Protecting the Protected Group: Circumventing Harmful Fairness

Omer Ben-Porat[1], **Fedor Sandomirskiy**[2,3], **Moshe Tennenholtz**[2]
[1]Tel-Aviv University
[2]Technion—Israel Institute of Technology
[3]Higher School of Economics, St. Petersburg, Russia
omerbenporat@mail.tau.ac.il, fedor.sandomirskiy@gmail.com, moshet@ie.technion.ac.il

Abstract

The recent literature on fair Machine Learning manifests that the choice of fairness constraints must be driven by the utilities of the population. However, virtually all previous work makes the unrealistic assumption that the exact underlying utilities of the population (representing private tastes of individuals) are known to the regulator that imposes the fairness constraint. In this paper we initiate the discussion of the mismatch, the unavoidable difference between the underlying utilities of the population and the utilities assumed by the regulator. We demonstrate that the mismatch can make the disadvantaged protected group worse off after imposing the fairness constraint and provide tools to design fairness constraints that help the disadvantaged group despite the mismatch.

1 Introduction

At first glance, algorithms may seem free of human biases such as sexism or racism. However, in many situations, they are not: the automated recruiting tool used by Amazon was favoring men (Doleac and Hansen 2016); judges in the US use the COMPAS algorithm to estimate the probability that the defendant will re-offend while this algorithm was accused of being biased against black people (Larson et al. 2016). See O'Neill (2016) for many more examples. These challenges call for imposing fairness constraints on algorithm design and, in particular, on machine-learned classifiers, which are the subjects of this paper. As a running example, consider a bank that gives loans to potential borrowers and is regulated by a policy-maker. The bank learns a decision rule (namely, a classifier) from historical data to decide for whom to approve or decline the loan to maximize its revenue that is increasing with the number of repaid loans. As repeatedly observed in the past, the resulting classifier may be biased against a protected group (e.g., ethnic minority). Hence, the regulator may wish to impose a fairness constraint on the bank.

Bias is deemed unjust. Beyond that, it affects the *welfare* of protected groups, as borrowers have preferences towards the different outcomes of the classification, captured by utility functions.[1] Consequently, the goal of imposing fairness constraints is to improve the welfare of the *disadvantaged* group (the originally-discriminated one). A natural approach by which the regulator can achieve this goal is by assuming a utility function and requiring the bank's classifier to equalize the protected groups' welfare. The two most popular fairness constraints, *Demographic Parity* (Agarwal et al. 2018; Dwork et al. 2012) and *Equal Opportunity* (Hardt et al. 2016) (henceforth DP and EO, respectively) are special cases of this approach for particular utility functions. For instance, DP is recovered by assuming that every agent gets a utility of one when receiving the loan, and zero otherwise.

The possibility that fairness may harm the well-being of those it is designed to protect, i.e., that it can harm the disadvantaged group's welfare, seems counter-intuitive. However, it is well known both theoretically and empirically that the most intuitive fairness constraint of Unawareness (which forbids using the sensitive attribute in classification) can be harmful (Corbett-Davies and Goel 2018; Dwork et al. 2018; Ustun, Liu, and Parkes 2019; Doleac and Hansen 2016). For example, Doleac and Hansen (2016) show that the "ban the box" policy, adopted by the United States and preventing employers from seeing applicants' criminal background, decreased the welfare of discriminated minorities (the chances of getting a job). Additionally, Liu et al. (2018) discovered that DP and EO are not free of the same flaw if we consider long-term consequences. They show that fairness constraints may force the bank to give loans to those members of the disadvantaged group who, otherwise, would not have received the loans due to the high probability of default. Therefore, such unqualified borrowers are likely to have problems with paying back the loan. This increased default ratio would harm the disadvantaged group's average credit score, thereby harming its welfare in the long run.

1.1 Our Contribution

In this paper, we uncover another mechanism underlying harmful fairness even in static settings:

> *Imposing a fairness constraint can make the disadvantaged group worse off if the fairness constraint and the utilities of the population mismatch.*

[1]We use the term utility for the well-being of a single individual and welfare for the aggregated well-being of groups of individuals.

Following the recent trends of fair ML literature, we assume that agents may have different preferences over classification outcomes, which are captured by utility functions. For example, borrowers may differ in their value for getting the loan depending on their access to alternative sources of money and on the purpose of borrowing. With utilities in hand, we can use social welfare to evaluate a group's well-being for any given classifier. To talk about the mismatch between utilities and fairness constraints, the latter has to be defined in utilitarian terms. As we described above, fairness constraints like DP and EO are naturally cast as equalizing *some* welfare functions of the protected groups. However, the difficulty in applying any welfare-based fairness constraint is that the utilities must be known to the designer, while these utilities represent individuals' private tastes. Hence, even if domain experts determine the utilities used in a fairness constraint (henceforth, *assumed utilities*), they can only approximate the actual utilities of the population (*underlying utilities* in what follows). This discrepancy leads to the following conclusion.

> *In practice, the mismatch between the underlying utilities of the population and the utilities assumed by the regulator is unavoidable.*

Together with the observation that the mismatch can make fairness harmful, this becomes a serious caution for regulators that design fairness constraints for a certain industry, e.g., banking. We complement this caution with a positive message. Naturally, a small discrepancy between underlying and assumed utilities must be innocuous. However, we characterize a much more applicable and promising connection; we show that

> *Fairness constraints help the disadvantaged group whenever the utilities assumed by the regulator and the underlying utilities of the population agree on which group is disadvantaged.*

Finally, we suggest additional ways to deal with the mismatch if the underlying utilities can be approximated from data.

