Delayed Impact of Fair Machine Learning

Lydia T. Liu [1] **Sarah Dean** [1] **Esther Rolf** [1] **Max Simchowitz** [1] **Moritz Hardt** [1]

Abstract

Fairness in machine learning has predominantly been studied in static classification settings without concern for how decisions change the underlying population over time. Conventional wisdom suggests that fairness criteria promote the long-term well-being of those groups they aim to protect. We study how static fairness criteria interact with temporal indicators of well-being, such as long-term improvement, stagnation, and decline in a variable of interest. We demonstrate that even in a one-step feedback model, common fairness criteria in general do not promote improvement over time, and may in fact cause harm in cases where an unconstrained objective would not. We completely characterize the delayed impact of three standard criteria, contrasting the regimes in which these exhibit qualitatively different behavior. In addition, we find that a natural form of measurement error broadens the regime in which fairness criteria perform favorably. Our results highlight the importance of measurement and temporal modeling in the evaluation of fairness criteria, suggesting a range of new challenges and trade-offs.

1. Introduction

Machine learning commonly considers static objectives defined on a snapshot of the population at one instant in time; consequential decisions, in contrast, reshape the population over time. Lending practices, for example, can shift the distribution of debt and wealth in the population. Job advertisements allocate opportunity. School admissions shape the level of education in a community.

Existing scholarship on fairness in automated decision-making criticizes unconstrained machine learning for its potential to *harm* historically underrepresented or disadvantaged groups in the population (Executive Office of the President, 2016; Barocas & Selbst, 2016). Consequently, a variety of *fairness criteria* have been proposed as constraints on standard learning objectives. Even though, in each case, these constraints are clearly intended to *protect* the disadvantaged group by an appeal to intuition, a rigorous argument to that effect is often lacking.

In this work, we formally examine under what circumstances fairness criteria do indeed promote the long-term well-being of disadvantaged groups measured in terms of a temporal variable of interest. Going beyond the standard classification setting, we introduce a one-step feedback model of decision-making that exposes how decisions change the underlying population over time.

Our running example is a hypothetical lending scenario. There are two groups in the population with features described by a summary statistic, such as a *credit score*, whose distribution differs between the two groups. The bank can choose thresholds for each group at which loans are offered. While group-dependent thresholds may face legal challenges (Ross & Yinger, 2006), they are generally inevitable for some of the criteria we examine. The impact of a lending decision has multiple facets. A default event not only diminishes profit for the bank, it also worsens the financial situation of the borrower as reflected in a subsequent decline in credit score. A successful lending outcome leads to profit for the bank and also to an increase in credit score for the borrower.

When thinking of one of the two groups as disadvantaged, it makes sense to ask what lending policies (choices of thresholds) lead to an expected improvement in the score distribution within that group. An unconstrained bank would maximize profit, choosing thresholds that meet a break-even point above which it is profitable to give out loans. One frequently proposed fairness criterion, sometimes called demographic parity, requires the bank to lend to both groups at an equal rate. Subject to this requirement the bank would continue to maximize profit to the extent possible. Another criterion, originally called equality of opportunity, equalizes the *true positive rates* between the two groups, thus requiring the bank to lend in both groups at an equal rate among individuals who repay their loan. Other criteria are natural, but for clarity we restrict our attention to these three.

[1]Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Berkeley, California, USA. Correspondence to: Lydia T. Liu <lydiatliu@berkeley.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Do these fairness criteria benefit the disadvantaged group? When do they show a clear advantage over unconstrained classification? Under what circumstances does profit maximization work in the interest of the individual? These are important questions that we begin to address in this work.

1.1. Contributions

We introduce a one-step feedback model that allows us to quantify the long-term impact of classification on different groups in the population. We represent each of the two groups A and B by a *score* distribution $\boldsymbol{\pi}_{\mathsf{A}}$ and $\boldsymbol{\pi}_{\mathsf{B}}$, respectively. The support of these distributions is a finite set $\mathcal{X}$ comprising the possible values that the score can assume. We think of the score as highlighting one variable of interest in a specific domain such that higher score values correspond to a higher probability of a positive outcome. An *institution* chooses selection policies $\boldsymbol{\tau}_{\mathsf{A}}, \boldsymbol{\tau}_{\mathsf{B}} \colon \mathcal{X} \to [0, 1]$ that assign to each value in $\mathcal{X}$ a number representing the rate of selection for that value. In our example, these policies specify the lending rate at a given credit score within a given group. The institution will always maximize their utility (see (1)) subject to either (a) no constraint, or (b) equality of selection rates, or (c) equality of true positive rates.

We assume the availability of a function $\boldsymbol{\Delta} \colon \mathcal{X} \to \mathbb{R}$ that provides the expected change in score for a selected individual at a given score. The central quantity we study is the expected difference $\Delta\boldsymbol{\mu}_{\mathsf{j}}$ in the mean score in group $j \in \{\mathsf{A}, \mathsf{B}\}$ that results from the selection policy. When modeling the problem, the expected mean difference can also absorb external factors such as "reversion to the mean" so long as they are mean-preserving. Qualitatively, we distinguish between *long-term improvement* ($\Delta\boldsymbol{\mu}_{\mathsf{j}} > 0$), *stagnation* ($\Delta\boldsymbol{\mu}_{\mathsf{j}} = 0$), and *decline* ($\Delta\boldsymbol{\mu}_{\mathsf{j}} < 0$).

Our findings can be summarized as follows.

1. Both fairness criteria (equal selection rates, equal true positive rates) can lead to all possible outcomes (improvement, stagnation, and decline) in natural parameter regimes. We provide a complete characterization of when each criterion leads to each outcome in section 3.

   - There are a class of settings where equal selection rates cause decline, whereas equal true positive rates do not (Theorem 3.5),
   - Under a mild assumption, the institution's optimal unconstrained selection policy can never lead to decline (Proposition 3.1).

2. We introduce the notion of an *outcome curve* (Figure 1) which succinctly describes the different regimes in which one criterion is preferable over the others.

3. We perform experiments on FICO credit score data from 2003 and show that under various models of bank utility and score change, the outcomes of applying fairness criteria are in line with our theoretical predictions.

4. We discuss how certain types of measurement error (e.g., the bank underestimating the repayment ability of the disadvantaged group) affect our comparison. We find that measurement error narrows the regime in which fairness criteria cause decline, suggesting that measurement should be a factor when motivating these criteria.

5. We consider alternatives to hard fairness constraints.

   - We evaluate the optimization problem where fairness criterion is a regularization term in the objective. Qualitatively, this leads to the same findings.
   - We discuss the possibility of optimizing for group score improvement $\Delta\boldsymbol{\mu}_{\mathsf{j}}$ directly subject to institution utility constraints. The resulting solution provides an interesting possible alternative to existing fairness criteria.

