Does mitigating ML's impact disparity require treatment disparity?

Zachary C. Lipton, Alexandra Chouldechova, Julian McAuley
Carnegie Mellon University
University of California, San Diego
zlipton@cmu.edu, achould@cmu.edu, jmcauley@cs.ucsd.edu

March 2, 2018

Abstract

Following related work in law and policy, two notions of disparity have come to shape the study of fairness in algorithmic decision-making. Algorithms exhibit *treatment disparity* if they formally treat members of protected subgroups differently; algorithms exhibit *impact disparity* when outcomes differ across subgroups, even if the correlation arises unintentionally. Naturally, we can achieve impact parity through purposeful treatment *dis*parity. In one thread of technical work, papers aim to reconcile the two forms of parity proposing *disparate learning processes* (DLPs). Here, the learning algorithm can see group membership during training but produce a classifier that is *group-blind* at test time. **In this paper**, we show theoretically that: (i) When other features correlate to group membership, DLPs *will* (indirectly) implement treatment disparity, undermining the policy desiderata they are designed to address; and (ii) When group membership is *partly* revealed by other features, DLPs induce within-class discrimination; and (iii) In general, DLPs provide a suboptimal trade-off between accuracy and impact parity. Based on our technical analysis, we argue that *transparent* treatment disparity is preferable to occluded methods for achieving impact parity. Experimental results on several real-world datasets highlight the practical consequences of applying DLPs vs. per-group thresholds.

1 Introduction

Effective decision-making requires decision-makers to distinguish between options given the available *features*. That much is unavoidable, unless we wish to make trivial decisions. In selection processes, such as hiring, university admissions, and loan approval, the options are *people*; the available features *include* (but are rarely limited to) direct evidence of qualifications; and the decisions, either *positive* or *negative*, consequentially impact lives.

Laws in many countries restrict the ways in which certain decisions can be made. In the United States, Title VII of the Civil Rights Act of 1964 [Civ, 1964], forbids employment decisions that discriminate on the basis of the following *protected characteristics*: *race, color, religion, sex,* and *national origin*. The interpretation of this law has led to two widely-referenced notions of discrimination: *disparate treatment* and *disparate impact*.

Disparate treatment addresses intentional discrimination. This includes: (i) decisions explicitly based on a protected characteristic; and (ii) intentional discrimination via proxy variables. For example, in the 1900s, literacy tests for voting eligibility were employed to disenfranchise racial minorities.

Disparate impact addresses facially neutral practices that might nevertheless have an "unjustified adverse impact on members of a protected class" [Civ, 1964]. Absent intentional discrimination, unequal outcomes can emerge due to correlations between protected and unprotected characteristics. For example, black defendants are sentenced to death more frequently than white defendants for the same crimes [Ford, 2014]. While this likely owes significantly to the racial biases of judges and juries, it also might owe, in part, to the correlation between race and wealth, and by extension, access to legal services. Unequal outcomes may not always signal unlawful discrimination. For example, the over-representation of Asian students in prestigious US colleges does not appear to entail pro-Asian discrimination. On the contrary, investigative reports suggest that it arises despite admissions policies that set higher bars for Asian applicants [Hartocollis and Saul, 2017].

Recently, owing to the increased use of machine learning (ML) to make (or assist in) consequential decisions, the topic of quantifying and mitigating ML-based discrimination has attracted interest among both practitioners and academics in both policy and ML. However, while the existing legal doctrine offers qualitative ideas expressed in prose, intervention in an ML-based system requires more concrete mathematical formalism.

Loosely inspired by the relevant legal concepts, technical papers have proposed several criteria to quantify discrimination. One criterion requires that the fraction given a positive decision be equal across different groups. Another criterion states that a classifier should be blind to the protected characteristic. Within the technical literature, these criteria are commonly referred to as *disparate impact* and *disparate treatment*, respectively.

In this paper, we will call these technical criteria *impact parity* and *treatment parity*, to distinguish them from their legal antecedents. The distinction between technical and legal terminology is important to maintain. While impact parity and treatment parity are inspired by legal concepts, **we contend that** technical approaches that achieve these criteria may fail to satisfy the underlying legal and ethical desiderata of the motivating policy problem.

We demonstrate how this disconnect arises in the context of a class of algorithms that we denote Disparate Learning Processes (DLPs). DLPs are methods for simultaneously satisfying treatment- and impact-parity criteria [Pedreshi et al., 2008, Kamishima et al., 2011, Zafar et al., 2017a]. DLPs operate according to the following principle: *The protected characteristic may be used during training, but is not available to the model at prediction time.* In the earliest such approach, Pedreshi et al. [2008] use the protected characteristic to winnow the set of acceptable rules from an expert system. In other papers, the protected characteristic is incorporated into the learning objective as either a regularizer or constraint or is used in preprocessing the training data [Kamiran and Calders, 2009, Kamiran et al., 2010, Zafar et al., 2017a].

These approaches are grounded in the premise that DLPs are acceptable in cases where using a protected characteristic as a direct input to the model would constitute *disparate treatment* and thus be impermissible. We call this premise into question on the following grounds:

1. When protected characteristics are redundantly encoded in the other features, sufficiently powerful DLPs can (indirectly) implement any form of treatment disparity.

2. When protected characteristics are partially encoded DLPs induce within-class discrimination based on irrelevant features, and can harm some members of the protected group.
3. While disparate treatment is by definition illegal, the legal status of treatment disparity is a subject of debate [Kim, 2017].
4. DLPs provide a suboptimal trade-off between accuracy and impact parity. The optimal way to trade off the two is to apply per-group thresholds, effecting treatment disparity.