1.2 Related Work on Economic Ideas in Fair Classification

Welfare-Equalizing, our approach to fairness, has a long history in normative economics (Pazner and Schmeidler 1978; Roemer 1986) (where it is known under the name of egalitarianism) and political philosophy (Rawls 2009); it was used for fair resource allocation without money transfers (Li and Xue 2013), in the field of cooperative games (Dutta and Ray 1989) and bargaining problems (Kalai and Smorodinsky 1975). In contrast to recent papers on the utilitarian approach to fair classification (Heidari et al. 2018; Heidari, Gummadi, and Krause 2019), which suggest maximizing the minimal welfare among the protected groups, we strengthen this desideratum by making it a normative requirement: the welfare must be equal among the subgroups defined by a sensitive attribute. This normative condition allows one to separate the *fairness constraint* (which may be imposed by a regulator) from the *selfish objective of the decision-maker* (a revenue-maximizing bank in our running example) and thus allows one to analyze how decisions change after imposing the fairness constraint. Another advantage of the Welfare-Equalizing concept is the simple threshold structure of the optimal fair classifier (similar to the one for Demographic Parity or Equal Opportunity (Corbett-Davies et al. 2017)), which makes it efficiently computable.

This work joins recent attempts (Rambachan et al. 2020a,b; Hu and Chen 2020; Elzayn and Fish 2020; Hossain, Mladenovic, and Shah 2020) to bring better economic understanding to fairness in ML; we address some of them here and refer the reader to Finocchiaro et al. (2020) for a comprehensive survey. Hu and Chen (2020) propose an optimization framework for fair classification and welfare analysis. In their modeling, a learner executes a soft-margin SVM with the additional constraint of group fairness: limiting the two groups' welfare discrepancy to a predefined quantity. They provide a sensitivity analysis, showing that applying stricter fairness constraints (decreasing the allowed discrepancy) can worsen welfare outcomes for both groups. Their findings are in line with ours, but our analysis is fundamentally different; in particular, they do not address the utility mismatch issue. Imposing fairness constraints on profit-maximizing entities, as we do in this paper, is an understudied point of view, as noted recently by Elzayn and Fish (2020).

There are also several recent attempts to harness economic principles to fair classification. For example, Gölz, Kahng, and Procaccia (2019) treat fair classification as an allocation of goods, where there is a fixed amount of resources to distribute. They examine the compatibility (or lack thereof) of Equalized Odds with axioms of fairness from the economic literature on fair division; see (Brandt et al. 2016) for a survey. Envy-freeness, the dominant fairness concept in economics, plays a crucial role in several recent papers at the intersection of economics and AI (Caragiannis et al. 2019; Cohler et al. 2011; Benade et al. 2018; Guruswami et al. 2005; Gal et al. 2016; Plaut and Roughgarden 2018), including several works on fair classification (Zafar et al. 2017; Balcan et al. 2019; Ustun, Liu, and Parkes 2019; Hossain, Mladenovic, and Shah 2020). However, these papers on fair classification focus on sample complexity and generalization (Balcan et al. 2019), or asserting that users favor treatment disparity (Ustun, Liu, and Parkes 2019; Zafar et al. 2017) in health applications.

The work most related to ours is the paper by Hossain, Mladenovic, and Shah (2020). Concurrently to and independently of our work, Hossain, Mladenovic, and Shah (2020) argue for group equability in fair classification, which coincides with our Welfare-Equalizing fairness constraint. They, too, show that their concept subsumes previously suggested fairness notions. However, there is a significant difference between the two works. First, Hossain, Mladenovic, and Shah (2020) are interested in learning the best classifier from data, and hence address issues of generalization from samples and differentiability. In contrast, we devote our paper to societal considerations of fair classification, and thus consider fairness as a post-processing step similarly to (Corbett-Davies et al. 2017; Hardt et al. 2016). Additionally, in contrast to Hossain, Mladenovic, and Shah (2020), our analysis focuses on the mismatch between the utilities assumed by the regulator and the actual, underlying utilities.

1.3 Paper Structure

In Section 2, we present our formal model, define the Welfare-Equalizing fairness framework, and prove structural results for optimal fair classifiers in Subsection 2.2. Section 3 deals with the implications of a mismatch between the population's underlying utilities and the utilities assumed by the regulator. Finally, Section 4 describes how to compute the bank-optimal fair classifier if the assumed utilities approximate the underlying utilities well enough.

2 Model

We consider a general classification problem, where agents have an ex-ante non-observable "quality" correlated with observable attributes. We keep using the metaphor of a bank that predicts the reliability of the population and makes lending decisions; however, the same setting captures student admissions, recruiting, assessing the recidivism risk for a criminal, etc.

There are three parties in the model: a heterogeneous population of potential borrowers; a bank that makes lending decisions based on the observable attributes of borrowers and cares only about its revenue; and a regulator that cares about fairness and can restrict the set of lending policies available to the bank by imposing a fairness constraint. We now present these parties formally.

Borrowers We assume that each potential borrower (henceforth borrower) is associated with a pair of observable attributes $(X, A) \in \mathcal{X} \times \{0, 1\}$. Here A is a binary[2] sensitive attribute (e.g., gender) and $X \in \mathcal{X}$ encodes all other characteristics of a borrower, e.g., employment history, salary, education, assets and so on. We do not impose any assumptions on $\mathcal{X}$. By $\{A = 0\}$ and $\{A = 1\}$ we denote the groups of all borrowers with the sensitive attribute equal to 0 or 1, respectively; we call $\{A = a\}$ a *protected group*. Furthermore, in addition to the observable attributes X and A, every borrower is also associated with an unobservable variable $Y \in \{0, 1\}$, which describes whether that borrower will pay back the loan or not. For brevity, we call borrowers with $Y = 1$ and $Y = 0$, *good* and *bad*, respectively. The statistical characteristics of the population are described by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$; so $X = X(\omega)$, $A = A(\omega)$ and $Y = Y(\omega)$ are random variables on $\omega \in \Omega$. By small letters (x, a, y) we denote realizations of X, A, and Y, i.e., generic elements of $\mathcal{X} \times \{0, 1\} \times \{0, 1\}$.

Each borrower (x, a, y) obtains a *utility* when receiving the loan. This utility can depend on x and a in many ways, but what is more critical is that it must depend on the non-observable quality of the borrower, namely $v = v(x, a, y)$. To simplify the presentation, we assume that the utility from a rejected application is zero and that the average utility of a borrower with given x and a is non-negative,[3] i.e., $\mathbb{E}[v(X, A, Y) \mid X = x, A = a] \geq 0$. However, both assumptions can be relaxed. We refer to v as the *underlying utility* of the population.