We focus on the impact of a selection policy over a single epoch. The motivation is that the designer of a system usually has an understanding of the time horizon after which the system is evaluated and possibly redesigned. Formally, nothing prevents us from repeatedly applying our model and tracing changes over multiple epochs. In reality, however, it is plausible that over greater time periods, economic background variables might dominate the effect of selection.

Reflecting on our findings, we argue that careful temporal modeling is necessary in order to accurately evaluate the impact of different fairness criteria on the population. Moreover, an understanding of measurement error is important in assessing the advantages of fairness criteria relative to unconstrained selection. Finally, the nuances of our characterization underline how intuition may be a poor guide in judging the long-term impact of fairness constraints.

1.2. Related work

Recent work by Hu & Chen (2018) considers a model for long-term outcomes in the labor market. They propose imposing the demographic parity constraint in a *temporary* labor market in order to provably achieve an equitable long-term equilibrium in the *permanent* labor market, reminiscent of economic arguments for affirmative action (e.g. Foster & Vohra, 1992; Coate & Loury, 1993). Our general framework is complementary to this type of domain specific approach.

Fuster et al. (2017) consider the problem of fairness in credit markets from a different perspective. Their goal is to study the effect of machine learning on interest rates in different groups at an equilibrium, under a static model without feedback.

Ensign et al. (2017) consider feedback loops in predictive policing, where the police more heavily monitor high crime

neighborhoods, thus further increasing the measured number of crimes in those neighborhoods. While the work addresses an important temporal phenomenon using the theory of urns, it is rather different from our one-step feedback model both conceptually and technically. Knowles et al. (2001) consider a more dynamic model in which individuals react to their probabilities of being searched.

Demographic parity and related formulations have been considered in numerous papers (e.g. Calders et al., 2009; Zafar et al., 2017). Hardt et al. (2016) introduced the equality of opportunity constraint and demonstrate limitations of a broad class of criteria. Kleinberg et al. (2017) and Chouldechova (2016) point out the tension between "calibration by group" and equal true/false positive rates. These trade-offs carry over to some extent to the case where we only equalize true positive rates (Pleiss et al., 2017).

A growing literature on fairness in the "bandits" setting of learning (see Joseph et al., 2016, *et seq.*) deals with online decision making that ought not to be confused with our one-step feedback setting. Finally, there has been much work in the social sciences on analyzing the effect of affirmative action (see e.g., Keith et al., 1985; Kalev et al., 2006).

2. Problem Setting

We consider two *groups* A and B, which comprise a g_{A} and $g_{\mathsf{B}} = 1 - g_{\mathsf{A}}$ fraction of the total population, and an *institution* which makes a binary decision for each individual in each group, called *selection*. Individuals in each group are assigned *scores* in $\mathcal{X} := [C]$, and the scores for group $\mathsf{j} \in \{\mathsf{A}, \mathsf{B}\}$ are distributed according $\boldsymbol{\pi}_{\mathsf{j}} \in \mathrm{Simplex}^{C-1}$. The institution selects a *policy* $\boldsymbol{\tau} := (\boldsymbol{\tau}_{\mathsf{A}}, \boldsymbol{\tau}_{\mathsf{B}}) \in [0,1]^{2C}$, where $\boldsymbol{\tau}_{\mathsf{j}}(x)$ corresponds to the probability the institution selects an individual in group j with score x. One should think of a score as an abstract quantity which summarizes how well an individual is suited to being selected; an example is provided at the end of this section.

We assume that the institution is utility-maximizing, but may impose certain constraints to ensure that the policy $\boldsymbol{\tau}$ is *fair*, in a sense described in Section 2.2. We assume that there exists a function $\boldsymbol{u} : \mathcal{X} \to \mathbb{R}$, such that the institution's expected utility for a policy $\boldsymbol{\tau}$ is given by

$$\mathcal{U}(\boldsymbol{\tau}) = \textstyle\sum_{\mathsf{j} \in \{\mathsf{A},\mathsf{B}\}} g_{\mathsf{j}} \sum_{x \in \mathcal{X}} \boldsymbol{\tau}_{\mathsf{j}}(x) \boldsymbol{\pi}_{\mathsf{j}}(x) \boldsymbol{u}(x). \tag{1}$$

Novel to this work, we focus on the effect of the selection policy $\boldsymbol{\tau}$ on the groups A and B. We quantify these *outcomes* in terms of an average effect that a policy $\boldsymbol{\tau}_{\mathsf{j}}$ has on group j. Formally, for a function $\boldsymbol{\Delta}(x) : \mathcal{X} \to \mathbb{R}$, we define the average change of the mean score $\boldsymbol{\mu}_{\mathsf{j}}$ for group j

$$\Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau}) := \textstyle\sum_{x \in \mathcal{X}} \boldsymbol{\pi}_{\mathsf{j}}(x) \boldsymbol{\tau}_{\mathsf{j}}(x) \boldsymbol{\Delta}(x)\,. \tag{2}$$

We remark that many of our results also go through if $\Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau})$ simply refers to an abstract change in well-being, not necessarily a change in the mean score. Lastly, we assume that the *success* of an individual is independent of their group given the score; that is, the score summarizes all relevant information about the success event, so there exists a function $\boldsymbol{\rho} : \mathcal{X} \to [0,1]$ such that individuals of score x succeed with probability $\boldsymbol{\rho}(x)$.

We introduce the specific domain of credit scores as a running example in the rest of the paper. Other examples showing the broad applicability of our model can be found in Appendix A.

Example 2.1 (Credit scores)**.** In the setting of loans, scores $x \in [C]$ represent credit scores, and the bank serves as the institution. The bank chooses to grant or refuse loans to individuals according to a policy $\boldsymbol{\tau}$. Both bank and personal utilities are given as functions of loan repayment, and therefore depend on the success probabilities $\boldsymbol{\rho}(x)$, representing the probability that any individual with credit score x can repay a loan within a fixed time frame. The expected utility to the bank is given by the expected return from a loan, which can be modeled as an affine function of $\boldsymbol{\rho}(x)$: $\boldsymbol{u}(x) = u_+ \boldsymbol{\rho}(x) + u_-(1 - \boldsymbol{\rho}(x))$, where u_+ denotes the profit when loans are repaid and u_- the loss when they are defaulted on. Individual outcomes of being granted a loan are based on whether or not an individual repays the loan, and a simple model for $\boldsymbol{\Delta}(x)$ may also be affine in $\boldsymbol{\rho}(x)$: $\boldsymbol{\Delta}(x) = c_+ \boldsymbol{\rho}(x) + c_-(1 - \boldsymbol{\rho}(x))$, modified accordingly at boundary states. The constant $c_+ > 0$ denotes the gain in credit score if loans are repaid and $c_- < 0$ is the score penalty in case of default.

