Protecting the Protected Group: Circumventing Harmful Fairness

Omer Ben-Porat* Fedor Sandomirskiy† Moshe Tennenholtz‡

May 28, 2019

Abstract

Machine Learning (ML) algorithms shape our lives. Banks use them to determine if we are good borrowers; IT companies delegate them recruitment decisions; police apply ML for crime-prediction, and judges base their verdicts on ML. However, real-world examples show that such automated decisions tend to discriminate protected groups. This generated a huge hype both in media and in the research community. Quite a few formal notions of fairness were proposed, which take a form of constraints a "fair" algorithm must satisfy. We focus on scenarios where fairness is imposed on a self-interested party (e.g., a bank that maximizes its revenue). We find that the disadvantaged protected group can be worse off after imposing a fairness constraint. We introduce a family of *Welfare-Equalizing* fairness constraints that equalize per-capita welfare of protected groups, and include *Demographic Parity* and *Equal Opportunity* as particular cases. In this family, we characterize conditions under which the fairness constraint helps the disadvantaged group. We also characterize the structure of the optimal *Welfare-Equalizing* classifier for the self-interested party, and provide an LP-based algorithm to compute it. Overall, our *Welfare-Equalizing* fairness approach provides a unified framework for discussing fairness in classification in the presence of a self-interested party.

1 Introduction

At first glance, algorithms may seem obviously free of human biases as sexism or racism. However, in many situations they are not: the automated recruiting tool used by Amazon was favoring men [8]; judges in the US use the COMPAS algorithm to estimate the probability that the defendant will re-offend while this algorithm is biased against black people [18]. See O'Neill [22] for many more examples. These challenges call for imposing fairness constraints on algorithms design and, in particular, on ML classifiers. Naturally, fairness is costly as confirmed by the literature on the price of fairness (see, e.g., [2] and [4]); hence, before imposing any fairness constraints it is crucial to understand who incurs these costs. It is not surprising that imposing constraints on the behavior of the decision-maker (the bank giving loans in our canonical example) may harm the well-being of the decision-maker. For instance, Corbett-Davies et al. [6] study this effect in the context of fair classification.

While the above is by now well known, we go further and ask: *can fairness harm the well-being of those whom it is designed to protect?*, i.e., can it harm the well-being of the disadvantaged

*Technion – Israeli Institute of Technology (omerbp@campus.technion.ac.il), Corresponding author

†Technion – Israeli Institute of Technology, and Higher School of Economics, St.Petersburg, Russia (fsandomirskiy@hse.ru)

‡Technion – Israeli Institute of Technology (moshet@ie.technion.ac.il)

group? Such a possibility might seem counter-intuitive, but keep in mind that the objective of the decision-maker is different from helping the disadvantaged group: if we impose a fairness constraint on a revenue-maximizing bank it will adjust its policy to accommodate the constraint but will still maximize its revenue; thus, it will try to reallocate the "price of fairness" to the borrowers, and, possibly, to the disadvantaged subgroup. Indeed, there are real-life examples showing that fairness can be harmful for welfare, e.g., Doleac and Hansen [8] show that the "ban the box" policy, adopted by the United States and preventing employers from seeing criminal backgrounds of applicants, decreased the chances of discriminated minorities (blacks and poor) to get a job. In this paper we take a utilitarian approach and model utility explicitly, by quantifying the well-being of agents for either label they obtain (i.e., whether they get the loan or not). Further, we model a self-interested decision-maker, who aims to maximize a quite general revenue-like objective function subject to the imposed fairness constraint. Our formal model is given in Section 2.

We then construct fairness criteria that are driven by welfare, through the notion of agents' utility function. In Section 3 we introduce a broad family of *Welfare-Equalizing fairness concepts*, equalizing the welfare among protected groups.[1] As we show, special cases from this family include popular fairness concepts (*Demographic Parity* and *Equal Opportunity*) that are obtained by aligning their statistical requirements with a particular selection of an agents' utility function. Most of the previous literature study fairness constraints that address welfare only implicitly, and are given in the language of classification – using a statistical metric (e.g., [16, 28]). In contrast, we draw on the approach of recent work [6, 12, 13] and focus on welfare as the key ingredient of fairness. Indeed, Our paper is the first one, where the language of utilities is used in full force, allowing us to analyze the structural properties of a broad family of rules.