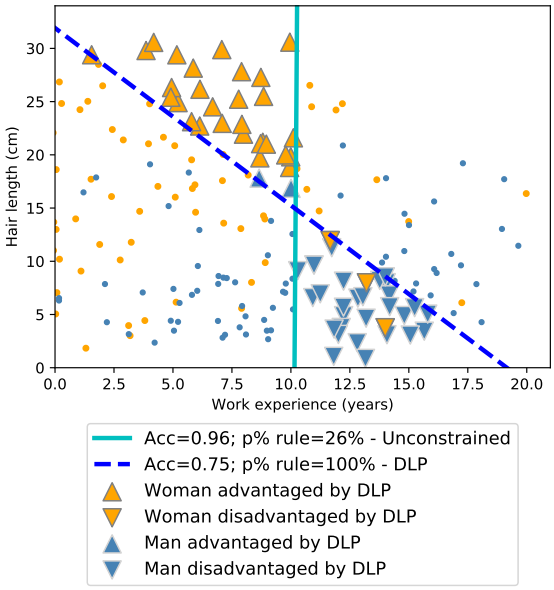

Figure 1: Demonstration of a DLP's undesirable side effects on a simple example of hiring data (see §4.1). Unconstrained classifier (vertical line) hires candidates based on work experience, yielding higher hiring rates for men than for women. A DLP (dashed diagonal) achieves parity by differentiating based on an irrelevant attribute (hair length). The DLP *hurts* some short-haired women, flipping their decisions to reject, and helps some long-haired men.

The capacity of DLPs to carry out treatment disparity *indirectly* casts doubt on whether, under the law, they would be viewed differently from approaches that *directly* apply treatment disparity. A recent California Law Review paper [Grimmelmann and Westreich, 2017] supports this view, illustrating their arguments in the "law-school hypothetical state" of Zootopia, where the protected groups are species.

> In our view, Title VII does not permit an employer to do indirectly what it could not do directly. An employer that explicitly selects applicants on the basis of species violates Title VII under a disparate treatment theory, regardless of whether species is correlated with job performance, and regardless of whether it bears animus against particular species. It is the selection "on the basis of" species that is the problem. An employer that uses home address to infer applicants' species and then selects applicants from particular

species does exactly the same, only in two steps rather than one. This too is a form of disparate treatment.

Organization The rest of this paper lays out our main arguments on theoretical (§2 & §3), empirical (§4), and qualitative (§6) grounds.

2 Disparate Learning Processes

To begin our formal description of the prior work, we'll introduce some notation. A dataset consists of n *examples*, or *data points* $\{\mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, each consisting of a feature vector $\mathbf{x}_i$ and a label y_i. A supervised learning algorithm $f : \mathcal{X}^n \times \mathcal{Y}^n \to (\mathcal{X} \to [0, 1])$ is a mapping from datasets to models. The learning algorithm produces a model $\hat{y} : \mathcal{X} \to \mathcal{Y}$, which given a feature vector $\mathbf{x}_i$, predicts the corresponding output y_i. In this discussion, we'll focus on binary classification, the setting in which the label y takes values from the set $\mathcal{Y} = \{0, 1\}$.

This paper considers probabilistic classifiers, which produce estimates $\hat{p}(\mathbf{x})$ of the conditional probability $\mathbb{P}(y = 1 \mid \mathbf{x})$ of the label given a feature vector $\mathbf{x}$. To make a prediction $\hat{y}(\mathbf{x}) \in \mathcal{Y}$ given an estimated probability $\hat{p}(\mathbf{x})$ a threshold rule is used: $\hat{y}_i = 1$ iff $\hat{p}_i > t$. The optimal choice of the threshold t depends on the performance metric being optimized. In our theoretical analysis we consider optimizing the *immediate utility*, as introduced in Corbett-Davies et al. [2017], of which classification accuracy (expected $0-1$ loss) is a special case. We will define this metric more precisely in the next section.

In formal descriptions of discrimination-aware ML, a dataset possesses a protected feature $z_i \in \mathcal{Z}$, making each example a three-tuple $(\mathbf{x}_i, y_i, z_i)$. The protected characteristic may be real-valued, like age, or categorical, like race or gender. The goal of many methods in discrimination-aware ML is not only to maximize accuracy, but also to ensure some form of impact parity. Following related work, we consider binary protected features that divide the set of examples into two groups a and b. Our analysis extends directly to settings with more than two groups.

Of the various measures of impact disparity that have been proposed, the two that are the most relevant here are the Calders-Verwer gap and the p-% rule. At a given threshold t, let $q_z = \frac{1}{n_z} \sum_{i:z_i=z} \mathbb{1}(\hat{p}_i > t)$, where $n_z = \sum_i^n \mathbb{1}(z_i = z)$. The **Calders-Verwer (CV) gap**,

$$CV = q_a - q_b,$$

is the difference between the proportions assigned to the positive class in the advantaged group (a) and the disadvantaged group (b) [Kamishima et al., 2011]. The p-% rule, as described in Zafar et al. [2017a], is a closely related metric that measures impact disparity as q_b/q_a. A classification is said to satisfy the **p-% rule** if $q_b/q_a \geq p/100$. This metric is motivated by a text on fair employment practices [Biddle, 2006], in which it is stated that cases where the ratio q_b/q_a is below 0.8 are problematic.

Many papers in discrimination-aware ML propose to optimize accuracy (or some other risk) subject to constraints on the resulting level of impact parity as assessed by some metric [Pedreshi et al., 2008, Kamiran et al., 2010, Dwork et al., 2017, Bechavod and Ligett, 2017, Hardt et al., 2016, Ritov et al., 2017]. Papers proposing DLPs [Pedreshi et al., 2008, Kamiran and Calders, 2009, Zafar et al., 2017a] take as a premise that using the protected feature z as a model input is impermissible

in this effort, as it amounts to *treatment disparity*. Discarding protected features, however, does not guarantee impact parity. As discussed in Dwork et al. [2012], even if the protected features are discarded, the model may still produce classifications that are correlated with z. Instead of discarding the protected features, DLPs incorporate z in the learning algorithm, but not in the classifier itself. Formally, a DLP is a learning algorithm described by the following mapping:

$$DLP : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \rightarrow (\mathcal{X} \rightarrow \mathcal{Y}).$$

Since z is not a direct input to the resulting model, DLPs achieve treatment parity. DLP papers suggest that, as a result, such models have better legal standing (vis-a-vis disparate treatment) than a model that uses z directly. As we show next (§3), under some circumstances, DLPs can (indirectly) realize any function achievable through treatment disparity. Moreover, as discussed in §1, certain recent legal scholarship suggests that DLPs, while satisfying the technical criterion of treatment parity, may nevertheless run afoul of disparate treatment [Grimmelmann and Westreich, 2017].