2.1. The Outcome Curve

We now introduce important outcome regimes, stated in terms of the change in average group score. A policy $(\boldsymbol{\tau}_{\mathsf{A}}, \boldsymbol{\tau}_{\mathsf{B}})$ is said to cause *active harm* to group j if $\Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau}_{\mathsf{j}}) < 0$, *stagnation* if $\Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau}_{\mathsf{j}}) = 0$, and *improvement* if $\Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau}_{\mathsf{j}}) > 0$. We denote the policy that maximizes the institution's utility in the absence of constraints as `MaxUtil`. Under our model, `MaxUtil` policies can be chosen in a standard fashion which applies the same threshold $\boldsymbol{\tau}^{\texttt{MaxUtil}}$ for both groups, and is agnostic to the distributions $\boldsymbol{\pi}_{\mathsf{A}}$ and $\boldsymbol{\pi}_{\mathsf{B}}$. Hence, if we define

$$\Delta\boldsymbol{\mu}_{\mathsf{j}}^{\texttt{MaxUtil}} := \Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau}^{\texttt{MaxUtil}}) \tag{3}$$

we say that a policy causes *relative harm* to group j if $\Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau}_{\mathsf{j}}) < \Delta\boldsymbol{\mu}_{\mathsf{j}}^{\texttt{MaxUtil}}$, and *relative improvement* if $\Delta\boldsymbol{\mu}_{\mathsf{j}}(\boldsymbol{\tau}_{\mathsf{j}}) > \Delta\boldsymbol{\mu}_{\mathsf{j}}^{\texttt{MaxUtil}}$. In particular, we focus on these outcomes for a disadvantaged group, and consider whether imposing a fairness constraint improves their outcomes relative to the `MaxUtil` strategy. From this point forward, we take A to be the disadvantaged or protected group.

Figure 1 displays the important outcome regimes in terms of *selection rates* $\beta_{\mathsf{j}} := \sum_{x \in \mathcal{X}} \boldsymbol{\pi}_{\mathsf{j}}(x) \boldsymbol{\tau}_{\mathsf{j}}(x)$. This succinct

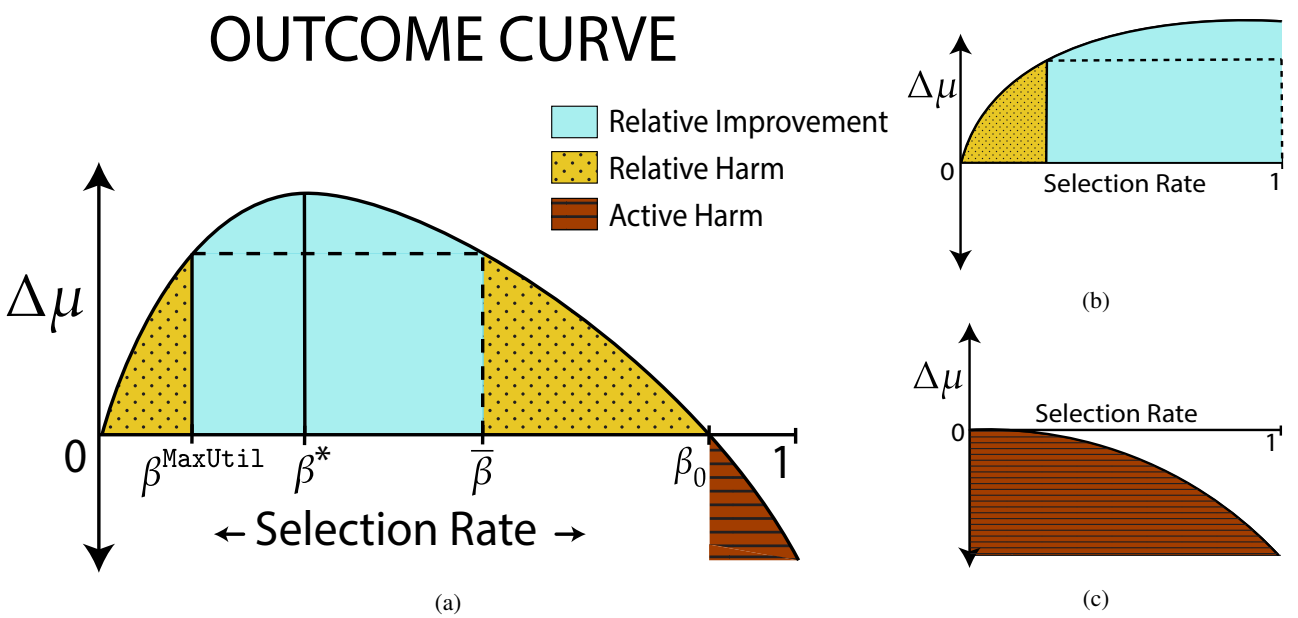

Figure 1. The above figure shows the *outcome curve*. The horizontal axis represents the selection rate for the population; the vertical axis represents the mean change in score. (a) depicts the full spectrum of outcome regimes, and colors indicate regions of active harm, relative harm, and no harm. In (b): a group that has much potential for gain, in (c): a group that has no potential for gain.

characterization is possible when considering decision rules based on (possibly randomized) score thresholding, in which all individuals with scores above a threshold are selected. In Appendix B, we justify the restriction to such *threshold policies* by showing it preserves optimality. In Appendix B.1, we show that the outcome curve is concave, thus implying that it takes the shape depicted in Figure 1. To explicitly connect selection rates to decision policies, we define the rate function $r_{\boldsymbol{\pi}}(\boldsymbol{\tau}_{\mathsf{j}})$ which returns the proportion of group j selected by the policy. We show that this function is invertible for a suitable class of threshold policies, and in fact the outcome curve is precisely the graph of the map from selection rate to outcome $\beta \mapsto \Delta\boldsymbol{\mu}_{\mathsf{A}}(r^{-1}_{\boldsymbol{\pi}_{\mathsf{A}}}(\beta))$. Next, we define the values of β that mark boundaries of the outcome regions.