Welfare-Equalizing fairness gives a general recipe to define the fairness concept specially tailored for a particular problem instance, if there is some understanding on how different outcomes affect the well-being of agents. For example, it suggests how to naturally extend the existing fairness notions if we have some additional information about borrowers (e.g., loans may differ in the amount of money, duration, payment schemes, etc.). In contrast to Heidari et al. [13], who unify existing fairness notions based on theories of justice from political philosophy, our unified framework of *Welfare-Equalizing* fairness allows one to analyze the properties of optimal fair rules and their implications for the well-being of protected groups. In Subsection 3.1 we characterize the optimal policy of the bank (self-interested decision-maker) when a *Welfare-Equalizing* fairness constraint is imposed on its behavior. Later, in Subsection 3.2 we describe how to compute an optimal *Welfare-Equalizing* classifier.

In Section 4 we study the implications of the class of *Welfare-Equalizing fairness concepts* and other fairness concepts on the *disadvantaged* group. We call the protected group having the lowest welfare before imposing a fairness constraint disadvantaged and say that the fairness constraint harms the protected group if imposing the constraint decreases its welfare. Our main result is that *Welfare-Equalizing fairness helps the disadvantaged group*. In comparison, the fairness constraint of *Unawareness* may harm both protected groups and the bank; *Demographic Parity* and *Equal Opportunity* could be harmful if their statistical constraint and the agents' utility function are misaligned. Later, we prove that we do not need to know the exact utilities of the agents in order

[1] An idea of maximizing the minimal welfare of the protected group was suggested in [12, 13] as a possible approach to fairness. We strengthen this desideratum by asking for equalizing welfare of subgroups defined by a protected attribute. This new normative condition allows one to separate the *constraint of fairness* from the *selfish objective of the decision-maker* and thus allows one to analyze how the decisions change after imposing this constraint.

to ensure that the disadvantaged group will not be harmed: the result turns out to be robust. If we use one welfare to define the fairness constraint and another one to measure its effect on the protected group, then we will conclude that fairness helps the disadvantaged group provided that the two welfare functions agree on which group is disadvantaged.

2 Model

We consider a general classification problem, where agents have ex-ante non-observable "quality" correlated with observable attributes. A classifier is predicting "quality" based on statistical data; misclassification is costly while a good guess is profitable. We will keep using a metaphor of a bank that predicts reliability of a borrower and makes a lending decision; however, the same setting captures student admissions, recruiting, accessing the recidivism risk for a criminal and so on.

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 the binary[2] protected 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 make any assumptions on $\mathcal{X}$. We also assume that each borrower has a *utility* for receiving the loan or not receiving it. Since utility is part of our approach to fairness, we defer its formal definition to Section 3.

The statistical characteristics of the population are described by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$; so $X = X(\omega)$ and $A = A(\omega)$ are random variables on Ω. Another random variable $Y = Y(\omega) \in \{0,1\}$ describes whether a given borrower will pay back if he or she is given the money. By its nature, Y is unobservable. We call borrowers with $Y = 1$ and $Y = 0$, *good* and *bad*, respectively.

Further, 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 attributes; we will denote it by $p(x, a) = \mathbb{P}(Y = 1 \mid X = x, A = a)$.

The banks makes lending decisions based on A and X 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[3] of $\alpha_+(X) > 0$ to the bank while each bad borrower leads to a loss of $\alpha_-(X) > 0$; we assume that $\alpha_\pm$ are bounded functions of X. The bank aims to maximize its revenue, which depends on the choice of a classifier c, and is defined by

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

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

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

[2]The assumption of dichotomy of A is made for simplicity. Extension of all results to a non-binary case (ethnicity) is straightforward. A more general setting where protected classes are defined by combinations of attributes ("a single black women with 3 children") can also be captured by introducing an auxiliary non-binary protected attribute that represents membership to each of the protected groups.