The bank We assume that the bank knows the joint distribution of (X, A, Y) from historical data. In particular, it knows the exact conditional probability of being a good borrower given the observable attributes; we denote it by $p(x, a) = \mathbb{P}(Y = 1 \mid X = x, A = a)$.[4]

The bank makes lending decisions based on X and A but without observing Y. It uses a classifier $c : \mathcal{X} \times \{0, 1\} \to [0, 1]$ where $c(x, a)$ is the probability of giving a loan to a population of borrowers with $X = x$ and $A = a$. Each loan given to a good borrower brings a revenue[5] of $\alpha_+(X) > 0$ to the bank while each bad borrower leads to a loss of $\alpha_-(X) > 0$; we assume that $\alpha_\pm(x)$ are bounded functions of $x \in \mathcal{X}$. The bank's revenue depends on the choice of a classifier c, and is defined by

$$R(c) = \mathbb{E}\left[c(X, A)\left(\alpha_+(X)Y - \alpha_-(X)(1 - Y)\right)\right]. \quad (1)$$

To ease notation, we define $t(x)$ and $r(x, a)$ such that for every $a \in \{0, 1\}, x \in \mathcal{X}$

$$t(x) \coloneqq \frac{\alpha_-(x)}{\alpha_+(x) + \alpha_-(x)}, \quad (2)$$

$$r(x, a) \coloneqq \left(\alpha_+(x) + \alpha_-(x)\right)\left(p(x, a) - t(x)\right); \quad (3)$$

hence, we can rewrite Equation (1) by taking a conditional expectation with respect to A, X as

$$\begin{aligned} R(c) &= \mathbb{E}\big[c(X,A)\big(\alpha_+(X)p(X,A) - \alpha_-(X)(1-p(X,A))\big)\big] \\ &= \mathbb{E}[r(X,A)c(X,A)]. \end{aligned} \quad (4)$$

The goal of the bank is to maximize $R(c)$ over the set of *feasible* classifiers, i.e., classifiers that satisfy the regulator's constraints.

The regulator The regulator evaluates the well-being of a group using its *welfare*: the expected utility of its members. For an underlying utility function v and a classifier c, the welfare of the subgroup $\{A = a\}$ is given by

$$W_{v,c}(a) = \mathbb{E}\left[v(X, A, Y)c(X, A) \mid A = a\right]. \quad (5)$$

The regulator aims to equalize welfare among the protected groups. However, as we discuss in the introduction, the underlying utility v is unknown to the regulator; thus, the regulator is forced to use a certain substitute u instead. We refer to u as the *assumed utility*; ideally, the assumed utility must be an approximation to the underlying one.

The objective of the regulator is captured by the following u-*Welfare-Equalizing* constraint (u-WE for abbreviation).

[2] The assumption of the dichotomy of A is made for simplicity. Extending our results to the non-binary case (ethnicity) is straightforward.

[3] Zero utility for not getting a loan is a normalization-condition: before borrowing money, everybody is at zero utility level. Non-negativity of v can be regarded as a rationality assumption on borrowers: no rational agent would apply for a loan if she/he expects that getting the loan brings negative utility while not getting gives 0.

[4] Indeed, this is aligned with previous works that consider fairness as a post-processing step (Corbett-Davies et al. 2017; Hardt et al. 2016).

[5] In contrast to the rest of the literature, we allow the bank's revenue to depend on the non-sensitive attribute X. This becomes important if X also encodes the type of loan a client is applying for, e.g., different borrowers may need a different amount of money and thus bring a different revenue/loss.

Definition 1 (u-WE classifier)**.** *Given a utility-function u, a classifier c is u-Welfare-Equalizing if*

$$W_{u,c}(0) = W_{u,c}(1), \tag{6}$$

i.e., if c equalizes the welfare among the two protected groups. The set of all such classifiers is denoted by $\mathrm{WE}(u)$.

Note that Welfare-Equalizing constraint is defined with respect to the assumed utility.

2.1 Special Cases of Welfare-Equalizing Fairness

The framework of Welfare-Equalizing fairness allows one to analyze existing fairness constraints in a unified manner. For instance,

- The fairness constraint of *Demographic Parity* (DP) (e.g., (Agarwal et al. 2018; Dwork et al. 2012)) requires that the fraction of those who receive loans in the two groups must be the same. Formally, a classifier c satisfies DP if $\mathbb{E}[c(X,A) \mid A=0] = \mathbb{E}[c(X,A) \mid A=1]$. It is a special case of WE fairness with $u(x,a,y) \equiv 1$ for any triplet (x,a,y).
- Motivated by drawbacks of DP, Hardt et al. (2016) concluded that good and bad borrowers within protected groups must be treated separately and introduced the concept of *Equal Opportunity* (EO). Under this fairness constraint, the fraction of good borrowers who get loans must be the same in the two subgroups. Formally, a classifier c satisfies EO if

  $$\mathbb{E}[c(X,A) \mid Y=1, A=0] = \mathbb{E}[c(X,A) \mid Y=1, A=1].$$

  We recover EO by setting $u(x,a,y) = y \cdot \beta_a$. The coefficients $\beta_a = 1/\mathbb{E}[Y \mid A=a]$ normalize the maximal possible welfare in each group to 1. Such a rescaling is known under the name of "relative welfare" and is commonly used in economics to make welfare or utilities among groups comparable (Kalai and Smorodinsky 1975).
- Borrowers can differ in the amount of money m they need. We can assume that information about m is encoded in X, so $m = m(X)$. Then, a straightforward generalization of EO is the following concept of *Heterogeneous*-EO given by $u(x,a,y) = y \cdot m(x) \cdot \beta_a$ with $\beta_a = 1/\mathbb{E}[m(x) \mid A=a]$. We can capture any other heterogeneity similarly (e.g., different interest rates, time-period, and payment schedules).

2.2 Structural Properties of the Bank-Optimal Classifiers

In this subsection, we analyze classifiers that are optimal for the bank. We first state the result of Corbett-Davies et al. (2017), who characterize the structure of the bank-optimal *unconstrained* classifier. Then, we build upon their results for the *constrained* case. Namely, we assume that the regulator imposes the u-WE fairness constraint on the bank and explore the structure of the bank-optimal classifiers. We show that the optimal classifiers have a generalized threshold structure, a fact that is extensively used in Sections 3 and 4.

