac{1}{2}} = \frac{14}{23} < \frac{2}{3}.$$

Continuity of TP then implies there exist a γ sufficiently close to $\frac{1}{6}$ and a bias B with an β satisfying $\beta(\overline{c}^*(w) - \overline{c}^*(b)) \in (0, \overline{EE}(l_1) - \frac{1}{3})$ and $\beta(c) = 0$ for all $c \leq 0$, such that, under (μ^*_{RE}, f^*_B), d^* equalizes true positive rates across groups. Moreover, β satisfying $\beta(\overline{c}^*(w) - \overline{c}^*(b)) \in (0, \overline{EE}(l_1) - \frac{1}{3})$ implies $(\overline{c}^*, d^*)$ is also an unfair equilibrium supported by (μ^*_{RE}, f^*_B).

References

Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz (2023) "Combining human expertise with artificial intelligence: Experimental evidence from radiology,"Technical report, National Bureau of Economic Research.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb (2018) *Prediction machines: the* *simple economics of artificial intelligence*: Harvard Business Press.

Aigner, Dennis J and Glen G Cain (1977) "Statistical theories of discrimination in labor markets," *Ilr Review*, 30 (2), 175–187.

Angelova, Victoria, Will S Dobbie, and Crystal Yang (2023) "Algorithmic recommendations and human discretion,"Technical report, National Bureau of Economic Research.

Arnold, David, Will Dobbie, and Peter Hull (2021) "Measuring racial discrimination in algorithms," in *AEA Papers and Proceedings*, 111, 49–54, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Arrow, Kenneth (1971) "The theory of discrimination."

Athey, Susan C, Kevin A Bryan, and Joshua S Gans (2020) "The allocation of decision authority to human and artificial intelligence," in *AEA Papers and Proceedings*, 110, 80–84, American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203.

Bardhi, Arjada, Yingni Guo, and Bruno Strulovici (2020) "Early-Career Discrimination: Spiraling or Self-Correcting?".

Barocas, Solon, Moritz Hardt, and Arvind Narayanan (2019) *Fairness and Machine Learning: Limitations and Opportunities*: fairmlbook.org, `http://www.fairmlbook.org`.

Bastani, Hamsa, Osbert Bastani, and Wichinpong Park Sinchaisri (2021) "Improving human decision-making with machine learning," *arXiv preprint arXiv:2108.08454*.

Blattner, Laura, Scott Nelson, and Jann Spiess (2021) "Unpacking the black box: Regulating algorithmic decisions," *arXiv preprint arXiv:2110.03443*.

Bohren, J Aislinn, Kareem Haggag, Alex Imas, and Devin G Pope (2023) "Inaccurate statistical discrimination: An identification problem," *Review of Economics and Statistics*, 1–45.

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg (2019) "The dynamics of discrimination: Theory and evidence," *American economic review*, 109 (10), 3395–3436.

Chambers, Christopher P and Federico Echenique (2021) “A characterisation of ‘Phelpsian’statistical discrimination,” *The Economic Journal*, 131 (637), 2018–2032.

Che, Yeon-Koo, Kyungmin Kim, and Weijie Zhong (2020) “Statistical discrimination in ratings-guided markets,” *arXiv preprint arXiv:2004.11531*.

Chiappa, Silvia (2019) “Path-specific counterfactual fairness,” in *Proceedings of the AAAI conference on artificial intelligence*, 33, 7801–7808.

Chouldechova, Alexandra (2017) “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big data*, 5 (2), 153–163.

Coate, Stephen and Glenn C Loury (1993) “Will affirmative-action policies eliminate negative stereotypes?” *The American Economic Review*, 1220–1240.

Corbett-Davies, Sam, J Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel (2023) “The measure and mismeasure of fairness,” *J. Mach. Learn. Res.*

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (2017) “Algorithmic decision making and the cost of fairness,” in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Cowgill, Bo and Catherine E Tucker (2020) “Algorithmic fairness and economics,” *Columbia Business School Research Paper*.