[3]In contrast to the rest of the literature, we allow dependence of bank's revenue on non-protected 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.

hence, we can rearrange Equation (1) by taking conditional expectation with respect to A, X and rewriting $R(c)$,

$$R(c) = \mathbb{E}\left[c(X,A)\big(\alpha_+(X)p(X,A) - \alpha_-(X)(1-p(X,A))\big)\right] = \mathbb{E}\left[r(X,A)c(X,A)\right]. \tag{2}$$

2.1 Optimal Unconstrained Classifier and Example

We now exemplify our setting and notations, by considering the *optimal* classifier. If the bank is free to choose any classifier, then the optimal c maximizing $R(c)$ given in Equation (2) has a simple form [6]. 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 $t(x)$. Consequently, the optimal lending policy is given by the threshold classifier c^*_{unc}; loans are given ($c^*_{\text{unc}}(x,a) = 1$) to all borrowers with $p(x,a) > t(x)$ and all borrowers with $p(x,a) \le t(x)$ are rejected ($c^*_{\text{unc}}(x,a) = 0$). [4]

This threshold behavior resembles the widespread usage of the credit score: a loan is given if the score is above some threshold. The probability $p(x,a)$ can thus be interpreted as an ideal credit score for unconstrained banks.

The following example illustrates that optimal unconstrained behavior of the bank can discriminate one of the groups even if *groups are of equal size and contain the same fraction of good borrowers.*

Example 1. Let A and X be binary, and let all four combinations be equally likely. The probability $p(x,a)$ of being a good borrower is given by the matrix

	$X=0$	$X=1$
$A=0$	0.4	0.6
$A=1$	0	1

.

We see in the group $A = 0$ the attribute X poorly separates good and bad borrowers, while it is a perfect predictor of creditworthiness for $A = 1$. If losses from a defaulting client are equal to the revenue from two borrowers paying back, e.g. $\alpha_- = 2 \cdot \alpha_+$, then we get the threshold $t = \frac{2}{3} \simeq 0.66$; thus, the optimal policy of the bank is giving loans to applicants with $(X, A) = (1, 1)$ only. As a result, *no loans are given in the group* $A = 0$, although the prior distribution shows that in total 1/2 of borrowers from each of the groups $A = 0$ and $A = 1$ are good. This contradicts intuitive understanding of fairness.

3 Welfare-Equalizing Fairness

The idea of a welfare-equalizing approach is to find the outcome that aligns the well-being of all the parties. The well-being of individuals is measured by means of a utility function (it is assumed that each agent can potentially assign a numerical value to each of the outcomes showing how valuable this outcome is for her) and well-being of groups by means of social welfare which aggregates utilities of the members. Welfare-equalizing, our approach to fairness, has a long history in normative economics [24, 27] (where it is known under the name of egalitarianism) and political philosophy [25]; it was used for fair resource allocation without money transfers [19], in the field of cooperative games [9] and bargaining problems [15].

[4] For definiteness, we assume that if the bank is indifferent to the two decisions (the knife-edge case $p(x,a) = t(x)$), it chooses the one with less loans given (e.g., this policy minimizes paperwork).

In order to apply a welfare-equalizing approach to our problem, we shall first agree how to measure welfare of individuals. We assume that each individual obtains a utility of u_+ for receiving the loan and u_- if his/her loan request is rejected. These utilities u_+ and u_-, on which we elaborate shortly, can depend on X and A (see example below, where X contains the amount of money the individual wishes to loan), but what is more important — they must depend on the non-observable quality of borrower, namely $u_\pm = u_\pm(x, a, y)$.

Indeed, while good borrowers benefit from receiving loans, bad borrowers do not. As argued in Liu et al. [20], when bad borrowers are given a loan they lose the borrowed money at the end (the property bought will be foreclosed by the bank) and also harm their credit history.

For a given utility-function $u = (u_+, u_-)$ and a classifier c, we define the utilitarian welfare of the whole population as

$$W_{u,c} = \mathbb{E}\left[u_+(X, A, Y)c(X, A) + u_-(X, A, Y)(1 - c(X, A))\right]. \tag{3}$$

Further, we denote by $W_{u,c}(a)$ the welfare of individuals in the protected group a, for $a \in \{0, 1\}$, which is simply the conditional expectation of $W_{u,c}$ given $A = a$. We are now ready to define the *Welfare-Equalizing* fairness condition.