Unconstrained classifier c^*_{unc} If the regulator imposes no constraint on the bank, i.e., the bank is free to choose any classifier, then the revenue-maximizing classifier has a simple form (Corbett-Davies et al. 2017). Only borrowers with $r(x,a) > 0$ are profitable for the bank, which is equivalent to the probability of paying back $p(x,a)$ being greater than $t(x)$ (recall that t and r are defined by Equations (2) and (3)). Consequently, the optimal lending policy is given by the following threshold classifier c^*_{unc}: all borrowers with $p(x,a) > t(x)$ get loans ($c^*_{\text{unc}}(x,a) = 1$) and all borrowers with $p(x,a) \le t(x)$ are rejected ($c^*_{\text{unc}}(x,a) = 0$).[6]

Constrained classifier $c^*_{\mathrm{WE}(u)}$ Consider the general, constrained case, where the regulator imposes u-WE fairness on the bank. For a fixed and given assumed utility u, we denote by $c^*_{\mathrm{WE}(u)}$ the classifier that maximizes the bank's revenue $R(c)$ (see Equation (1)) among all u-WE classifiers $c \in \mathrm{WE}(u)$. The set of u-WE classifiers is non-empty since $0 \in \mathrm{WE}(u)$ and, therefore, the bank's optimization problem is well-defined.

To ease notation, we denote by $\overline{u}(x,a)$ the assumed utility of a borrower associated with (x,a) averaged over the possible values of Y,

$$\begin{aligned}\overline{u}(x,a) &= \mathbb{E}[u(X,A,Y) \mid X=x, A=a] \\ &= u(x,a,1)p(x,a) + u(x,a,0)(1-p(x,a)).\end{aligned}$$

Further, we denote by $R^*_a(w)$ the maximal revenue that the bank could extract from the group $\{A=a\}$ at the (assumed) welfare level $W_{u,c}(a) = w$. Formally,

$$R^*_a(w) = \max_{\{c\colon \mathcal{X} \to [0,1] \mid W_{u,c}(a) = w\}} \mathbb{E}\left[r(X,A) \cdot c(X) \mid A=a\right],$$

where r is given by Equation (3). The following Proposition 2 shows that the bank-optimal constrained classifier always exists and reveals its structure.

Proposition 2. *The bank-optimal u-WE classifier $c^*_{\mathrm{WE}(u)}$ exists. Furthermore, each optimal classifier has the following form:*

$$c^*_{\mathrm{WE}(u)}(x,a) = \begin{cases} 1 & r(x,a) > \lambda_a \overline{u}(x,a) \\ 0 & r(x,a) < \lambda_a \overline{u}(x,a) \\ \tau_a(x) & r(x,a) = \lambda_a \overline{u}(x,a) \end{cases}. \tag{7}$$

The group-dependent thresholds λ_a, $a \in \{0,1\}$ belong to the super-gradient[7] of the subgroup-optimal revenue $R^*_a(w)$ (a concave function of w) computed at the welfare level w^* maximizing the total bank's revenue $\mathbb{P}(A=0)R^*_0(w) + \mathbb{P}(A=1)R^*_1(w)$. The functions $\tau_a : \mathcal{X} \to [0,1]$ are arbitrary[8] up to the constraint that $c^*_{\mathrm{WE}(u)}$ provides the desired welfare level w^* for both groups, $w^* = W_{u,c^*_{\mathrm{WE}(u)}}(0) = W_{u,c^*_{\mathrm{WE}(u)}}(1)$.

[6] For definiteness, we assume that if the bank finds the two decisions equally profitable (the knife-edge case $p(x,a) = t(x)$), it chooses the one with fewer loans given (e.g., this policy minimizes paperwork).

[7] For a concave function $f = f(t)$, $t \in [t_0, t_1]$, the super-gradient $\partial_t f$ is the set of all $q \in \mathbb{R}$ such that $f(t') \le f(t) + q(t'-t)$ for all t'. If f is continuous, then for any t the super-gradient is non-empty, see Rockafellar (2015).

[8] In particular, there always exists $c^*_{\mathrm{WE}(u)}$ with constant τ_a, i.e., independent of x.

Proof sketch of Proposition 2. The revenue maximization over $c \in \mathrm{WE}(u)$ can be represented as a two-stage procedure. In the first stage, we find the revenue-maximizing classifier in each of the subgroups $\{A = a\}$ given the welfare level w; in the second stage, we optimize over w. The welfare constraint in the first stage can be internalized using the Lagrangian approach; the corresponding Lagrange multipliers λ_a are equal to the "shadow prices", i.e., the derivatives of the value functions $R^*_a(w)$ with respect to w. This internalization reduces finding the subgroup optimal classifier to the unconstrained problem; thus, the optimal classifier has a threshold structure similarly to c^*_{unc}. This structure is inherited by $c^*_{\mathrm{WE}(u)}$. Since the resulting linear program is infinite-dimensional and $R^*_a(w)$ may be non-differentiable, the formal proof requires some functional-analytic arguments presented in the appendix. □

For the special cases of Demographic Parity and Equal Opportunity, the explicit form of the optimal classifiers was obtained by Corbett-Davies et al. (2017). Their result becomes an immediate corollary of Proposition 2; see the appendix.

3 Mismatch of Fairness and Underlying Utilities

As we discussed in the introduction, the regulator aiming to equalize welfare among protected groups unavoidably assumes a certain approximation u of the underlying utilities v. For example, u can be determined by the domain experts, while v reflects the private tastes of the population and hence is not observable directly. We refer to the fact that u is different from v as a *mismatch*. This section explores how imposing the Welfare-Equalizing fairness constraint with respect to u affects the underlying welfare, which is measured by v.