Definition 2.1 (Selection rates of interest)**.** Given the protected group A, the following selection rates are of interest in distinguishing between qualitatively different classes of outcomes (Figure 1). We define $\beta^{\texttt{MaxUtil}}$ as the selection rate for A under `MaxUtil`; β_0 as the harm threshold, such that $\Delta\boldsymbol{\mu}_{\mathsf{A}}(r^{-1}_{\boldsymbol{\pi}_{\mathsf{A}}}(\beta_0)) = 0$; β^* as the selection rate such that $\Delta\boldsymbol{\mu}_{\mathsf{A}}$ is maximized; $\overline{\beta}$ as the outcome-complement of the `MaxUtil` selection rate, $\Delta\boldsymbol{\mu}_{\mathsf{A}} r^{-1}_{\boldsymbol{\pi}_{\mathsf{A}}}(\overline{\beta})) = \Delta\boldsymbol{\mu}_{\mathsf{A}}(r^{-1}_{\boldsymbol{\pi}_{\mathsf{A}}}(\beta^{\texttt{MaxUtil}}))$ with $\overline{\beta} > \beta^{\texttt{MaxUtil}}$.

2.2. Decision Rules and Fairness Criteria

We will consider policies that maximize the institution's total expected utility, potentially subject to a constraint: $\boldsymbol{\tau} \in \mathcal{C} \in [0,1]^{2C}$ which enforces some notion of "fairness". Formally, the institution selects $\tau_* \in \operatorname{argmax} \mathcal{U}(\boldsymbol{\tau})$ s.t. $\boldsymbol{\tau} \in \mathcal{C}$. We consider the three following constraints:

Definition 2.2 (Fairness criteria)**.** The *maximum utility* (`MaxUtil`) policy corresponds to the null-constraint $\mathcal{C} = [0,1]^{2C}$, so that the institution is free to focus solely on utility. The *demographic parity* (`DemParity`) policy results in equal selection rates between both groups. Formally, the constraint is $\mathcal{C} = \left\{(\boldsymbol{\tau}_{\mathsf{A}}, \boldsymbol{\tau}_{\mathsf{B}}) : \sum_{x\in\mathcal{X}} \boldsymbol{\pi}_{\mathsf{A}}(x)\boldsymbol{\tau}_{\mathsf{A}} = \sum_{x\in\mathcal{X}} \boldsymbol{\pi}_{\mathsf{B}}(x)\boldsymbol{\tau}_{\mathsf{B}}\right\}$. The *equal opportunity* (`EqOpt`) policy results in equal true positive rates (TPR) between both group, where TPR is defined as $\mathrm{TPR}_{\mathsf{j}}(\boldsymbol{\tau}) := \frac{\sum_{x\in\mathcal{X}} \boldsymbol{\pi}_{\mathsf{j}}(x)\boldsymbol{\rho}(x)\boldsymbol{\tau}(x)}{\sum_{x\in\mathcal{X}} \boldsymbol{\pi}_{\mathsf{j}}(x)\boldsymbol{\rho}(x)}$. `EqOpt` ensures that the conditional probability of selection given that the individual will be successful is independent of the population, formally enforced by the constraint $\mathcal{C} = \{(\boldsymbol{\tau}_{\mathsf{A}}, \boldsymbol{\tau}_{\mathsf{B}}) : \mathrm{TPR}_{\mathsf{A}}(\boldsymbol{\tau}_{\mathsf{A}}) = \mathrm{TPR}_{\mathsf{B}}(\boldsymbol{\tau}_{\mathsf{B}})\}$.

Just as the expected outcome $\Delta\boldsymbol{\mu}$ can be expressed in terms of selection rate for threshold policies, so can the total utility $\mathcal{U}$. In the unconstrained case, $\mathcal{U}$ varies independently over the selection rates for group A and B; however, in the presence of fairness constraints the selection rate for one group determines the allowable selection rate for the other. The selection rates must be equal for `DemParity`, but for `EqOpt` we can define a *transfer function*, $G^{(\mathsf{A}\to\mathsf{B})}$, which for every loan rate β in group A gives the loan rate in group B that has the same true positive rate. Therefore, when considering threshold policies, decision rules amount to maximizing functions of single parameters. This idea is expressed in

Figure 2, and underpins the results to follow.

3. Results

In order to clearly characterize the outcome of applying fairness constraints, we make the following assumption.

Assumption 1 (Institution utilities)**.** *The institution's individual utility function is more stringent than the expected score changes,* $\boldsymbol{u}(x) > 0 \implies \boldsymbol{\Delta}(x) > 0$*. (For the linear form presented in Example 2.1,* $\frac{u_-}{u_+} < \frac{c_-}{c_+}$ *is necessary and sufficient.)*

This simplifying assumption quantifies the intuitive notion that institutions take a greater risk by accepting than the individual does by applying. For example, in the credit setting, a bank loses the amount loaned in the case of a default, but makes only interest in case of a payback. Using Assumption 1, we can restrict the position of `MaxUtil` on the outcome curve in the following sense.

Proposition 3.1 (`MaxUtil` does not cause active harm)**.** *Under Assumption 1,* $0 \leq \Delta\boldsymbol{\mu}^{\texttt{MaxUtil}} \leq \Delta\boldsymbol{\mu}^{*}$.

We direct the reader to Appendix F for the proof of the above proposition, and all subsequent theorems presented in this section.

3.1. Prospects and Pitfalls of Fairness Criteria

We begin by characterizing general settings under which fairness criteria act to improve outcomes over unconstrained `MaxUtil` strategies. For this result, we will assume that group A is disadvantaged in the sense that the `MaxUtil` acceptance rate for B is large compared to relevant acceptance rates for A.

Theorem 3.2 (Fairness criteria can cause relative improvement)**.** *(a) Under the assumption that* $\beta_{\mathsf{A}}^{\texttt{MaxUtil}} < \overline{\beta}$ *and* $\beta_{\mathsf{B}}^{\texttt{MaxUtil}} > \beta_{\mathsf{A}}^{\texttt{MaxUtil}}$*, there exist population proportions* $g_0 < g_1 < 1$ *such that, for all* $g_{\mathsf{A}} \in [g_0, g_1]$*,* $\beta_{\mathsf{A}}^{\texttt{MaxUtil}} < \beta_{\mathsf{A}}^{\texttt{DemParity}} < \overline{\beta}$*. That is,* `DemParity` *causes relative improvement.*

(b) Under the assumption that there exist $\beta_{\mathsf{A}}^{\texttt{MaxUtil}} < \beta < \beta' < \overline{\beta}$ *such that* $\beta_{\mathsf{B}}^{\texttt{MaxUtil}} > G^{(\mathsf{A}\to\mathsf{B})}(\beta), G^{(\mathsf{A}\to\mathsf{B})}(\beta')$*, there exist population proportions* $g_2 < g_3 < 1$ *such that, for all* $g_{\mathsf{A}} \in [g_2, g_3]$*,* $\beta_{\mathsf{A}}^{\texttt{MaxUtil}} < \beta_{\mathsf{A}}^{\texttt{EqOpt}} < \overline{\beta}$*. That is,* `EqOpt` *causes relative improvement.*

This result gives the conditions under which we can guarantee the existence of settings in which fairness criteria cause improvement relative to `MaxUtil`. Relying on machinery proved in the appendix, the result follows from comparing the position of optima on the utility curve to the outcome curve. Figure 2 displays a illustrative example of both the outcome curve and the institutions' utility $\mathcal{U}$ as a function of the selection rates in group A. In the utility function (1), the contributions of each group are weighted by their population proportions g_{j}, and thus the resulting selection rates are sensitive to these proportions.