Definition 2. *Given a utility-function $u = (u_+, u_-)$, a classifier c is u-Welfare-Equalizing if*

$$W_{u,c}(0) = W_{u,c}(1), \tag{4}$$

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

Welfare-Equalizing fairness (hereinafter denoted by WE for abbreviation) allows one to analyze (existing) fairness concepts in a unified way and also opens some new possibilities. For instance, [5]

- The fairness concept of *Demographic Parity* (e.g., [1, 10]), hereinafter denoted DP for brevity, imposes constraints on the outcomes of decisions within the two groups: it 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 $u_+ \equiv 1$ and $u_- \equiv 0$.

- Motivated by drawbacks of DP, Hardt et al. [11] concluded that good and bad borrowers within protected groups must be treated separately and introduced the concept of *Equal Opportunity* (hereinafter EO). Under this fairness concept, the fraction of good borrowers who get loans must be the same in the two subgroups. Formally, a classifier c satisfies EO if

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

  We recover EO by setting

  $$u_+(y) = \begin{cases} 1, & y = 1 \\ 0, & y = 0 \end{cases}, \qquad u_- \equiv 0.$$

[5] We provide a brief introduction to some popular fairness concept in Section 8.

- If a classifier is u-WE with respect to all utility functions $u_+(y=1)=\alpha$, $u_+(y=0)=\beta$, and $u_-\equiv 0$ for $\alpha\geq\beta\geq 0$, we get a fairness concept called *Equalized odds* [11]. It combines the condition of EO with equalizing false-positive rates: $\mathbb{E}\left[c(X,A)\mid Y=y, A=0\right]=\mathbb{E}\left[c(X,A)\mid Y=y, A=1\right]$, for $y\in\{0,1\}$.

- 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

$$u_+(x,y)=\begin{cases} m(x), & y=1 \\ 0, & y=0 \end{cases}, \qquad u_-\equiv 0.$$

  We can capture any other heterogeneity in a similar way (e.g., different interest rates, time-period, and payment schedules).

Notice that in all the above examples u_- is identically zero. This is not a coincidence: without loss of generality, $u_-\equiv 0$ for any "reasonable" WE fairness concept. Typically, and under all the fairness criteria above, the *zero classifier* $c\equiv 0$ (giving no loans at all) is considered fair. The following lemma shows that if the zero classifier is u-WE, then we can shift the utilities to have $u_-=0$.

Lemma 3. *If the zero classifier belongs to* $\mathrm{WE}(u)$ *with* $u=(u_+,u_-)$*, then*

$$\mathrm{WE}(u)=\mathrm{WE}(u'), \quad \textit{where } u'=(u'_+,\,0) \textit{ with } u'_+=u_+-u_-$$

Hence $u_-\equiv 0$ is a normalization-condition: prior to borrowing money everybody is at zero utility level, and then the *WE approach equalizes utilitarian gains from borrowing.* Motivated by Lemma 3, we henceforth assume that

$$u=(u_+,0) \quad \text{and} \quad u_+\geq 0.$$

Non-negativity of u_+ can be regarded as a rationality assumption on borrowers: no rational agent applies for a loan if she/he expects that getting the loan brings negative utility while not getting it gives 0. In what follows, we drop "+" in u_+ and identify u_+ and $u=(u_+,0)$ for brevity; so $W_{u,c}=\mathbb{E}\left[u(X,A,Y)c(X,A)\right]=\mathbb{E}\left[u_+(X,A,Y)c(X,A)\right]$.

3.1 Structural Properties of the Optimal Welfare-Equalizing classifier

In this subsection, we explore the structure of optimal WE classifiers. The computational problem is deferred to the next subsection.

For a fixed 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. The following proposition shows that any such classifier is a generalized threshold classifier.

To ease notation, we denote by $\overline{u}(x,a)$ the average utility of a borrower associated with (x,a) $\overline{u}(x,a)=\mathbb{E}[u\mid X=x, A=a]=u(x,a,1)p(x,a)+u(x,a,0)(1-p(x,a))$. Further, we denote by $R^*_a(w)$ the maximal revenue that the bank could extract from the group $\{A=a\}$ at the welfare level $W_{u,c}(a)=w$. Formally, $R^*_a(w)=\max_{\{c\colon\mathcal{X}\to[0,1]\,\mid\,W_{u,c}(a)=w\}}\mathbb{E}\left[r(X,A)\cdot c(X)\mid A=a\right]$.