3 Theoretical Analysis

In this section we introduce our theoretical arguments. We present a set of simple theoretical results that demonstrate the optimality of treatment disparity, and highlight properties of DLPs. Our optimality results are all derived in the population or *infinite data* setting, where we assume knowledge of the true conditional probability function $p_{Y|X,Z}(\mathbf{x}, z) \equiv \mathbb{P}(Y = 1 \mid X = \mathbf{x}, Z = z)$. The main results of this section can be summarized as follows:

1. Direct treatment disparity on the basis of z is the optimal strategy for maximizing classification accuracy[1] subject to CV and p-% constraints.
2. When X fully encodes Z, a sufficiently powerful DLP is equivalent to treatment disparity.

In the next section (§4), we empirically demonstrate a related point:

(3) When X only partially encodes Z, a DLP may be suboptimal and can induce intra-group disparity on the basis of otherwise irrelevant features correlated with Z.

Treatment disparity is optimal Absent impact parity constraints, the Bayes-optimal decision rule for minimizing expected $0-1$ loss (i.e., maximizing accuracy) is given by

$$d^*_{\text{uncon}}(\mathbf{x}, z) = \begin{cases} 1 & p_{Y|X,Z}(\mathbf{x}, z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}.$$

In this section, we show that the optimal decision rule in the CV and p-% constrained problems has a similar form. The optimal decision rule will again be based on thresholding $p_{Y|X,Z}(\mathbf{x}, z)$, but at *group-specific thresholds*.

[1] Our results are all presented in terms of a more general performance metric, of which classification accuracy is a special case.

These rules can be thought of as operationalizing the following mechanism: Suppose that we start with the classifications of the unconstrained rule $d^*_{\text{uncon}}(\mathbf{x}, z)$, and this results in a CV gap of $q_a - q_b > \gamma$. To reduce the CV gap to γ we have two mechanisms: (i) We can flip predictions of cases in group b from 0 to 1, and (ii) we can flip predictions of cases in group a from 1 to 0. The optimal strategy is to perform these flips on group b cases that have the highest value of $p_{Y|X,Z}(\mathbf{x}, z)$ and group a cases that have the lowest value of $p_{Y|X,Z}(\mathbf{x}, z)$.

The results in this section adapt the work of Corbett-Davies et al. [2017], who establish optimal decision rules d under different kinds of fairness-related constraints. In that work, the authors characterize the optimal decision rule $d = d(\mathbf{x}, z)$ that maximizes the *immediate utility* $u(d, c) = \mathbb{E}[Yd(X, Z) - cd(X, Z)]$ for $(0 < c < 1)$, under different parity criteria. We begin with a lemma showing that expected classification accuracy has the functional form of an immediate utility function.

Lemma 1. *Optimizing classification accuracy is equivalent to optimizing immediate utility with* $c = 0.5$.

Proof. The expected accuracy of a binary decision rule $d(X)$ can be written as $\mathbb{E}[Yd(X) + (1 - Y)(1 - d(X))]$. Expanding and rearranging this expression gives

$$\begin{aligned}\mathbb{E}[Yd(X) + (1 - Y)(1 - d(X))] &= \mathbb{E}(2Yd(X) - d(X)) + \mathbb{E}(Y) + 1 \\ &= 2u(d, 0.5) + \mathbb{E}(Y) + 1.\end{aligned}$$

The only term in this expression that depends on d is the immediate utility, u. Thus the decision rule that maximizes u also maximizes accuracy. □

Before proceeding to the main technical results, we note that the results in this section are closely related to the concurrent independent work of Menon and Williamson [2018]. In their paper, the authors derive Bayes-optimal decision rules under the same parity constraints we consider here, working instead with the *cost-sensitive risk*,

$$\text{CS}(d; c) = \pi(1 - c)\text{FNR}(d) + (1 - \pi)c\text{FPR}(d),$$

where $\pi = \mathbb{P}(Y = 1)$. One can show that $u(d, c) = -\text{CS}(d; c) + \pi(1 - c)$, and hence the problem of maximizing immediate utility considered here is equivalent to minimizing cost-sensitive risk as in Menon and Williamson [2018]. In our case, it will be more convenient to work with the immediate utility.

For the next set of results, we follow Corbett-Davies et al. [2017] and assume that $p_{Y|X,Z}(X, Z)$, viewed as a random variable, has positive density on $[0, 1]$. This ensures that the optimal rules are unique and deterministic by disallowing point-masses of probability that would necessitate tie-breaking among observations with equal probability. The first result that we state is a direct corollary of two results in Corbett-Davies et al. [2017]. It considers the case where we desire exact parity, i.e., that $q_a = q_b$.

Corollary 2. *The optimal decision rules* d^* *under various parity constraints have the following form and are unique up to a set of probability zero:*

1. *Among rules satisfying statistical parity (the 100% rule), the optimum is*

$$d^*(\mathbf{x}, z) = \begin{cases} 1 & p_{Y|X,Z}(\mathbf{x}, z) \geq t_z \\ 0 & \textit{otherwise} \end{cases},$$

where $t_z \in [0,1]$ are constants that depend only on group membership z.

2. *Among rules that have equal false positive rates across groups, the optimum is*

$$d^*(\mathbf{x}, z) = \begin{cases} 1 & p_{Y|X,Z}(\mathbf{x}, z) \geq s_z \\ 0 & otherwise \end{cases},$$

where s_z are constants that depend only on group membership z (but are different from t_z).

3. *(1) and (2) continue to hold even in the resource-constrained setting where the overall proportion of cases classified as positive is constrained.*

Proof. (1) and (2) are direct corollaries of Lemma 1 combined with Theorem 3.2 and Prop 3.3 of Corbett-Davies et al. [2017]. □

The next set of results establishes optimality under general p-% and CV rules.

Proposition 3. *Under the same assumptions as above, the optimum among rules that satisfy the CV constraint $0 \leq q_a - q_b < \gamma$ or the p-% rule also has the form*

$$d^*(\mathbf{x}, z) = \begin{cases} 1 & p_{Y|X,Z}(\mathbf{x}, z) \geq t_z \\ 0 & otherwise \end{cases},$$

where $t_z \in [0,1]$ are constants that depend on the group membership z, and on the choice of constraint parameter γ or p. The thresholds t_z are different for the CV constraint and p-% rule.