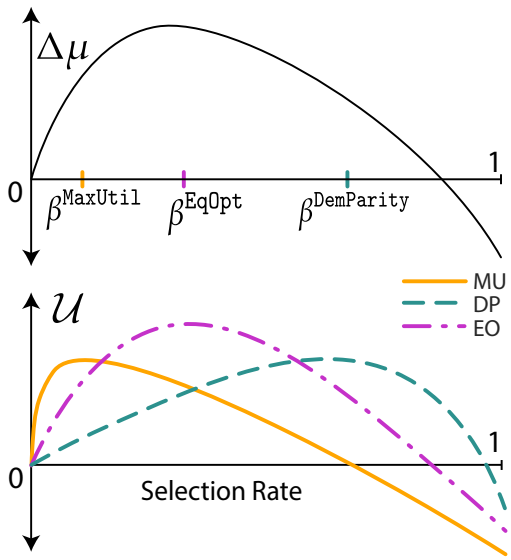

Figure 2. Both outcomes $\Delta\boldsymbol{\mu}$ and institution utilities $\mathcal{U}$ can be plotted as a function of selection rate for one group. The maxima of the utility curves determine the selection rates resulting from various decision rules.

As we see in the remainder of this section, fairness criteria can achieve nearly any position along the outcome curve under the right conditions. This fact comes from the potential mismatch between the outcomes, controlled by $\boldsymbol{\Delta}$, and the institution's utility $\boldsymbol{u}$.

The next theorem implies that `DemParity` can be bad for long term well-being of the protected group by being over-generous, under the mild assumption that $\Delta\boldsymbol{\mu}_{\mathsf{A}}(\beta_{\mathsf{B}}^{\texttt{MaxUtil}}) < 0$:

Theorem 3.3 (`DemParity` can cause harm by being over-eager)**.** *Fix a selection rate* β*. Assume that* $\beta_{\mathsf{B}}^{\texttt{MaxUtil}} > \beta > \beta_{\mathsf{A}}^{\texttt{MaxUtil}}$*. Then, there exists a population proportion* g_0 *such that, for all* $g_{\mathsf{A}} \in [0, g_0]$*,* $\beta_{\mathsf{A}}^{\texttt{DemParity}} > \beta$*. In particular, when* $\beta = \beta_0$*,* `DemParity` *causes active harm, and when* $\beta = \overline{\beta}$*,* `DemParity` *causes relative harm.*

The assumption $\Delta\boldsymbol{\mu}_{\mathsf{A}}(\beta_{\mathsf{B}}^{\texttt{MaxUtil}}) < 0$ implies that a policy which selects individuals from group A at the selection rate that `MaxUtil` would have used for group B necessarily lowers average score in A. This is one natural notion of protected group A's 'disadvantage' relative to group B. In this case, `DemParity` penalizes the scores of group A even more than a naive `MaxUtil` policy, as long as group proportion g_{A} is small enough. Again, small g_{A} is another notion of group disadvantage.

Using credit scores as an example, Theorem 3.3 tells us that an overly aggressive fairness criterion will give too many loans to people in a protected group who cannot pay them back, hurting the group's credit scores on average. In the following theorem, we show that an analogous result holds for EqOpt.

Theorem 3.4 (EqOpt can cause harm by being over-eager)**.** *Suppose that* $\beta_{\mathsf{B}}^{\texttt{MaxUtil}} > G^{(\mathsf{A}\to\mathsf{B})}(\beta)$ *and* $\beta > \beta_{\mathsf{A}}^{\texttt{MaxUtil}}$. *Then, there exists a population proportion* g_0 *such that, for all* $g_{\mathsf{A}} \in [0, g_0]$, $\beta_{\mathsf{A}}^{\texttt{EqOpt}} > \beta$. *In particular, when* $\beta = \beta_0$, EqOpt *causes active harm, and when* $\beta = \overline{\beta}$, EqOpt *causes relative harm.*

We remark that in Theorem 3.4, we rely on the *transfer function*, $G^{(\mathsf{A}\to\mathsf{B})}$, which for every loan rate β in group A gives the loan rate in group B that has the same true positive rate. Notice that if $G^{(\mathsf{A}\to\mathsf{B})}$ were the identity function, Theorems 3.3 and Theorem 3.4 would be exactly the same. Indeed, our framework (detailed in Appendix E) unifies the analyses for a large class of fairness constraints that includes DemParity and EqOpt as specific cases, and allows us to derive results about impact on $\Delta\boldsymbol{\mu}$ using general techniques. In the next section, we present further results that compare the fairness criteria, demonstrating the usefulness of our technical framework.

3.2. Comparing EqOpt and DemParity

Our analysis of the acceptance rates of EqOpt and DemParity in Appendix C suggests that it is difficult to compare DemParity and EqOpt without knowing the full distributions $\boldsymbol{\pi}_{\mathsf{A}}, \boldsymbol{\pi}_{\mathsf{B}}$, which is necessary to compute the transfer function $G^{(\mathsf{A}\to\mathsf{B})}$. In fact, we have found that settings exist both in which DemParity causes harm while EqOpt causes improvement and in which DemParity causes improvement while EqOpt causes harm. There cannot be one general rule as to which fairness criteria provides better outcomes in all settings. We now present simple sufficient conditions on the geometry of the distributions for which EqOpt is always better than DemParity in terms of $\Delta\boldsymbol{\mu}_{\mathsf{A}}$.