Proposition 4. *The optimal u-WE classifier $c^*_{\mathrm{WE}(u)}$ exists and has the following form:*

$$c^*_{\mathrm{WE}(u)}(x,a) = \begin{cases} 1 & r(x,a) > \lambda_a \overline{u}(x,a) \\ 0 & r(x,a) < \lambda_a \overline{u}(x,a) \\ \tau_a(x) & r(x,a) = \lambda_a \overline{u}(x,a) \end{cases}. \tag{5}$$

Group-dependent thresholds λ_a, $a \in \{0,1\}$ belong to the super-gradient[6] *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 $\mathbb{P}(A=0)R^*_0(w) + \mathbb{P}(A=1)R^*_1(w)$.*

Functions $\tau_a : \mathcal{X} \to [0,1]$ are arbitrary[7] *up to the constraint that $c^*_{\mathrm{WE}(u)}$ provides the desired welfare level w^* for both groups: $w^* = W_{u,c^*_{\mathrm{WE}(u)}}(0) = W_{u,c^*_{\mathrm{WE}(u)}}(1)$.*

Proof sketch of Proposition 4. The revenue maximization over $c \in \mathrm{WE}(u)$ can be represented as a two-stage procedure: 1) find the revenue-maximizing classifier in each of the subgroups $A = a$ given the welfare level w; 2) optimize over w. The welfare constraint in 1) can be internalized using the Lagrangian approach; the corresponding Largange 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 provides the threshold structure for the optimal classifier from 1) inherited by $c^*_{\mathrm{WE}(u)}$.

Since the linear program 1) is infinite-dimensional and $R^*_a(w)$ may be non-differentiable, the formal proof requires some functional-analytic arguments presented in the Appendix. □

Examples Proposition 4 immediately provides the optimal classifiers under DP and EO: for the former, it is given by the "additive perturbation" of the optimal unconstrained classifier, while the latter is obtained by "multipliciative perturbation."

More precisely, for the optimal classifier c^*_{parity} satisfying DP, there exist group-dependent constants $\lambda_a \in \mathbb{R}$, $a \in \{0,1\}$ such that

$$c_{\text{parity}}(x,a) = \begin{cases} 1 & p(x,a) > t(x,a) + \frac{\lambda_a}{\alpha_+(x)+\alpha_-(x)} \\ 0 & p(x,a) < t(x,a) + \frac{\lambda_a}{\alpha_+(x)+\alpha_-(x)} \end{cases}.$$

Indeed, this recovers the result of Corbett-Davies et al. [6]. For the optimal EO classifier $c^*_{\text{opportunity}}$, there exist $\lambda_a \in \mathbb{R}$, $a \in \{0,1\}$ such that

$$c^*_{\text{opportunity}}(x,a) = \begin{cases} 1 & p(x,a)\left(1 - \frac{\lambda_a}{\alpha_+(x)+\alpha_-(x)}\right) > t(x,a) \\ 0 & p(x,a)\left(1 - \frac{\lambda_a}{\alpha_+(x)+\alpha_-(x)}\right) < t(x,a) \end{cases}.$$

3.2 Computing the Optimal Welfare-Equalizing Classifier

After discussing the structure of the optimal WE classifier, we now explain how it can be computed in two common cases. We handle finite $\mathcal{X}$ and describe how to extend the analysis of the finite case to infinite $\mathcal{X}$ under mild assumptions.

[6] 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 $f(t') \leq f(t) + q(t'-t)$ for all t'. If f is continuous, then for any t the super-gradient is non-empty, see [26].