D’Amour, Alexander, Hansa Srinivasan, James Atwood, Pallavi Baljekar, David Sculley, and Yoni Halpern (2020) “Fairness is not static: deeper understanding of long term fairness via simulation studies,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 525–534.

Deb, Rahul, Ludovic Renou et al. (2022) *Which wage distributions are consistent with statistical discrimination?*: University of Toronto, Department of Economics.

Donahue, Kate, Alexandra Chouldechova, and Krishnaram Kenthapadi (2022) “Human-algorithm collaboration: Achieving complementarity and avoiding unfairness,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1639–1656.

Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel (2012) "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Echenique, Federico and Anqi Li (2022) "Rationally inattentive statistical discrimination: Arrow meets Phelps," *arXiv preprint arXiv:2212.08219*.

Eliaz, Kfir and Ran Spiegler (2019) "The model selection curse," *American Economic Review: Insights*, 1 (2), 127–140.

Escudé, Matteo, Paula Onuchic, Ludvig Sinander, and Quitzé Valenzuela-Stookey (2022) "Statistical discrimination and statistical informativeness," *arXiv preprint arXiv:2205.07128*.

Fang, Hanming and Andrea Moro (2011) "Theories of statistical discrimination and affirmative action: A survey," *Handbook of social economics*, 1, 133–200.

Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian (2015) "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Fosgerau, Mogens, Rajiv Sethi, and Jörgen Weibull (2023) "Equilibrium screening and categorical inequality," *American Economic Journal: Microeconomics*, 15 (3), 201–242.

Frankel, Alex and Navin Kartik (2022) "Improving information from manipulable data," *Journal of the European Economic Association*, 20 (1), 79–115.

Fu, Runshan, Manmohan Aseri, Param Vir Singh, and Kannan Srinivasan (2022) ""Un" fair machine learning algorithms," *Management Science*, 68 (6), 4173–4195.

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther (2022) "Predictably unequal? The effects of machine learning on credit markets," *The Journal of Finance*, 77 (1), 5–47.

Gillis, Talia, Bryce McLaughlin, and Jann Spiess (2021) "On the fairness of machine-assisted human decisions," *arXiv preprint arXiv:2110.15310*.

Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters (2016a) "Strategic classification," in *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, 111–122.

Hardt, Moritz, Eric Price, and Nati Srebro (2016b) "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, 29.

Heidari, Hoda, Claudio Ferrari, Krishna Gummadi, and Andreas Krause (2018) "Fairness behind a veil of ignorance: A welfare analysis for automated decision making," *Advances in neural information processing systems*, 31.

Hu, Lily and Yiling Chen (2018) "A short-term intervention for long-term fairness in the labor market," in *Proceedings of the 2018 World Wide Web Conference*, 1389–1398.

——— (2020) "Fair classification and social welfare," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 535–545.

Hu, Lily, Nicole Immorlica, and Jennifer Wortman Vaughan (2019) "The disparate effects of strategic manipulation," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 259–268.

Jarosch, Gregor and Laura Pilossoph (2019) "Statistical discrimination and duration dependence in the job finding rate," *The Review of Economic Studies*, 86 (4), 1631–1665.

Jung, Christopher, Sampath Kannan, Changhwa Lee, Mallesh Pai, Aaron Roth, and Rakesh Vohra (2020) "Fair prediction with endogenous behavior," in *Proceedings of the 21st ACM Conference on Economics and Computation*, 677–678.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018a) "Human decisions and machine predictions," *The Quarterly Journal of Economics*, 133 (1), 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein (2018b) "Discrimination in the Age of Algorithms," *Journal of Legal Analysis*, 10, 113–174.

——— (2020) "Algorithms as discrimination detectors," *Proceedings of the National Academy of Sciences*, 117 (48), 30096–30100.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2016) "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*.

Kleinberg, Jon and Manish Raghavan (2020) "How do classifiers induce agents to invest effort strategically?" *ACM Transactions on Economics and Computation (TEAC)*, 8 (4), 1–23.

Komiyama, Junpei and Shunya Noda (2020) "On statistical discrimination as a failure of social learning: A multi-armed bandit approach," *arXiv preprint arXiv:2010.01079*.