Theorem 3.5 (EqOpt may avoid active harm where DemParity fails)**.** *Fix a selection rate* β. *Suppose* $\boldsymbol{\pi}_{\mathsf{A}}, \boldsymbol{\pi}_{\mathsf{B}}$ *are identical up to a translation with* $\boldsymbol{\mu}_{\mathsf{A}} < \boldsymbol{\mu}_{\mathsf{B}}$, *i.e.* $\boldsymbol{\pi}_{\mathsf{A}}(x) = \boldsymbol{\pi}_{\mathsf{B}}(x + (\boldsymbol{\mu}_{\mathsf{B}} - \boldsymbol{\mu}_{\mathsf{A}}))$. *For simplicity, take* $\boldsymbol{\rho}(x)$ *to be linear in* x. *Suppose*

$$\beta^0 > \sum_{x > \boldsymbol{\mu}_{\mathsf{A}}} \boldsymbol{\pi}_{\mathsf{A}}.$$

Then there exists an interval $[g_1, g_2] \subseteq [0, 1]$, *such that* $\forall g_{\mathsf{A}} > g_1$, $\beta^{\texttt{EqOpt}} < \beta$ *while* $\forall g_{\mathsf{A}} < g_2$, $\beta^{\texttt{DemParity}} > \beta$. *In particular, when* $\beta = \beta_0$, *this implies* DemParity *causes active harm but* EqOpt *causes improvement for* $g_{\mathsf{A}} \in [g_1, g_2]$, *but for any* g_{A} *such that* DemParity *causes improvement,* EqOpt *also causes improvement.*

To interpret the conditions under which Corollary 3.5 holds, consider when we might have $\beta_0 > \sum_{x > \boldsymbol{\mu}_{\mathsf{A}}} \boldsymbol{\pi}_{\mathsf{A}}$. This is precisely when $\Delta\boldsymbol{\mu}_{\mathsf{A}}(\sum_{x > \boldsymbol{\mu}_{\mathsf{A}}} \boldsymbol{\pi}_{\mathsf{A}}) > 0$, that is, $\Delta\boldsymbol{\mu}_{\mathsf{A}} > 0$ for a policy that selects every individual whose score is above the group A mean, which is reasonable in reality. Indeed, the converse would imply that group A has such low scores that even selecting all above average individuals in A would hurt the average score. In such a case, Corollary 3.5 suggests that EqOpt is better than DemParity at avoiding active harm, because it is more conservative. A natural question then is: can EqOpt cause relative harm by being too stingy?

Theorem 3.6 (DemParity never loans less than MaxUtil, but EqOpt might)**.** *Recall the definition of the TPR functions* ω_{j}, *and suppose that the* MaxUtil *policy* $\boldsymbol{\tau}^{\texttt{MaxUtil}}$ *is such that* $\beta_{\mathsf{A}}^{\texttt{MaxUtil}} < \beta_{\mathsf{B}}^{\texttt{MaxUtil}}$ *and* $\mathrm{TPR}_{\mathsf{A}}(\boldsymbol{\tau}^{\texttt{MaxUtil}}) > \mathrm{TPR}_{\mathsf{B}}(\boldsymbol{\tau}^{\texttt{MaxUtil}})$. *Then* $\beta_{\mathsf{A}}^{\texttt{EqOpt}} < \beta_{\mathsf{A}}^{\texttt{MaxUtil}} < \beta_{\mathsf{A}}^{\texttt{DemParity}}$. *That is,* EqOpt *causes relative harm by selecting at a rate lower than* MaxUtil.

The above theorem shows that DemParity is never stingier than MaxUtil to the protected group A, as long as a A is disadvantaged in the sense that MaxUtil selects a larger proportion of B than A. On the other hand, EqOpt can select less of group A than MaxUtil, and by definition, cause relative harm. This is a surprising result about EqOpt, and this phenomenon arises from high levels of in-group inequality for group A. Moreover, we show in Appendix F that there are parameter settings where the conditions in Theorem 3.6 are satisfied even under a stringent notion of disadvantage we call CDF domination, described therein.

4. Relaxations of Constrained Fairness

Regularized fairness: In many cases, it may be unrealistic for an institution to ensure that fairness constraints are met exactly. However, one can consider "soft" formulations of fairness constraints which either penalized the differences in acceptance rate (DemParity) or the differences in TPR (EqOpt). In Appendix E, we formulate these soft constraints as regularized objectives. For example, a soft-DemParity can be rendered as

$$\max_{\boldsymbol{\tau} := \boldsymbol{\tau}_{\mathsf{A}}, \boldsymbol{\tau}_{\mathsf{B}}} \mathcal{U}(\boldsymbol{\tau}) - \lambda\Phi(\langle \boldsymbol{\pi}_{\mathsf{A}}, \boldsymbol{\tau}_{\mathsf{A}} \rangle - \langle \boldsymbol{\pi}_{\mathsf{B}}, \boldsymbol{\tau}_{\mathsf{B}} \rangle)\,, \qquad (4)$$

where $\lambda > 0$ is a regularization parameter, and $\Phi(t)$ is a convex regularization function. We show that the solutions to these objectives are threshold policies, and can be fully characterized in terms of the group-wise selection rate. We also make rigorous the notion that policies which solve the soft-constraint objective interpolate between MaxUtil policies at $\lambda = 0$ and hard-constrained policies (DemParity or

EqOpt) as $\lambda \to \infty$. This fact is clearly demonstrated by the form of the solutions in the special case of the regularization function $\Phi(t) = |t|$, provided in the appendix.

Fairness under measurement error: Next, consider the implications of an institution with imperfect knowledge of scores. Under a simple model in which the estimate of an individual's score $X \sim \boldsymbol{\pi}$ is prone to errors $e(X)$ such that $X + e(X) := \widehat{X} \sim \widehat{\boldsymbol{\pi}}$. Constraining the error to be negative results in the setting that scores are systematically *underestimated*. In this setting, it is equivalent to consider the CDF of underestimated distribution $\widehat{\boldsymbol{\pi}}$ to be *dominated* by the CDF true distribution $\boldsymbol{\pi}$, that is $\sum_{x \geq c} \widehat{\boldsymbol{\pi}}(x) \leq \sum_{x \geq c} \boldsymbol{\pi}(x)$ for all $c \in [C]$. Then we can compare the institution's behavior under this estimation to its behavior under the truth.

Proposition 4.1 (Underestimation causes underselection)**.** *Fix the distribution of* B *as* $\boldsymbol{\pi}_{\mathsf{B}}$ *and let* β *be the acceptance rate of* A *when the institution makes the decision using perfect knowledge of the distribution* $\boldsymbol{\pi}_{\mathsf{A}}$*. Denote* $\widehat{\beta}$ *as the acceptance rate when the group is instead taken as* $\widehat{\boldsymbol{\pi}}_{\mathsf{A}}$*. Then* $\beta_{\mathsf{A}}^{\texttt{MaxUtil}} > \widehat{\beta}_{\mathsf{A}}^{\texttt{MaxUtil}}$ *and* $\beta_{\mathsf{A}}^{\texttt{DemParity}} > \widehat{\beta}_{\mathsf{A}}^{\texttt{DemParity}}$*. If the errors are further such that the true TPR dominates the estimated TPR, it is also true that* $\beta_{\mathsf{A}}^{\texttt{EqOpt}} > \widehat{\beta}_{\mathsf{A}}^{\texttt{EqOpt}}$.