We use the situation that exists before imposing the fairness constraint as the benchmark. We say that a group $\{A = a\}$ is *v-disadvantaged* if under the bank-optimal unconstrained classifier, the welfare of $\{A = a\}$ is lower than the welfare of the other group $\{A = 1 - a\}$. Formally, the group $\{A = a\}$ is v-disadvantaged if

$$W_{v,c^*_{\text{unc}}}(a) < W_{v,c^*_{\text{unc}}}(1-a),$$

where W is defined in Equation (5) and c^*_{unc} is the bank-optimal unconstrained classifier from Subsection 2.2. When the underlying utility v is clear from the context, we say that the group is disadvantaged and omit the dependence on v.

Ideally, imposing WE-fairness (or any other fairness constraint) should improve the welfare of the disadvantaged group. However, as we show next, this is not always the case. We say that a fairness constraint *harms* the group $\{A = a\}$ if $W_{v,c^*}(a) < W_{v,c^*_{\text{unc}}}(a)$, where c^* is the bank-optimal classifier after imposing that fairness constraint. Put differently, the fairness constraint harms the group $\{A = a\}$ if the welfare of the group (measured with respect to underlying utilities) decreases after imposing the fairness constraint.

Harmful mismatch We now demonstrate that the mismatch between assumed and underlying utilities can make the fairness constraint harmful to the disadvantaged group.

Example 3. Let the underlying utility-function be $v(x, a, y) \equiv 1$, i.e., all borrowers equally benefit from receiving loans. However, the regulator does not know the underlying utilities and decides to impose the fairness constraint of EO. Equivalently, the regulator assumes the utility $u(x, a, y) = y \cdot \beta_a$, for normalizing coefficients β_a (as we explained in Subsection 2.1). Notice that the underlying utility is the one associated with DP, and the regulator's assumed utility is the one associated with EO. Since $v \neq u$, there is a mismatch.

Suppose that $\mathcal{X} = \{0, 1, 2\}$, and all the combinations of (x, a) have the same probability of $\frac{1}{6}$. Furthermore, assume that the fraction $p(x, a)$ of good borrowers is given by the table

	$x=0$	$x=1$	$x=2$
$a=0$	$3/4$	$3/4$	$1/4$
$a=1$	1	0	0

.

In addition, let the revenue of the bank from a paid-back loan be $\alpha_+(x) = 1$, and the loss from a borrower's default be $\alpha_-(x) = 2$ for every $x \in \mathcal{X}$.

We first want to determine which group is disadvantaged. The threshold for the bank-optimal unconstrained classifier c^*_{unc} equals $t(x) = \frac{2}{3}$. Hence, under c^*_{unc}, only borrowers with $x = 0$ receive loans in the group $\{A = 1\}$. In $\{A = 0\}$, borrowers with $x \in \{0, 1\}$ receive loans since in such cases $t(x) = \frac{2}{3} < \frac{3}{4} = p(x, 0)$; however, loans are not given to borrowers with $x = 2$ since $t(2) = \frac{2}{3} > \frac{1}{4} = p(2, 0)$. Consequently, $\{A = 1\}$ is disadvantaged: $W_{v,c^*_{\text{unc}}}(0) = \frac{2}{3}$ compared to $W_{v,c^*_{\text{unc}}}(1) = \frac{1}{3}$.

Next, let us examine how imposing the fairness constraint of EO changes the outcome of the bank-optimal classifier. The proportion of loans given by c^*_{unc} to good borrowers in $\{A = 0\}$ is equal to the welfare of this group with respect to the assumed utility u, namely

$$W_{u,c^*_{\text{unc}}}(0) = \mathbb{E}\left[c^*_{\text{unc}}(X, A) \mid Y = 1, A = 0\right] = \frac{6}{7}.$$

In contrast, for $\{A = 1\}$ we have

$$W_{u,c^*_{\text{unc}}}(1) = \mathbb{E}\left[c^*_{\text{unc}}(X, A) \mid Y = 1, A = 1\right] = 1.$$

By imposing u-WE-fairness, the regulator requires the bank to equalize these two quantities. To do so, the bank-optimal constrained classifier can either increase the amount of loans given to $\{A = 0\}$ by approving some applications of $x = 2$ or decrease the number of loans given to $\{A = 1\}$. However, giving loans to $x = 2$ in $\{A = 0\}$ is too costly: the cost $\frac{3}{4}\alpha_- - \frac{1}{4}\alpha_+$ is not compensated by the benefit $1 \cdot \alpha_+$ from giving the same amount of loans to good borrowers with $x = 0$ in $\{A = 1\}$. Therefore, the bank-optimal constrained classifier coincides with c^*_{unc} in $\{A = 0\}$ and gives fewer loans to $x = 0$ in the group $\{A = 1\}$: $c^*_{\mathrm{WE}(u)}(0, 1) = \frac{6}{7}$. Consequently, the bank equalizes the proportion of loans given to good borrowers in both groups: $W_{u,c^*_{\mathrm{WE}(u)}}(0) = W_{u,c^*_{\mathrm{WE}(u)}}(1) = \frac{6}{7}$. However, since the underlying utility is measured by v, not u, we get that $W_{v,c^*_{\mathrm{WE}(u)}}(1) = \frac{2}{7} < W_{v,c^*_{\text{unc}}}(1) = \frac{1}{3}$ and the disadvantaged group is harmed by imposing the fairness constraint.

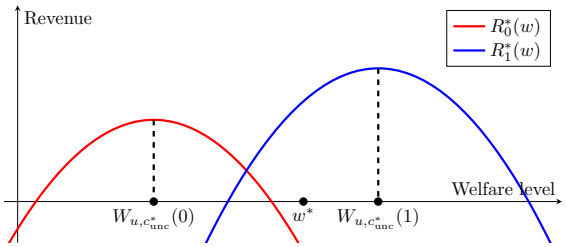

Figure 1: Intuition for the proof of Lemma 5. The red curve illustrates the revenue $R_0^*(w)$ from the disadvantaged group $\{A = 0\}$ at every possible welfare level, and the blue curve illustrates $R_1^*(w)$ from the advantaged group $\{A = 1\}$. The subgroup revenues $R_a^*(w)$ are concave functions of w that attain their maxima at a welfare level of $W_{u,c^*_{\text{unc}}}(a)$. The total revenue is $\mathbb{P}(A = 0)R_0^*(w) + \mathbb{P}(A = 1)R_1^*(w)$, which is maximized at $w^* = W_{u,c^*_{\text{WE}(u)}}$. Noticeably, it always lies between the two maxima. In this illustration, $\mathbb{P}(A = 0) = \frac{1}{3}$ and $\mathbb{P}(A = 1) = \frac{2}{3}$.