Proof. Suppose that the optimal solution under the CV or p-% rule constraint classifies proportions q_a and q_b of the advantaged and disadvantaged groups, respectively, to the positive class. As shown in Corbett-Davies et al. [2017], we can rewrite the immediate utility as

$$u(d, 0.5) = \mathbb{E}[d(X, Z)(p_{Y|X,Z} - 0.5)].$$

From this expression, it is clear that the utility will be maximized precisely when $d^*(X, Z) = 1$ for the q_z proportion of individuals in each group that have the highest values of $p_{Y|X,Z}$. Since the optimal values of q_z may differ between the CV-constrained solution and the p-% solution, the optimal thresholds may differ as well. □

The final result in this section shows that a decision rule that does not directly use z as an input variable or for determining the thresholds will have lower accuracy than the optimal rule that uses this information. That is, we show that DLPs are suboptimal for trading off between accuracy and impact parity.

Theorem 4. *Let $d^*(\mathbf{x}, z)$ be the optimal decision rule under a the CV-γ or p-% constraint. Let $d_{DLP}(\mathbf{x})$ be the optimal solution to a DLP. If $d(\mathbf{x}, z)$ and $d_{DLP}(\mathbf{x})$ satisfy CV or p-% constraints with the same q_a and q_b, the DLP solution results in lower or equal accuracy. (equal only if the solutions are the same.)*

Proof. From Proposition 3, we know that the unique accuracy-optimizing solution is given by

$$d^*(\mathbf{x}, z) = \begin{cases} 1 & p_{Y|X,Z}(\mathbf{x}, z) \geq t_z \\ 0 & \text{otherwise} \end{cases},$$

where t_z is the 1 - q_z quantile of $p_{Y|X,Z}$. The difference in immediate utility between the two decision rules can be expressed as follows:

$$\begin{aligned}
&\mathbb{E}[d^*(X,Z)(p_{Y|X,Z} - 0.5)] - \mathbb{E}[d_{DLP}(X)(p_{Y|X,Z} - 0.5)] \\
&= \mathbb{E}[(d^*(X,Z) - d_{DLP}(X))(p_{Y|X,Z} - 0.5)] \\
&= \mathbb{E}[p_{Y|X,Z} - 0.5 \mid d^* = 1, d_{DLP} = 0]\mathbb{P}(d^* = 1, d_{DLP} = 0) \\
&\quad - \mathbb{E}[p_{Y|X,Z} - 0.5 \mid d^* = 0, d_{DLP} = 1]\mathbb{P}(d^* = 0, d_{DLP} = 1) \\
&= \big(\mathbb{E}[p_{Y|X,Z} - 0.5 \mid d^* = 1, d_{DLP} = 0] \\
&\quad - \mathbb{E}[p_{Y|X,Z} - 0.5 \mid d^* = 0, d_{DLP} = 1]\big)\mathbb{P}(d^* = 1, d_{DLP} = 0) \\
&\geq 0
\end{aligned}$$

The final inequality follows from the observation that $d^*(X, Z) = 1$ for the highest values of $p_{Y|X,Z}$, so $p_{Y|X,Z}$ is stochastically greater on the event $\{d^* = 1, d_{DLP} = 0\}$ than on $\{d^* = 0, d_{DLP} = 1\}$. Note that equality holds only if $\mathbb{P}(d^* = 1, d_{DLP} = 0) = 0$, i.e., if the two rules are equivalent with probability 1. □

All of the results in this section continue to hold under "do no harm" constraints, where we require that any individual in the disadvantaged group who was classified as positive under the unconstrained rule $d_{\text{uncons}}(\mathbf{x}, z)$ remains positively classified. This corresponds to the setting where the proportion of cases in the disadvantaged group classified as positive is constrained to be no lower than the proportion under the unconstrained rule $d_{\text{uncons}}(\mathbf{x}, z)$ (or no lower than some fixed value $q_a^{\min}$). Such constraint impose an upper bound on the optimal thresholds t_b, but do not change the structure of the optimal rules.

Functional equivalence when protected characteristic is redundantly encoded Consider the case where the protected feature z is redundantly encoded in the other features $\mathbf{x}$. More precisely, suppose that there exists a known subcomputation g such that $z = g(\mathbf{x})$. This allows for any function of the data $f(\mathbf{x}, z)$ to be represented as a function of $\mathbf{x}$ alone via $\tilde{f}(\mathbf{x}) = f(\mathbf{x}, g(\mathbf{x}))$. While it remains the case that $\tilde{f}(\mathbf{x})$ does not directly use z as an input variable—and thus satisfies treatment parity—$\tilde{f}$ should be no less legally suspect from a *disparate treatment* perspective than the original function f that uses z directly. The main difference for the purpose of our discussion is that $\tilde{f}$, resulting from a DLP, may technically satisfy treatment parity, while f does not.

Within-class discrimination when protected characteristic is partially redundantly encoded When the protected characteristic is partially encoded in the other features, disparate treatment may induce within-class discrimination by applying the benefit of the affirmative action unevenly, and can even harm some members of the protected class. In the following section, we demonstrate this phenomenon empirically using synthetic data, university admissions data, and several public datasets. The ease of producing such examples might convince the reader that the highly varied effects of intervention with a DLP on members of the disadvantaged group raise serious questions about the usefulness of DLPs.