Because fairness criteria encourage a higher selection rate for disadvantaged groups (Theorem 3.2), systematic underestimation widens the regime of their applicability. Furthermore, since the estimated MaxUtil policy underloans, the region for relative improvement in the outcome curve (Figure 1) is larger, corresponding to more regimes under which fairness criteria can yield favorable outcomes. Thus potential measurement error should be a factor when motivating these criteria.

Outcome-based alternative: As explained in the preceding sections, fairness criteria may actively harm disadvantaged groups. It is thus natural to consider a modified decision rule which involves the explicit maximization of $\Delta\boldsymbol{\mu}_{\mathsf{A}}$. In this case, imagine that the institution's primary goal is to aid the disadvantaged group, subject to a limited profit loss compared to the maximum possible expected profit $\mathcal{U}^{\texttt{MaxUtil}}$. The corresponding problem is as follows.

$$\max_{\boldsymbol{\tau}_{\mathsf{A}}} \Delta\boldsymbol{\mu}_{\mathsf{A}}(\boldsymbol{\tau}_{\mathsf{A}}) \text{ s.t. } \mathcal{U}_{\mathsf{A}}^{\texttt{MaxUtil}} - \mathcal{U}(\boldsymbol{\tau}) < \delta\,. \qquad (5)$$

Unlike the fairness constrained objective, this objective no longer depends on group B and instead depends on our model of the mean score change in group A, $\Delta\boldsymbol{\mu}_{\mathsf{A}}$.

Proposition 4.2 (Outcome-based solution)**.** *In the above setting, the optimal bank policy* $\boldsymbol{\tau}_{\mathsf{A}}$ *is a threshold policy with selection rate* $\beta = \min\{\beta^*, \beta^{\max}\}$*, where* β^* *is the outcome-optimal loan rate and* $\beta^{\max}$ *is the maximum loan rate under the bank's "budget".*

The above formulation's advantage over fairness constraints is that it directly optimizes the outcome of A and can be approximately implemented given reasonable ability to predict outcomes. Importantly, this objective shifts the focus to outcome modeling, highlighting the importance of domain specific knowledge. Future work can consider strategies that are robust to outcome model errors.

5. Simulations

We examine the outcomes induced by fairness constraints in the context of FICO scores for two race groups. FICO scores are a proprietary classifier widely used in the United States to predict credit worthiness. Our FICO data is based on a sample of 301,536 TransUnion TransRisk scores from 2003 (US Federal Reserve, 2007), preprocessed by Hardt et al. (2016). These scores, corresponding to x in our model, range from 300 to 850 and are meant to predict credit risk. Empirical data labeled by race allows us to estimate the distributions π_{j}, where j represents race, which is restricted to two values: white non-Hispanic (labeled "white" in figures), and black. Using national demographic data, we set the population proportions to be 18% and 82%.

Individuals were labeled as defaulted if they failed to pay a debt for at least 90 days on at least one account in the ensuing 18-24 month period; we use this data to estimate the success probability given score, $\boldsymbol{\rho}_{\mathsf{j}}(x)$, which we allow to vary by group to match the empirical data. Our outcome curve framework allows for this relaxation; however, this discrepancy can also be attributed to group-dependent mismeasurement of score, and adjusting the scores accordingly would allow for a single $\boldsymbol{\rho}(x)$. We use the success probabilities to define the affine utility and score change functions defined in Example 2.1. We model individual penalties as a score drop of $c_- = -150$ in the case of a default, and in increase of $c_+ = 75$ in the case of successful repayment.

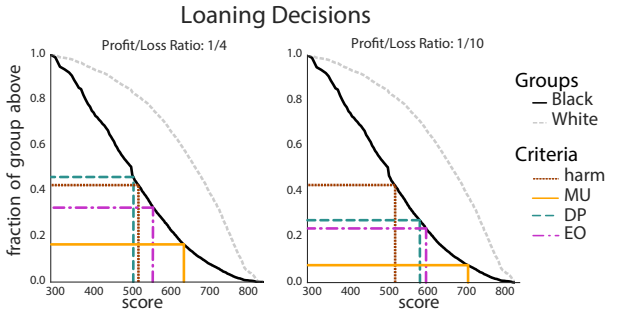

Figure 3. The empirical CDFs of both groups are plotted along with the decision thresholds resulting from MaxUtil, DemParity, and EqOpt for a model with bank utilities set to (a) $\frac{u_-}{u_+} = -4$ and (b) $\frac{u_-}{u_+} = -10$. The threshold for active harm is displayed; in (a) DemParity causes active harm while in (b) it does not. EqOpt and MaxUtil never cause active harm.

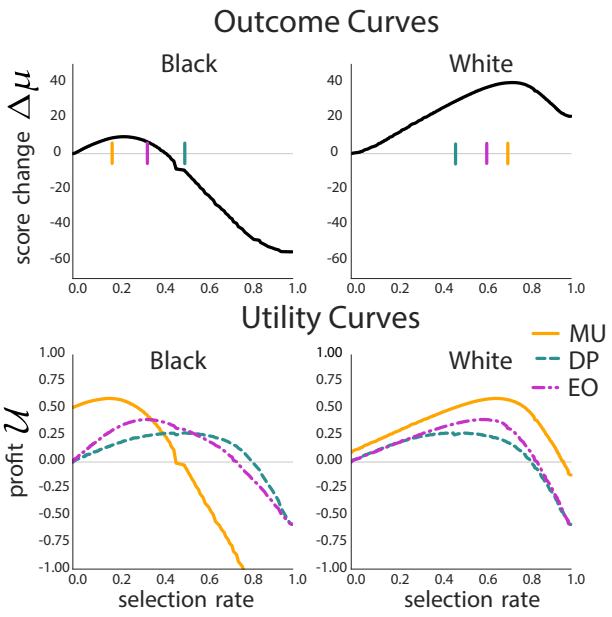

Figure 4. The outcome and utility curves are plotted for both groups against the group selection rates. The relative positions of the utility maxima determine the position of the decision rule thresholds. We hold $\frac{u_-}{u_+} = -4$ as fixed.

In Figure 3, we display the empirical CDFs along with selection rates resulting from different loaning strategies for two different settings of bank utilities. In the case that the bank experiences a loss/profit ratio of $\frac{u_-}{u_+} = -10$, no fairness criteria surpass the active harm rate β_0; however, in the case of $\frac{u_-}{u_+} = -4$, `DemParity` overloans, in line with the statement in Theorem 3.3.