3.1 Can WE-fairness Help the Disadvantaged Group Despite The Mismatch?

In this subsection, we examine when imposing WE-fairness could help the disadvantaged group, even in the presence of a mismatch.[9] A natural observation is that if the assumed and underlying utilities "almost" match, the results of imposing u-WE-fairness or v-WE-fairness should be roughly the same. We postpone such quantitative statements to the end of this subsection and first address easy-to-check general conditions guaranteeing that fairness is not harmful. Theorem 4 below shows that only a lenient condition is required to assure that imposing WE-fairness benefits the disadvantaged group.

Theorem 4. *If v and u agree on which group is disadvantaged, then the u-WE classifier weakly increases v-welfare of the disadvantaged group, i.e.,*

$$\begin{cases} W_{u,c^*_{\text{unc}}}(a) < W_{u,c^*_{\text{unc}}}(1-a) \\ W_{v,c^*_{\text{unc}}}(a) < W_{v,c^*_{\text{unc}}}(1-a) \end{cases} \implies W_{v,c^*_{\text{unc}}}(a) \le W_{v,c^*_{\text{WE}(u)}}(a).$$

At first glance, this theorem may look rather intuitive. However, the claim is non-trivial even if there is no mismatch, i.e., when $v \equiv u$. To see this, recall that the WE-fairness constraint is imposed on the bank: a self-interested party, which is going to find the revenue-optimal way to satisfy the constraint. One possible way to achieve welfare equality is to give no loans to both protected groups thus harming both of them. As we show in the appendix, such an undesired behavior is possible when Unawareness is imposed. However, it never happens under the WE-fairness; we use this inherent property of WE-fairness as a tool to prove Theorem 4. Momentarily, let us assume that there is no mismatch, i.e., that the assumed utility and the underlying one are exactly the same. In such a case, the following Lemma 5 suggests that not only imposing WE-fairness always improves the welfare of the disadvantaged group, but also that every *individual* in the disadvantaged group is weakly better off.

Lemma 5 (Matching utilities)**.** *The bank-optimal u-WE classifier makes the u-disadvantaged protected group $\{A = a\}$ weakly better off at the expense of the advantaged group. Formally,*

$$W_{u,c^*_{\text{unc}}}(a) < W_{u,c^*_{\text{unc}}}(1-a) \implies \begin{cases} W_{u,c^*_{\text{unc}}}(a) \le W_{u,c^*_{\text{WE}(u)}}(a) \\ W_{u,c^*_{\text{unc}}}(1-a) \ge W_{u,c^*_{\text{WE}(u)}}(1-a) \end{cases}.$$

Moreover, any borrower from the u-disadvantaged group who receives a loan under the unconstrained classifier, receives it under the bank-optimal u-WE one.[10] *Formally, for every $x \in \mathcal{X}$ it holds that*

$$W_{u,c^*_{\text{unc}}}(a) < W_{u,c^*_{\text{unc}}}(1-a) \implies c^*_{\text{unc}}(x,a) \le c^*_{\text{WE}(u)}(x,a).$$

Proof of Lemma 5. By Proposition 2, the welfare level $w^* = W_{u,c^*_{\text{WE}(u)}}(0) = W_{u,c^*_{\text{WE}(u)}}(1)$ achieved by the u-WE classifier maximizes the revenue $\mathbb{P}(A = 0)R_0^*(w) + \mathbb{P}(A = 1)R_1^*(w)$ as a function of welfare level w. The sub-group revenues $R_a^*(w)$, $a \in \{0, 1\}$ are concave functions; thus the welfare level w^* lies between their maxima. These maxima are attained at the welfare levels of the bank-optimal unconstrained classifier; therefore, the welfare level w^* is between $W_{u,c^*_{\text{unc}}}(a)$ and $W_{u,c^*_{\text{unc}}}(1-a)$. See Figure 1 for illustration.

The second part of the lemma, the individual guarantees, follow from the threshold structure of the bank-optimal constrained classifier $c^*_{\text{WE}(u)}$, which we identified in Proposition 2. Since the welfare level w^* is above the maximum of $R_a^*(w)$ (for $\{A = a\}$ being the disadvantaged group), its super-gradient contains $\lambda_a \le 0$; therefore, $c^*_{\text{unc}}(x,a)$, which corresponds to $\lambda_a = 0$, is below $c^*_{\text{WE}(u)}(x,a)$. □

Equipped with Lemma 5, we are ready to prove Theorem 4.

Proof of Theorem 4. We apply Lemma 5 to the assumed utility function u. By the second part of the lemma, after imposing the u-WE constraint every borrower $x \in \mathcal{X}$ from the u-disadvantaged group $\{A = a\}$ who received loans under the unconstrained classifier still receives them. Namely, $c^*_{\text{unc}}(x,a) \le c^*_{\text{WE}(u)}(x,a)$ for all $x \in \mathcal{X}$. Multiplying both sides by the actual underlying utility $v(x,a,y)$, substituting $X = x$ and $Y = y$, and taking expectation, we

[9] We remind the reader that "helping" and "harming" is always with respect to the actual underlying utility.

[10] While our paper is focused on group notions of fairness, we stress that this result provides stronger "individual" guarantees.

get $W_{v,c^*_{\text{unc}}}(a) < W_{v,c^*_{\text{WE}(u)}}(a)$. In other words, the bank-optimal classifier $c^*_{\text{WE}(u)}$ for the utilities u assumed by the regulator improves the welfare of the group $\{A = a\}$ with respect to the underlying utilities v. Since v and u agree on the disadvantaged group's identity, this classifier improves the disadvantaged group's underlying welfare. □

In the presence of a mismatch, it is easy to see that a weak converse to Theorem 4 holds, i.e., fairness always makes the disadvantaged group weakly worse off. To avoid a mismatch, one can use a simple quantitative criterion. We say that u is an α-approximation of v for some $\alpha \geq 1$ if for all x, a, and y, $\frac{1}{\alpha} \leq \frac{v(x,a,y)}{u(x,a,y)} \leq \alpha$. Theorem 4 implies the following quantitative criterion on how well the regulator should approximate the underlying utilities to help the disadvantaged group.