4 Empirical Analysis

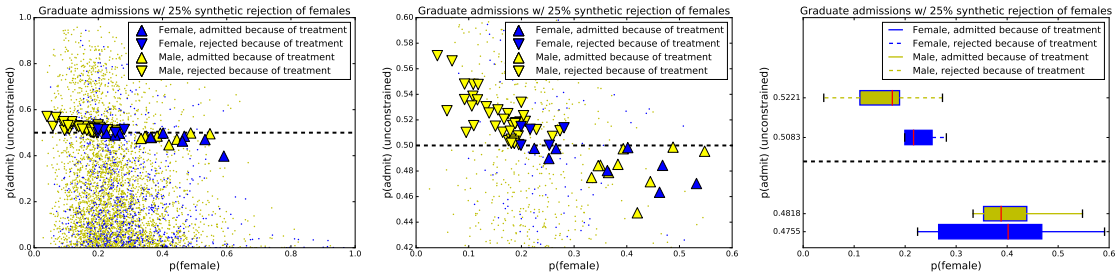

Figure 2: (left) probability of the sensitive variable versus (unconstrained) admission probability, on unseen test data. Downward triangles indicate individuals who are rejected *only* after applying the DLP, while upward triangles indicate individuals accepted *only* by the DLP. The remaining ∼4,000 blue/yellow dots indicate people whose decisions are not altered. Note that most students benefiting from the DLP are males who 'look like' females based on other features, whereas females who 'look like' males are hurt by the DLP. Detail view (center) and summary statistics (right) of the same plot.

This preceding analysis demonstrates several theoretical advantages to increasing impact parity via treatment disparity:

- **Optimality:** As demonstrated for CV score and for p-% rule, intervention via per-group thresholds maximizes accuracy subject to an impact parity constraint.
- **Rational ordering:** Within each group, individuals with higher probability of belonging to the positive class are always assigned to the positive class ahead of those with lower probabilities.
- **Does no harm to the protected group:** The treatment disparity intervention can be constrained to only benefit members of the disadvantaged class.

In order to avoid treatment disparity, DLPs attempt to produce a classifier that satisfies the parity constraints, by relying upon the proxy features to satisfy the parity metric. Typically, this is accomplished either by introducing constraints to a convex optimization problem, or by adding a regularization term and tuning the corresponding hyper-parameter. Because the CV score and p-% rule are non-convex in model parameters (scores only change when a point crosses the decision boundary), Kamishima et al. [2011], Zafar et al. [2017a] introduce convex surrogates aimed at reducing the correlation between the sensitive feature and the prediction.

These approaches presume that the proxy variables contain information about the sensitive attribute. Otherwise, the parity could only be satisfied via a trivial solution (e.g. assign either *everyone* or *nobody* to the positive class). So we must consider two scenarios: (i) the proxy variables $\mathbf{x}$ fully encode z, in which case, a sufficiently powerful DLP will implicitly reconstruct z, because this gives the optimal solution to the impact-constrained objective; and (ii) $\mathbf{x}$ doesn't fully capture z, in which case the DLP may be sub-optimal, violate rational ordering within groups, and harm members of the disadvantaged group.

4.1 Synthetic data example: work experience and hair length in hiring

To begin, we illustrate our arguments empirically with a simple synthetic data experiment. To construct the data, we sample $n_{\text{all}} = 2000$ total observations from the data-generating process described below. 70% of the observations are used for training, and the remaining 30% are reserved for model testing.

$$
\begin{aligned}
z_i &\sim \text{Bernoulli}(0.5) \\
\text{hair\_length}_i \mid z_i = 1 &\sim 35 \cdot \text{Beta}(2, 2) \\
\text{hair\_length}_i \mid z_i = 0 &\sim 35 \cdot \text{Beta}(2, 7) \\
\text{work\_exp}_i \mid z_i &\sim \text{Poisson}(25 + 6z_i) - \text{Normal}(20, \sigma = 0.2) \\
y_i \mid \text{work\_exp} &\sim 2 \cdot \text{Bernoulli}(p_i) - 1, \text{ where} \\
p_i &= 1/\left(1 + \exp[-(-25.5 + 2.5\text{work\_exp})]\right)
\end{aligned}
$$

This data-generating process has the following key properties: (i) the historical hiring process was based solely on the number of years of work experience; (ii) because women on average have fewer years of work experience than men (5 years vs. 11), men have been hired at a much higher rate than women; and (iii) women have longer hair than men, a fact that was irrelevant to historical hiring practice.

Figure 1 shows the test set results of applying a DLP to the available historical data to equalize hiring rates between men and women. We apply the DLP proposed by Zafar et al. [2017a], using code available from the authors.[2] While the DLP successfully equalizes hiring rates (satisfying a 100-% rule), it does so through a problematic within-class discrimination mechanism. The DLP rule advantages individuals with longer hair over those with shorter hair and considerably longer work experience. We find that several women who would have been hired under historical practices, owing to their 11+ years of work experience, would not be hired under the DLP due to their short hair (i.e., their male-like characteristics captured in $\mathbf{x}$). Similarly, several men, who would not have been hired based on work experience alone, are advantaged by the DLP due to their longer hair (i.e., their 'female-like' characteristics in $\mathbf{x}$). The DLP violates rational ordering, and harms some of the most qualified individuals in the protected group. Group parity is achieved at the cost of individual unfairness.

Granted, we might not expect factors such as hair length to knowingly be used as inputs to a typical hiring algorithm. We construct this toy example to illustrate a more general point: since DLPs do not have direct access to the protected feature, they must infer from the other features which people are most likely to belong to each subgroup. Using the protected feature directly can yield more reasonable policies: For example, by applying per-group thresholds, we could hire the highest rated individuals in each group, rather than distorting rankings within groups based on how female/male individuals *appear* to be from their other features.

[2] https://github.com/mbilalzafar/fair-classification/

4.2 Case study: Gender bias in CS graduate admissions

For our next example, we demonstrate a similar result but this time by analyzing real data from the Master's admissions process of a large public university. We consider a sample of ∼9,000 students considered for admission over an 11-year period spanning 2006-2016. Half of the examples are withheld for testing. The available attributes include basic demographic information, such as country of origin, interest areas, and gender, as well as quantitative information such as GRE scores. Our data also includes a label in the form of a decision provided by an admissions committee.[3]

Based on a superficial analysis, we did not observe any gender bias (the admissions rates for male and female applicants are within 1% of each other). So, to demonstrate the effects of DLPs, we corrupt the data with *synthetic discrimination*. Of all women who were admitted, i.e., $z_i = b, y_i = 1$, we flip 25% of those labels to 0: giving noisy labels $\bar{y}_i = y_i \cdot \eta$, for $\eta \sim Bernoulli(.25)$. This simulates a setting in which the training data exhibits a historical bias.