These results are further examined in Figure 4, which displays the normalized outcome curves and the utility curves for both the white and the black group. To plot the `MaxUtil` utility curves, the group that is not on display has selection rate fixed at $\beta^{\texttt{MaxUtil}}$. In this figure, the top panel corresponds to the average change in credit scores for each group under different loaning rates β; the bottom panels shows the corresponding *total* utility $\mathcal{U}$ (summed over both groups and weighted by group population sizes) for the bank.

Figure 4 highlights that the position of the utility optima in the lower panel determines the loan (selection) rates. In this specific instance, the utility and change ratios are fairly close, $\frac{u_-}{u_+} = -4$ and $\frac{c_-}{c_+} = -2$, meaning that the bank's profit motivations align with individual outcomes to some extent. Here, we can see that `EqOpt` loans much closer to optimal than `DemParity`, similar to the setting suggested by Theorem 3.2.

Although one might hope for decisions made under fairness constraints to positively affect the black group, we observe the opposite behavior. The `MaxUtil` policy (solid orange line) and the `EqOpt` policy result in similar expected credit score change for the black group. However, `DemParity` (dashed green line) causes a negative expected credit score change in the black group, corresponding to active harm. For the white group, the bank utility curve has almost the same shape under the fairness criteria as it does under `MaxUtil`, the main difference being that fairness criteria lowers the total expected profit from this group.

This behavior stems from a discrepancy in the outcome and profit curves for each population. While incentives for the bank and positive results for individuals are somewhat aligned for the majority group, under fairness constraints, they are more heavily misaligned in the minority group, as seen in graphs (left) in Figure 4. We remark that in other settings where the *unconstrained* profit maximization is misaligned with individual outcomes (e.g., when $\frac{u_-}{u_+} = -10$), fairness criteria may perform more favorably for the minority group by pulling the utility curve into a shape consistent with the outcome curve.

By analyzing the resulting effects of `MaxUtil`, `DemParity`, and `EqOpt` on actual credit score lending data, we show the applicability of our model to real-world applications. In particular, results shown in Section 3 hold empirically for the FICO TransUnion TransRisk scores.

6. Conclusion and Future Work

We argue that without a careful model of delayed outcomes, we cannot foresee the impact a fairness criterion would have if enforced as a constraint on a classification system. However, if such an accurate outcome model is available, we show that there are more direct ways to optimize for positive outcomes than via existing fairness criteria.

Our formal framework exposes a concise, yet expressive way to model outcomes via the expected change in a variable of interest caused by an institutional decision. This leads to the natural concept of an outcome curve that allows us to interpret and compare solutions effectively. In essence, the formalism we propose requires us to understand the two-variable causal mechanism that translates decisions to outcomes. Depending on the application, such an understanding might necessitate greater domain knowledge and additional research into the specifics of the application. This is consistent with much scholarship that points to the context-sensitive nature of fairness in machine learning.

An interesting direction for future work is to consider other characteristics of impact beyond the change in population *mean*. Variance and individual-level outcomes are natural and important considerations. Moreover, it would be interesting to understand the robustness of outcome optimization to modeling and measurement errors.

Acknowledgements

We thank Lily Hu, Aaron Roth, and Cathy O'Neil for discussions and feedback on an earlier version of the manuscript. We thank the students of CS294: Fairness in Machine Learning (Fall 2017, University of California, Berkeley) for inspiring class discussions and comments on a presentation that was a precursor of this work. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1752814.

References

Barocas, Solon and Selbst, Andrew D. Big data's disparate impact. *California Law Review*, 104, 2016.

Calders, Toon, Kamiran, Faisal, and Pechenizkiy, Mykola. Building classifiers with independency constraints. In *Proc. IEEE ICDMW*, ICDMW '09, pp. 13–18, 2009.

Chouldechova, Alexandra. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *FATML*, 2016.

Coate, Stephen and Loury, Glenn. Will affirmative-action policies eliminate negative stereotypes? 83:1220–40, 02 1993.

Ensign, Danielle, Friedler, Sorelle A, Neville, Scott, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Runaway feedback loops in predictive policing. *arXiv preprint arXiv:1706.09847*, 2017.

Executive Office of the President. Big data: A report on algorithmic systems, opportunity, and civil rights. Technical report, White House, May 2016.

Foster, Dean P and Vohra, Rakesh V. An economic argument for affirmative action. *Rationality and Society*, 4(2):176–188, 1992.

Fuster, Andreas, Goldsmith-Pinkham, Paul, Ramadorai, Tarun, and Walther, Ansgar. Predictably unequal? the effects of machine learning on credit markets. *SSRN*, 2017.

Hardt, Moritz, Price, Eric, and Srebro, Nati. Equality of opportunity in supervised learning. In *Proc.* 30*th NIPS*, 2016.

Hu, Lily and Chen, Yiling. A short-term intervention for long-term fairness in the labor market. In *Proc.* 27*th WWW*, 2018.

Joseph, Matthew, Kearns, Michael, Morgenstern, Jamie H, and Roth, Aaron. Fairness in learning: Classic and contextual bandits. In *Proc.* 30*th NIPS*, pp. 325–333, 2016.

Kalev, Alexandra, Dobbin, Frank, and Kelly, Erin. Best Practices or Best Guesses? Assessing the Efficacy of Corporate Affirmative Action and Diversity Policies. *American Sociological Review*, 71(4):589–617, 2006.

Keith, Stephen N., Bell, Robert M., Swanson, August G., and Williams, Albert P. Effects of affirmative action in medical schools. *New England Journal of Medicine*, 313 (24):1519–1525, 1985.

Kleinberg, Jon M., Mullainathan, Sendhil, and Raghavan, Manish. Inherent trade-offs in the fair determination of risk scores. *Proc.* 8*th ITCS*, 2017.

Knowles, John, Persico, Nicola, and Todd, Petra. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1):203–229, 2001.

Pleiss, Geoff, Raghavan, Manish, Wu, Felix, Kleinberg, Jon, and Weinberger, Kilian Q. On fairness and calibration. In *Advances in Neural Information Processing Systems 30*, pp. 5684–5693, 2017.

Ross, Stephen and Yinger, John. *The Color of Credit: Mortgage Discrimination, Research Methodology, and Fair-Lending Enforcement*. MIT Press, Cambridge, 2006.

US Federal Reserve. Report to the congress on credit scoring and its effects on the availability and affordability of credit, 2007.

Zafar, Muhammad Bilal, Valera, Isabel, Rogriguez, Manuel Gomez, and Gummadi, Krishna P. Fairness Constraints: Mechanisms for Fair Classification. In *Proc.* 20*th AISTATS*, pp. 962–970. PMLR, 2017.