Corollary 6. *If the utility u assumed by the regulator is an α-approximation of the underlying utility v and the gap between the welfare of the groups with respect to u is big enough, namely, $\frac{W_{u,c^*_{\text{unc}}}(0)}{W_{u,c^*_{\text{unc}}}(1)} \in \left(0, \frac{1}{\alpha^2}\right) \cup \left(\alpha^2, +\infty\right)$, then u-WE classifier helps the v-disadvantaged group.*

4 Computing Bank-Optimal Welfare-Equalizing Classifiers

In this section, we provide evidence for the applicability of WE-fairness by developing tools for computing bank-optimal WE classifiers. Our goal is to show how the bank can use the assumed utility proposed by the regulator to compute approximately optimal classifiers. We first discuss the case where the underlying and assumed utilities match (i.e., $v \equiv u$) and the other objects are fully known (revenues and losses $\alpha_{\pm}(x)$, and the probability $p(x, a)$ of paying back). Later on, we relax this assumption. Due to space considerations and our desire to focus on the conceptual assets of the paper, we defer most of the analysis to the appendix, as well as an elaborated version of formal statements.

Consider the case where $u \equiv v$, and both $\alpha_{\pm}$ and p are known. If the data is tabular, i.e., $\mathcal{X}$ is relatively small (say several thousand different borrower types), we can compute the bank-optimal u-WE-classifier $c^*_{\text{WE}(u)}$ explicitly by standard LP-methods. For large sets of attributes, e.g., multidimensional or continuous, the size of the LP "explodes", and a different approach should be taken. In this case, we use the structural insights from Proposition 2: the bank-optimal WE-classifier $c^*_{\text{WE}(u)}$ is parameterized by the two thresholds λ_a for $a \in \{0, 1\}$; therefore, to compute it we can restrict our attention to a finite-dimensional parametric family of classifiers. Due to the large-scale nature of the problem, we shall seek efficient algorithms for computing approximately bank-optimal WE classifiers, where these approximated solutions are defined as follows.

Definition 7. *A classifier c is $(\varepsilon, \varepsilon')$ bank-optimal u-WE if $R(c) \geq R(c^*_{\text{WE}(u)}) - \varepsilon$ and $|W_{u,c}(0) - W_{u,c}(1)| \leq \varepsilon'$.*

Notice that such classifiers are doubly approximated: they approximate the revenue of the (exact) bank-optimal WE classifier, and also approximately equalize the welfare of the two classes. In the appendix, we demonstrate how the bank can efficiently find a classifier that approximates the revenue and (exactly) equalizes the welfare. Namely, we show how one can apply the ternary search method (as in Hardt et al. (2016)) to find $(\varepsilon, 0)$ u-WE-classifier in $O\left(\log^2\left(\frac{1}{\varepsilon}\right)\right)$ runtime.

Next, we get rid of the full information assumption. Recent papers (Balcan et al. 2019; Hossain, Mladenovic, and Shah 2020) propose convex relaxations for imposing fairness constraints in settings like ours, which includes generalization bounds. However, since an extensive body of literature deals with estimating real-valued functions (ranging from linear regression to deep learning), we take a different approach. We suggest that the bank employs the assumed utility given by the regulator, and describe its performance guarantees in terms of the "quality" of u. This perspective has been adopted recently for several other ML problems (Medina and Vassilvitskii 2017; Lykouris and Vassilvitskii 2018; Purohit, Svitkina, and Kumar 2018).

For simplicity, we assume that $|r(x, a)|$, $u(x, a, y)$, and $v(x, a, y)$ are all upper-bounded by 1.

Proposition 8. *Fix a small $\delta > 0$ and assume that the bank has access to a sample of $(X, A, Y, \alpha_{\pm}, v)$ and to estimators u and $\hat{r}$ such that $\mathbb{E}[|u - v|] \leq \eta_u$ and $\mathbb{E}[|\hat{r} - r|] \leq \eta_r$ for small enough η_u and η_r. Then, a $(\varepsilon, \varepsilon)$ bank-optimal v-WE classifier with*

$$\varepsilon = 2\sqrt{6\left(\frac{1}{\mathbb{P}(A=0)} + \frac{1}{\mathbb{P}(A=1)}\right)\max\{\eta_u, \eta_r\}}$$

can be computed with probability $1 - \delta$ on a sample of size

$$O\left(\frac{1}{\max\{\eta_u, \eta_r\}}\left(\log\frac{1}{\delta} + \log\log\frac{1}{\max\{\eta_u, \eta_r\}}\right)\right).$$

5 Conclusions

Our paper draws on the economic approach to fair classification. It initiates the discussion of the impact that the regulator's misconception about the population characteristics may have on the protected groups' well-being. Beyond that, we believe that our WE-fairness can serve as an anchor for grounding other fairness stances.

We have prioritized clarity over generality and focused on binary classification. Nevertheless, our results are more general. The key technical Proposition 2 can be extended to multiple classes (e.g., several loans with different periods and interest rates).[11] All other results (the mismatch analysis and the algorithms) are, essentially, corollaries of Proposition 2 and extend straightforwardly. Noisy utilities and revenues can be assumed for free if we interpret u, v, and r in all formulas as the conditional expectations for a given triplet (X, A, Y). Allowing for utilities with negative expectations also does not alter the statements. We see considerable scope for follow-up work. One prominent direction is to understand how the "price of fairness" is distributed among the parties, e.g., by how much the bank's revenue and the advantaged group's welfare drop.

[11] As in the binary case, there is one threshold λ_a per group a, but now the revenue r and the utilities $\overline{u}$ depend on the class. The optimal WE-classifier selects a class c with the maximal $r - \lambda_a \cdot \overline{u}$.