We then train three logistic regressors: (1) To predict the (prejudice-corrupted) labels from the non-sensitive features $\{x_i, \bar{y}_i\}$; (2) The same model, applying the fairness constraint of Zafar et al. [2017a]; and (3) A logistic regressor that predicts the sensitive feature from the non-sensitive features $\{x_i, z_i\}$. The data contains limited information that can predict gender, though such predictions can be made better than random (AUC=0.59) due to different rates of gender imbalance across (e.g.) countries and interest areas.

Figure 2 (left) shapes our basic intuition for what is happening here: Considering the probability of admission for the unconstrained classifier (y-axis), students whose decisions are 'flipped' (after applying the fairness constraint) tend to be those close to the decision boundary. Furthermore, students *predicted* to be male (x-axis) tend to be flipped to the negative class (left half of plot) while students *predicted* to be female tend to be flipped to the positive class (right half of plot). This is shown in detail in Figure 2 (center and right). However, of the 19 students whose decisions are flipped to 'admit,' the majority (10) are males, each of whom has 'female-like' characteristics according to their other features, as demonstrated in our synthetic hair-length example. Demonstrated here with real-world data, the DLP both disrupts the within-group ordering and violates the *do no harm* principle by disadvantaging some women who, but for the DLP, would have been admitted.

4.2.1 Comparison with Treatment Disparity

To demonstrate the better performance of per-group thresholding (violating treatment parity), we implement a simple decision scheme and compare its performance to the DLP. Assuming that our model gives us calibrated probabilities, and that this is all the information available to the decision maker, it's easy to derive the optimal thresholds for maximizing expected accuracy under linear constraints on the proportions of predicted positives, like the CV-gap or p-% rule.

Our thresholding rule for maximizing accuracy subject to a p-% rule works as follows: Recall that the p-% rule requires that $q_b/q_a > p/100$. We can rewrite this as:

$$\frac{p}{100} q_a - q_b < 0$$

[3] These decisions do not precisely determine whether a student is made an offer, but rather represent an 'above-the-bar' assessment that is used to guide admissions decisions, and can be considered as a binary label.

Table 1: Comparison between unconstrained classification, DLPs, and thresholding schemes. Note that the p-% rules from Zafar et al. [2017a] were the strongest that could be obtained with their method; on complex datasets p-% rules of 100% are rarely obtained in practice, due to their specific approximation scheme. Employee and Customer datasets are from IBM, the others are UCI datasets.

basic statistics					**naïve (unconstrained) classification**			**fair (constrained) classification [Zafar et al., 2017a]**			**optimal threshold**
dataset	**%prot.**	**%prot. in +'ve**	**%non-prot. in +'ve**	**label p-%**	**acc.**	**prot./non-prot. in positive**	**p-%**	**acc.**	**prot./non-prot. in positive**	**p-%**	**p-% at const. acc.**
a	b	c	d	e	f	g	h	i	j	k	l
Income	66.9%	30.6%	10.9%	35.8%	0.85	8% / 25%	31%	0.85	7% / 24%	29%	52.9%
Marketing	60.2%	14.1%	10.1%	71.9%	0.89	3% / 4%	82%	0.89	3% / 3%	102%	100.3%
Credit	60.4%	24.1%	20.8%	86.0%	0.82	10% / 12%	88%	0.74	21% / 25%	85%	100.0%
Employee	45.8%	19.2%	12.5%	65.0%	0.87	8% / 12%	65%	0.86	8% / 11%	69%	100.4%
Customer	48.3%	33.0%	19.7%	59.7%	0.80	15% / 30%	49%	0.79	16% / 19%	84%	100.2%

Like the CV-gap, the p-% rule imposes a linear constraint. We denote the quantity $\frac{p}{100}q_a - q_b$ as the p-gap. To maximize accuracy subject to satisfying the p-% rule, we construct a score that quantifies reduction in p-gap per reduction in accuracy. Starting from the accuracy-maximizing predictions (thresholded at .5), we then flip those predictions which close the gap fastest:

1. Assign each example with $\{\tilde{y}_i = 0, z_i = b\}$ or $\{\tilde{y}_i = 1, z_i = a\}$, a score c_i equal to the reduction in the CV-gap divided by the reduction in accuracy:
   (a) For each example in group a with initial $\hat{y}_i = 1$, $c_i = \frac{p}{100 n_a (2\hat{p}_i - 1)}$.
   (b) For each example in group b with initial $\tilde{y}_i = 0$, $c_i = \frac{1}{n_b (1 - 2\hat{p}_i)}$.
2. Flip examples in descending order according to this score until the desired CV-score is reached.

These scores do not change after each iteration. So the greedy policy is optimal.

This dataset revealed that the method due to Zafar et al. [2017a] cannot produce any specified p-%-rule. On the admissions data, their algorithm maxes out at a p-% rule of 77.59%, compared to a p-% rule of 71.44% by naïve classification (on unseen test data). Both have similar accuracy: given that both positive labels and female applicants are a minority, assigning negative labels to males close to the boundary impacts accuracy very little. Both methods had accuracy of around 78% on this data. Critically though, by applying an optimal thresholding strategy, we were able to obtain the same accuracy as the method of Zafar et al. [2017a], but with a higher p-% rule of 78.34%; subject to a $< 1\%$ drop in accuracy we can achieve a p-% rule of $\sim 100\%$. Similarly, we

Table 2: Statistics of public datasets.

dataset	source	prot. feature	prediction target	n
Income	UCI [Kohavi, 1996]	Gender (female)	income > \$50k	32,561
Marketing	UCI [Moro et al., 2014]	Status (married)	customer subscribes	45,211
Credit	UCI [Yeh and Lien, 2009]	Gender (female)	credit card default	30,000
Employee Attr.	IBM [ibm]	Status (married)	employee attrition	1,470
Customer Attr.	IBM [ibm]	Status (married)	customer attrition	7,043

could achieve a modest improvement in accuracy ($<$0.1%) while maintaining the same p-% rule as the method of Zafar et al. [2017a].

4.3 Examples on public datasets

Finally, for reproducibility, we repeated our experiments from Section 4.2 on a variety of public datasets.[4] Again we compare applying our simple thresholding scheme against the fairness constraint of Zafar et al. [2017a], considering a binary outcome and a single protected feature. Basic info about these datasets (including the prediction target and protected feature) is shown in Table 2.