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

When $\mathcal{X}$ is finite, we can use the following linear program (LP) to find $c^*_{\mathrm{WE}(u)}$.

$$\begin{aligned} &\text{maximize} \sum_{x,a} \mathbb{P}(X = x, A = a) \cdot r(x,a) \cdot c(x,a) \\ &\text{subject to} \sum_{x} \mathbb{P}(X = x \mid A = 0) \cdot \overline{u}(x,0) \cdot c(x,0) = \sum_{x} \mathbb{P}(X = x \mid A = 1) \cdot \overline{u}(x,1) \cdot c(x,1) \\ &\qquad\qquad 0 \le c(x,a) \le 1, x \in \mathcal{X}, a \in A \end{aligned} \tag{6}$$

The objective function of the LP is $R(c)$, the bank's revenue, defined formally in Equation (2). We wish to maximize it over the class of functions $\{\mathcal{X} \times A \to [0,1]\}$. The first constraint is welfare-equalizing (see Equation (4)): the welfare of group $A = 0$ should be equal to the welfare of group $A = 1$. The remaining constraints are range constraints, assuring that the classifier is indeed in the $[0,1]$ segment for every pair (x,a). We can thus compute $c^*_{\mathrm{WE}(u)}$ efficiently, in polynomial time in $|\mathcal{X}|$.

In case of infinite $\mathcal{X}$, under standard regularity assumptions on $r(x,a)$ and $\overline{u}(x,a)$, one can simply discretize the space and execute the LP in Equation (6) to find an approximately optimal revenue, with approximately equal welfare. This process can also be slightly modified to allow approximate revenue but (exact) equal welfare: let c be a classifier that approximates both revenue and the welfare difference between the two groups. If $W_{u,c}(0) \le W_{u,c}(1)$ we define c' by $c'(x,0) = c(x,0)$ and $c'(x,1) = c(x,1) \cdot \frac{W_{u,c}(0)}{W_{u,c}(1)}$ (the case of $W_{u,c'}(0) > W_{u,c'}(1)$ is symmetric). Indeed, this normalization equalizes the welfare of the two groups while not damaging the revenue too much.

4 Implications of fairness constraints

In this section we discuss the implications of applying WE fairness. We call a group $\{A = a\}$ the *disadvantaged* protected group (or u-disadvantaged to emphasize the dependence on u) if under the optimal unconstrained classifier, the welfare of $\{A = a\}$ is lower than welfare of the other group $\{A = 1 - a\}$. Formally,

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

where W is defined in Equation (3). We say that a *fairness constraint harms the group* $\{A = a\}$ if there are cases where $W_{u,c^*}(a) < W_{u,c^*_{\mathrm{unc}}}(a)$, and where c^* is the optimal fair classifier.

In order to put the analysis of WE fairness in context, we present and analyze *Unawareness*, perhaps the most intuitive fairness criterion. A classifier c satisfies unawareness if $c(x,0) = c(x,1)$ for all $x \in \mathcal{X}$. Informally, to deliver a "fair" outcome to both groups $A = 0$ and $A = 1$, the classifier c must ignore the protected attribute A. We begin by showing that *Unawareness* can make all three parties strictly worse off: both groups and the bank, *regardless* of the utility function u. The following example shows this phenomenon when information losses caused by unawareness are significant: the interpretation of non-protected attributes X depends on the protected[8] attribute A and without knowing A the classifier cannot achieve good separation of good and bad borrowers thus giving loans becomes too risky.

Example 5. Let $u \ge 0$ be an arbitrary utility function, and assume that $\mathcal{X} = \{0,1\}$ and all the four combinations of attributes (x,a) are equally likely in the population. The fraction $p(x,a)$ of good

[8] For example, having children correlates with spending more time at work for men and has negative correlation for women [23].

borrowers is given by the following matrix:

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

Notice that the fraction of good and bad borrowers is the same in both groups. Further, $X = 1$ is a positive signal about the quality of a borrower in the group $\{A = 0\}$ and a negative one for $\{A = 1\}$.

The optimal unconstrained classifier c^*_{unc} gives loans to agents with $p(x, a) \geq t(x)$; thus, if the threshold t is between $\frac{1}{3}$ and $\frac{2}{3}$ (for example, $t = \frac{3}{5}$ if the revenue α_+ from giving money to a good borrower equals $\frac{2}{3}$ of the losses α_- from a bad one), then agents with (X, A) equal to $(0, 0)$ or $(1, 1)$ get loans under c^*_{unc}. So one half of the members in each group receives loans and the bank gets a positive revenue.