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

Li, Danielle, Lindsey R Raymond, and Peter Bergman (2020) "Hiring as exploration,"Technical report, National Bureau of Economic Research.

Liang, Annie, Jay Lu, and Xiaosheng Mu (2023) "Algorithm Design: A Fairness-Accuracy Frontier," *arXiv preprint arXiv:2112.09975*.

Liu, Lydia T, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt (2018) "Delayed impact of fair machine learning," in *International Conference on Machine Learning*, 3150–3158, PMLR.

Liu, Lydia T, Ashia Wilson, Nika Haghtalab, Adam Tauman Kalai, Christian Borgs, and Jennifer Chayes (2020) "The disparate equilibria of algorithmic decision making when individuals invest rationally," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 381–391.

Mailath, George J, Larry Samuelson, and Avner Shaked (2000) "Endogenous inequality in integrated labor markets with two-sided search," *American Economic Review*, 91 (1), 46–72.

Martin, Daniel and Philip Marx (2022) "A Robust Test of Prejudice for Discrimination Experiments," *Management Science*, 68 (6), 4527–4536.

Marx, Philip (2022) "An absolute test of racial prejudice," *The Journal of Law, Economics, and Organization*, 38 (1), 42–91.

McLaughlin, Bryce and Jann Spiess (2022) "Algorithmic assistance with recommendation-dependent preferences," *arXiv preprint arXiv:2208.07626*.

Moro, Andrea and Peter Norman (2004) "A general equilibrium model of statistical discrimination," *Journal of economic theory*, 114 (1), 1–30.

Mouzannar, Hussein, Mesrob I Ohannessian, and Nathan Srebro (2019) "From fair decision making to social equality," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 359–368.

Onuchic, Paula (2022) "Recent contributions to theories of discrimination," *arXiv preprint arXiv:2205.05994*.

Onuchic, Paula and Debraj Ray (2023) "Signaling and discrimination in collaborative projects," *American Economic Review*, 113 (1), 210–252.

Penn, Elizabeth Maggie and John W Patty (2023) "Algorithms, Incentives, and Democracy," *arXiv preprint arXiv:2307.02319*.

Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt (2020) "Performative prediction," in *International Conference on Machine Learning*, 7599–7609, PMLR.

Phelps, Edmund S (1972) "The statistical theory of racism and sexism," *The american economic review*, 62 (4), 659–661.

Pope, Devin G and Justin R Sydnor (2011) "Implementing anti-discrimination policies in statistical profiling models," *American Economic Journal: Economic Policy*, 3 (3), 206–231.

Puranik, Bhagyashree, Upamanyu Madhow, and Ramtin Pedarsani (2022) "A Dynamic Decision-Making Framework Promoting Long-Term Fairness," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 547–556.

Rambachan, Ashesh, Jon Kleinberg, Sendhil Mullainathan, and Jens Ludwig (2020) "An economic approach to regulating algorithms,"Technical report, National Bureau of Economic Research.

Rosén, Åsa (1997) "An equilibrium search-matching model of discrimination," *European Economic Review*, 41 (8), 1589–1613.

Segal, Meirav, Anne-Marie George, and Christos Dimitrakakis (2023) "Policy Fairness and Unknown Bias Dynamics in Sequential Allocations," in *Proceedings of* *the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–10.

Shimao, Hajime, Warut Khern-am nuai, Karthik Natarajan Kannan, and Maxime C Cohen (2023) "Addressing fairness in machine learning predictions: Strategic best-response fair discriminant removed algorithm," *Available at SSRN 3389631*.

Williams, Joshua and J Zico Kolter (2019) "Dynamic modeling and equilibria in fair decision making," *arXiv preprint arXiv:1911.06837*.

Yang, Crystal S and Will Dobbie (2020) "Equal protection under algorithms: A new statistical and legal framework," *Mich. L. Rev.*, 119, 291.

Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork (2013) "Learning fair representations," in *International conference on machine learning*, 325–333, PMLR.