The protocol we follow is the same as in Section 4.2. Each of these datasets exhibits a certain degree of bias w.r.t. the protected characteristic (Table 1), so no synthetic discrimination is applied. In Table 1, we compare (1) The p-% rule obtained using the classifier of Zafar et al. [2017a] compared to that of a naïve classifier (column k vs. column h); and (2) The p-% rule obtained when applying our thresholding strategy from Section 4.2.1. As before, half of the data are withheld for testing.

First, we note that in most cases, the method of Zafar et al. [2017a] increases the p-% rule (column k vs. h), while maintaining an accuracy similar to that of unconstrained classification (column i vs. f). One exception is the UCI-Credit dataset, in which *both* the accuracy and the p-% rule simultaneously decrease; although this is against our expectations, note that the optimization technique of Zafar et al. [2017a] is an approximation scheme and does not offer accuracy guarantees in practice (nor can it in general achieve a p-% rule of 100%). However these details are implementation-specific and not the focus of this paper.

Second, as in Section 4.2.1, we note that the optimal thresholding strategy is able to offer a strictly larger p-% rule (column l vs. k) at a given accuracy (in this case, the accuracy from column i). In most cases, we can obtain a p-% rule of (close to) 100% at the given accuracy.

We emphasize that the goal of our experiments is not to 'beat' the method of Zafar et al. [2017a], or even to comment on any specific discrimination-aware classification scheme. Rather, we emphasize that *any* DLP is fundamentally upper-bounded (in terms of the p-% rule/accuracy trade-off) by simple schemes that explicitly consider the protected feature. Not only do our experiments validate this claim, but they also reveal that the practical difference between the two schemes is *large*: the two schemes make strikingly different decisions, and while 'hiding' the protected feature from the classifier may be conceptually desirable, practitioners of such schemes should be aware of the consequences of doing so.

[4] Code and data available on http://jmcauley.ucsd.edu/code/fairness/

5 Related Work Beyond DLPs

In this section we provide a brief overview of some of the other approaches that have been put forth for trading off between classification performance and impact disparity. One common approach consists of preprocessing or "massaging" the training data to reduce the dependence between the resulting model predictions and the sensitive attribute [Kamiran and Calders, 2009, 2012, Feldman et al., 2015, Adler et al., 2016, Johndrow and Lum, 2017]. These methods differ both in terms of what variables are affected by the data processing, and the degree of independence that is achieved. For instance, Kamiran and Calders [2009] propose flipping the negative labels of some observations in the disadvantaged class. Zemel et al. [2013] proposes learning representations—in this case, cluster assignments—of each example such that each example maps to a cluster with some probability, seeking parity in the proportion of each group assigned to each cluster. Feldman et al. [2015] also investigates transformations of the features X into a new set of features that are constructed to be marginally independent from Z. Johndrow and Lum [2017] demonstrate how to construct transformations to ensure that the derived features are jointly independent of Z, and show that this produces distributional parity of the resulting fitted model.

A second common approach is to modify existing classification methods either through post-hoc corrections or in the training stage to constrain the level of impact parity in the resulting model. Kamishima et al. [2011], Goh et al. [2016], Calders and Verwer [2010], Kamiran et al. [2010] consider modifications to methods such as SVM, logistic regression, Naive Bayes, and decision trees. Agarwal et al. [2017] show how impact parity constraints can be framed as a cost-sensitive classification problem.

6 Discussion

Now that we have expressed our primary technical arguments, we return to the critical questions driving research in discrimination-aware classification. The arguments in this portion of the paper address both the difficulty of communicating desiderata across disciplines, the relationship between classification and decision-making, and promising research directions in discrimination-aware ML beyond narrow considerations of the two parity measures that we focused on in earlier sections.

6.1 Coming to terms with treatment disparity

At present, academic work in law and machine learning tends to take place in a disjoint set of journals, and with notable exceptions researchers typically attend a disjoint set of conferences. These communities occasionally intersect when some paper, such as the widely-influential California Law Review article due to Barocas and Selbst [2016], reaches a cross-disciplinary audience and captures popular attention. However, the subsequent interdisciplinary work is often again siloed by discipline. Technical work tends to be published in technical conferences, where the peer-reviewers may be ill-equipped to identify shortcomings in problem formulation.

This sort of disciplinary isolation enables legal terms such as *disparate treatment* and *disparate impact* to be overloaded to mean something different from their legal antecedents. As a result,

methodological solutions guided by technical interpretations of legal criteria might nevertheless turn out to be incompatible with the policy desiderata they are designed to satisfy. Such solutions are unlikely to be adopted in practice. In the present context, we argue that (i) DLPs would enjoy no better legal standing than explicit treatment disparity if tested under the law (as supported in Grimmelmann and Westreich [2017]); and, (ii) some form of treatment disparity may already be tolerated under the law in order to ensure more fair outcomes. The latter view is supported by Pauline Kim in her paper *Data-driven discrimination at work* [Kim, 2017]:

> A formalist reading of Title VII might appear to prohibit any use of variables capturing sensitive characteristics in a data model. Certainly, a simple model that relied on race or other protected characteristics as the basis for adverse decisions would run afoul of Title VII's prohibitions. However, when dealing with a complex statistical model involving multiple variables, the appropriate treatment of these sensitive variables is more complicated. If the goal is to reduce biased outcomes, then a simple prohibition on using data about race or sex could be either wholly ineffective or actually counterproductive due to the existence of class proxies and the risk of omitted variable bias. Instead, avoiding classification bias may sometimes call for excluding sensitive demographic variables and at other times call for including them. Any response to biased data models must be sensitive to these nuances.

On the balance of these considerations, there are several compelling reasons for practitioners to promote equality more transparently through direct treatment disparity, rather than through hidden changes to the learning algorithm. As articulated earlier, treatment disparity approaches have three principal advantages over DLPs: they optimally trade accuracy for representativeness, preserve rankings among members of each group (as compared to the unconstrained scores), and do no harm to members of the disadvantaged group.