Acknowledgements

We are grateful to Lily Hu, Nikita Kalinin, Alexander Nesterov, Margarita Niyazova, and Ivan Susin for insightful discussions and pointers to the relevant literature. We also thank the anonymous reviewers for their helpful comments and clarifications and Lillian Bluestein for proofreading.

The work of M. Tennenholtz is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 740435). F. Sandomirskiy is partially supported by the Lady Davis Foundation, by Grant 19-01-00762 of the Russian Foundation for Basic Research, and by Basic Research Program of the National Research University Higher School of Economics. This work was done while O. Ben-Porat and F. Sandomirskiy were at the Technion—Israel Institute of Science and were partially funded by the European Union's Horizon 2020 research and innovation programme (grant agreement n° 740435).

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453* .

Balcan, M.-F. F.; Dick, T.; Noothigattu, R.; and Procaccia, A. D. 2019. Envy-free classification. In *Advances in Neural Information Processing Systems*, 1238–1248.

Benade, G.; Kazachkov, A. M.; Procaccia, A. D.; and Psomas, C.-A. 2018. How to make envy vanish over time. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, 593–610. ACM.

Brandt, F.; Conitzer, V.; Endriss, U.; Lang, J.; and Procaccia, A. D. 2016. *Handbook of computational social choice*. Cambridge University Press.

Caragiannis, I.; Kurokawa, D.; Moulin, H.; Procaccia, A. D.; Shah, N.; and Wang, J. 2019. The unreasonable fairness of maximum Nash welfare. *ACM Transactions on Economics and Computation (TEAC)* .

Cohler, Y. J.; Lai, J. K.; Parkes, D. C.; and Procaccia, A. D. 2011. Optimal envy-free cake cutting. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Corbett-Davies, S.; and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* .

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797–806. ACM.

Doleac, J. L.; and Hansen, B. 2016. Does "ban the box" help or hurt low-skilled workers? Statistical discrimination and employment outcomes when criminal histories are hidden. Technical report, National Bureau of Economic Research.

Dutta, B.; and Ray, D. 1989. A concept of egalitarianism under participation constraints. *Econometrica: Journal of the Econometric Society* 615–635.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. ACM.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. 2018. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, 119–133.

Elzayn, H.; and Fish, B. 2020. The effects of competition and regulation on error inequality in data-driven markets. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 669–679. Association for Computing Machinery.

Finocchiaro, J.; Maio, R.; Monachou, F.; Patro, G. K.; Raghavan, M.; Stoica, A.-A.; and Tsirtsis, S. 2020. Fairness and Discrimination in Mechanism Design and Machine Learning. In *AI for Social Good Workshop*.

Gal, Y. K.; Mash, M.; Procaccia, A. D.; and Zick, Y. 2016. Which is the fairest (rent division) of them all? In *Proceedings of the 2016 ACM Conference on Economics and Computation*, 67–84. ACM.

Gölz, P.; Kahng, A.; and Procaccia, A. D. 2019. Paradoxes in Fair Machine Learning. In *Advances in Neural Information Processing Systems*, 8340–8350.

Guruswami, V.; Hartline, J. D.; Karlin, A. R.; Kempe, D.; Kenyon, C.; McSherry, F.; and McSherry, F. 2005. On profit-maximizing envy-free pricing. In *ACM-SIAM Symposium on Discrete Algorithms*.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.

Heidari, H.; Ferrari, C.; Gummadi, K.; and Krause, A. 2018. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Advances in Neural Information Processing Systems*, 1265–1276.

Heidari, H.; Gummadi, M. L. K. P.; and Krause, A. 2019. Moral framework for understanding fair ML through economic models of equality of opportunity. In *FAT* 19*, 181–190.

Hossain, S.; Mladenovic, A.; and Shah, N. 2020. Designing fairly fair classifiers via economic fairness notions. In *Proceedings of The Web Conference 2020*, 1559–1569.

Hu, L.; and Chen, Y. 2020. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 535–545. Association for Computing Machinery. ISBN 9781450369367.

Kalai, E.; and Smorodinsky, M. 1975. Other solutions to Nash's bargaining problem. *Econometrica* 43(3): 513–518.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9.

Li, J.; and Xue, J. 2013. Egalitarian division under Leontief preferences. *Economic Theory* 54(3): 597–622.

Liu, L.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed impact of fair machine learning. In *International Conference on Machine Learning*, 3156–3164.

Lykouris, T.; and Vassilvitskii, S. 2018. Competitive caching with machine learned advice. *arXiv preprint arXiv:1802.05399* .

Medina, A. M.; and Vassilvitskii, S. 2017. Revenue optimization with approximate bid predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 1856–1864.

O'Neill, C. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishers.

Pazner, E. A.; and Schmeidler, D. 1978. Egalitarian equivalent allocations: A new concept of economic equity. *The Quarterly Journal of Economics* 92(4): 671–687.

Plaut, B.; and Roughgarden, T. 2018. Almost envy-freeness with general valuations. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, 2584–2603. Society for Industrial and Applied Mathematics.

Purohit, M.; Svitkina, Z.; and Kumar, R. 2018. Improving online algorithms via ml predictions. In *Advances in Neural Information Processing Systems*, 9661–9670.

Rambachan, A.; Kleinberg, J.; Ludwig, J.; and Mullainathan, S. 2020a. An Economic Perspective on Algorithmic Fairness. In *AEA Papers and Proceedings*, volume 110, 91–95.

Rambachan, A.; Kleinberg, J.; Mullainathan, S.; and Ludwig, J. 2020b. An economic approach to regulating algorithms. Technical report, National Bureau of Economic Research.

Rawls, J. 2009. *A theory of justice*. Harvard University Press.

Rockafellar, R. T. 2015. *Convex analysis*. Princeton University Press.

Roemer, J. E. 1986. Equality of resources implies equality of welfare. *The Quarterly Journal of Economics* 101(4): 751–784.

Ustun, B.; Liu, Y.; and Parkes, D. 2019. Fairness without harm: Decoupled classifiers with preference guarantees. In *International Conference on Machine Learning*, 6373–6382.

Zafar, M. B.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017. From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, 229–239.