After imposing *Unawareness*, the optimal classifier c^*_{unaware} compares the average fraction of good borrowers with given x, namely $\overline{p}(x) = \mathbb{P}(A = 0)p(x, 0) + \mathbb{P}(A = 1)p(x, 1)$, with $t(x)$. In our example $\overline{p}(x) = \frac{1}{2}$ for every x, and thus no loans are given for $t > 1/2$; thus for $t \in (1/2, 2/3)$ unawareness pushes the welfare of both groups to zero as well as the bank's revenue.

Having demonstrated that fairness constraints can be harmful for those it tries to protect, we now develop the general theory for WE fairness concepts. We show that if the concept of fairness matches the way we are measuring the welfare of borrowers, then *WE fairness always helps the disadvantaged group.*

Theorem 6. *The optimal u-WE classifier makes the u-disadvantaged protected group weakly better off. Formally,*

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

Moreover, any borrower from the u-disadvantaged group who receives a loan under the unconstrained classifier, receives it under optimal u-WE.[9] *Formally, for all $x \in \mathcal{X}$ it holds that*

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

Proof. By Proposition 4, the welfare $w^* = W_{u,c^*_{\text{WE}(u)}}(0) = W_{u,c^*_{\text{WE}(u)}}(1)$ achieved by u-WE classifier maximizes the total 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 and thus w^* lies between their maxima. These maxima are attained at welfare of the optimal unconstrained classifier and thus w^* is between $W_{u,c^*_{\text{unc}}}(a),\ a \in \{0, 1\}$. See Figure 1 for illustration.

Individual guarantees follow from the threshold structure in Equation (5) of $c^*_{\text{WE}(u)}$: since w^* is above the maximum of $R^*_a(w)$, the subgradient contains $\lambda_a \leq 0$ and thus $c^*_{\text{unc}}(x, a)$ (which corresponds to zero λ_a) is below $c^*_{\text{WE}(u)}(x, a)$. □

Theorem 6 emphasizes an important aspect of WE fairness: the well-being of the disadvantaged protected group will always (weakly) increase under the optimal classifier after imposing WE fairness, while the well-being of the advantageous protected group will always (weakly) decrease; hence, the optimal u-WE classifier balances the well-being of the three parties: the two groups and the bank. Theorem 6 provides insights about EO too: if the borrower's utility function is $u(x, a, y) = y$, then

[9] While our paper is focused on statistical notions of fairness, we stress that this result provides stronger "individual" guarantees that no particular agent from the disadvantaged group will be harmed by fairness.

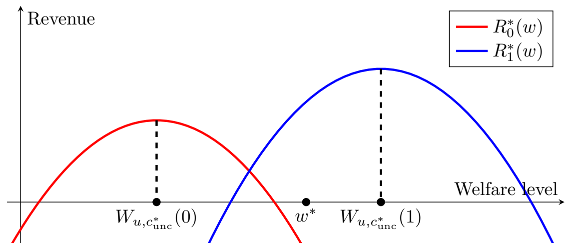

Figure 1: Intuition for the proof of Theorem 6. The red curve illustrates the revenue $R_0^*(w)$ from the disadvantaged group $\{A = 0\}$ at every possible welfare level, and the blue curve illustrates $R_1^*(w)$ from the advantageous group $\{A = 1\}$. The subgroup revenues $R_a^*(w)$ are concave functions 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 $\mathbb{P}(A = 1) = \frac{2}{3}$.

EO guarantees to make the disadvantaged better off. However, if we impose EO and use a different utility function to measure welfare, that may no longer hold. As for DP, even under $u(x, a, y) = y$ we are not guaranteed to improve the welfare of the disadvantaged group.

Ultimately, we note that in some cases the exact borrowers' utility function u is unknown, and a theoretical approximation $\tilde{u}$ is used instead. One criticism of our model could be that, under the unobservable utility scenario, the WE fairness is not meaningful. Theorem 6 ensures that such a fairness concept weakly increases $\tilde{u}$-welfare of the $\tilde{u}$-disadvantaged group, but we would like to ensure that the same holds with respect to "real" but unobserved utilities u. The second statement of Theorem 6 together with non-negativity of u imply robust guarantees of this kind.