In addition to these three properties, treatment disparity has another advantage: by setting class-dependent thresholds, it's much easier to quantify intuitively how treatment disparity impacts individuals. Having such a quantity to reason about might help policy-makers to decide what magnitude of treatment disparity best trades off group equality and individual fairness. With indirect methods for increasing impact parity, it might be harder to reason about the intervention. As an example, it seems harder to express policy in terms of the value of a regularization coefficient (compared to a threshold).

Several key challenges remain. The theoretical arguments in this paper demonstrate that thresholding approaches are optimal in the setting where we assume complete knowledge of the data-generating distribution. It is not always clear how best to realize these gains in practice, where imbalanced or unrepresentative data sets can pose a significant obstacle to accurate estimation. Furthermore, some of our results are tailored to the CV or the p-% rule notions of group parity. As shown in Hardt et al. [2016], Woodworth et al. [2017] and Dwork et al. [2017], satisfying other parity criteria can be more difficult.

6.2 Separating estimation and decision-making

In the context of algorithmic, or algorithm-supported decision-making, it's often useful to obtain not just a classification, but also an accurate probability estimate. These estimates could then be

incorporated into the decision-theoretic part of the pipeline where appropriate measures could be taken to align decisions with social desiderata. By intervening at the modeling phase, DLPs distort the predicted probabilities themselves. It's not clear what the outputs of the resulting classifiers actually signify. In unconstrained learning approaches, even if the label itself may reflect historical prejudice, one at least knows what is being estimated. This leaves open the possibility of intervening at decision time to promote more equal outcomes.

While the distinction between building a model and making decisions is stated clearly in the first modern work on discrimination-aware classification [Pedreshi et al., 2008], this distinction is frequently muddled in subsequent papers. For example, Kamiran and Calders [2009] state that "a learned model may exhibit unlawfully prejudiced behavior." The conflation of modeling and decision-making may lead to counterproductive modifications to learning algorithms that do not adequately account for how models are actually used in practice. For example, these papers often assume that decision makers desire to optimize accuracy and hence that decisions will be made by thresholding probability estimates at 0.5. That's not typically how ML works in real-world applications. First, due to differences in the cost of false positives and false negatives, accuracy is seldom the most task-relevant metric. Furthermore, decision-makers are often faced with a multi-objective problem that entails considerations beyond what the algorithm is designed to predict.

6.3 Fairness beyond disparate impact

How best to quantify discrimination and unfairness remains an important open question. The CV scores and $p-\%$ rules addressed in this paper offer one set of definitions, but there are many other notions of fairness to which our results do not directly apply. For example, equality of opportunity as introduced in Hardt et al. [2016]—requiring equality of true positive rates across groups—has received considerable attention. Other notions of fairness and the trade-offs between them have been studied by Joseph et al. [2016], Kleinberg et al. [2016], Chouldechova [2017], Berk et al. [2017], Ritov et al. [2017]. In a recent work, Zafar et al. [2017b] depart from parity-based definitions and propose instead a preference-based notion of fairness. Dwork et al. [2017] address the problem of how best to incorporate information about protected characteristics for several of these other fairness criteria.

Problematically, research into fairness in ML is often motivated by the case in which our ground-truth data is itself biased. It is not clear how to assess many of these other fairness criteria in the presence of biased data. Characterizing different forms of data bias and their impacts on algorithmic auditing remains an important outstanding challenge.

Even if we accept that the solution for many proportional representation problems will take the form of treatment disparity in favor of the disadvantaged group, questions remain of "how much?" and "when?". How can we say when treatment disparity is correcting for recent discrimination manifested, e.g. as biased labels? When does it amount to affirmative action, correcting instead for historical discrimination? And when is treatment disparity itself actually tantamount to discrimination, as when used to exclude qualified Asian students from higher education [Hartocollis and Saul, 2017]. For now, it seems that these judgments must be exogenously specified by persons cognizant of the social context in which algorithms operate.

Recent work on identifying proxy discrimination [Datta et al., 2017] and causal formulations of

fairness [Nabi and Shpitser, 2017, Kilbertus et al., 2017, Kusner et al., 2017] offer some promising approaches to framing such problems. To answer these questions, it would help to have a better understanding of by what mechanisms and to what degree the data has been influenced by prejudice. Perhaps data mining and machine learning have some role to play in asking these questions? The answers could guide decisions about where and how strongly to intervene.

References

IBM Watson analytics blog. https://www.ibm.com/communities/analytics/watson-analytics-blog/.

Civil rights act of 1964, 1964. Accessed on September 11th, 2017.

Philip Adler, Casey Falk, Sorelle A Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models by obscuring features. *arXiv preprint arXiv:1602.07043*, 2016.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, and John Langford. A reductions approach to fair classification. 2017.

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

Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044*, 2017.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

Dan Biddle. *Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing*. Gower Publishing, Ltd., 2006.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 2010.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *arXiv preprint arXiv:1701.08230*, 2017.

Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613*, 2017.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, 2015.

Matt Ford. Racism and the execution chamber. 2014. URL https://www.theatlantic.com/politics/archive/2014/06/race-and-the-death-penalty/373081/.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *NIPS*, pages 2415–2423, 2016.

James Grimmelmann and Daniel Westreich. Incomprehensible discrimination. *California Law Review*, 2017.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *NIPS*, 2016.

Anemona Hartocollis and Stepanie Saul. Affirmative action battle has a new focus: Asian americans. 2017. URL https://www.nytimes.com/2017/08/02/us/affirmative-action-battle-has-a-new-focus-asian-americans.html?mcubz=1.

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 2016.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *Computer, Control and Communication*, 2009.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, 2010.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *ICDM Workshops*, 2011.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.02744*, 2017.

Pauline T Kim. Data-driven discrimination at work. *William & Mary Law Review*, 58(3), 2017.

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

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *KDD*, 1996.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Aditya Menon and Robert Williamson. The cost of fairness in binary classification. In *Fairness, Accountability and Transparency*, 2018.

S. Moro, P. Cortez, and P. Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 2014.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. *arXiv preprint arXiv:1705.10378*, 2017.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *KDD*, 2008.

Ya'acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parity as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519*, 2017.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

I. C. Yeh and C. H. Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 2009.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. *arXiv preprint arXiv:1707.00010*, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.