Corollary 7. *If $\tilde{u}$ and u agree on which group is disadvantaged, then the $\tilde{u}$-WE classifier weakly increases u-welfare of the disadvantaged group, i.e.,*

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

5 Conclusions

In this paper we draw on an economic approach to design a family of fairness constraints. In particular, we assume an agents' utility function and a selfish decision-maker, and design a fairness constraint under which the disadvantaged group is always better off. We sketched the structure of the optimal WE fairness for the bank, showing that our proposed family contains celebrated concepts as special cases. We also show that the WE fairness approached is robust to imperfect information of agents' utility.

We see considerable scope for follow-up work. One prominent direction is the following: our model is Bayesian, and assumes a distribution over agents and their competence. Indeed, this is assumed to neutralize possible bias in data. It would be interesting to consider WE fairness in prior-free settings, where the WE constraint (as well as the bank's revenue) should hold up to some approximation factor. This process might also include learning the utility function u.

Another promising direction is to estimate the "price of fairness" and its distribution among the parties involved. How much of the bank's revenue can drop because of fairness constraint? How much the welfare of a advantaged protected group can decrease?

Acknowledgements

The authors are grateful to Nikita Kalinin, Alexander Nesterov, Margarita Niyazova, and Ivan Susin for insightful discussions and pointers to the relevant literature.

The work of O. Ben-Porat and 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 F. Sandomirskiy was at the Technion – Israeli Institute of Science and was partially funded by the European Union's Horizon 2020 research and innovation programme (grant agreement n° 740435).

References

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

[2] D. Bertsimas, V. F. Farias, and N. Trichakis. The price of fairness. *Operations research*, 59(1): 17–31, 2011.

[3] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[4] I. Caragiannis, C. Kaklamanis, P. Kanellopoulos, and M. Kyropoulou. The efficiency of fair division. *Theory of Computing Systems*, 50(4):589–610, 2012.

[5] P. Collinson. How an eu gender equality ruling widenedinequality. *The Guardian*, 14 January 2017.

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

[7] J. Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. *San Fransico, CA: Reuters. Retrieved on October*, 9:2018, 2018.

[8] J. L. Doleac and B. Hansen. 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, 2016.

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

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226. ACM, 2012.

[11] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

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

[13] H. Heidari, M. L. K. P. Gummadi, and A. Krause. Moral framework for understanding fair ml through economic models of equality of opportunity. In *FAT19: Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190, 2019.

[14] B. C. John. *A Course in functional analysis.* New York: Springer-Verlag, 1985.

[15] E. Kalai, M. Smorodinsky, et al. Other solutions to nash's bargaining problem. *Econometrica*, 43(3):513–518, 1975.

[16] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[17] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[18] J. Larson, S. Mattu, L. Kirchner, and J. Angwin. How we analyzed the compas recidivism algorithm. *ProPublica (5 2016)*, 9, 2016.

[19] J. Li and J. Xue. Egalitarian division under leontief preferences. *Economic Theory*, 54(3): 597–622, 2013.

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

[21] G. Meyers and I. Van Hoyweghen. Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. *Science as Culture*, 27(4):413–438, 2018.

[22] C. O'Neill. Weapons of math destruction. *How Big Data Increases Inequality and Threatens Democracy*, 2016.

[23] K. Parker and W. Wang. Modern parenthood. *Pew Research Center's Social & Demographic Trends Project*, 14, 2013.

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

[25] J. Rawls. *A theory of justice.* Harvard university press, 2009.

[26] R. T. Rockafellar. *Convex analysis.* Princeton university press, 2015.

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

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

6 Omitted Proofs

Proof of Lemma 3. For any classifier c we have $W_{u,c}(a) = W_{u',c}(a) + \mathbb{E}[u_- \mid A = a]$. Since $0 \in \mathrm{WE}(u)$, the second summands coincide for $a \in \{0, 1\}$: $\mathbb{E}[u_- \mid A = 0] = \mathbb{E}[u_- \mid A = 1]$. Thus $W_{u,c}(0) = W_{u,c}(1)$ if and only if $W_{u',c}(0) = W_{u',c}(1)$, which is equivalent to $\mathrm{WE}(u) = \mathrm{WE}(u')$. $\square$

Proof of Proposition 4. The revenue maximization under the WE constraint can be split into two subsequent maximization problems:

$$\max_{c \in \mathrm{WE}(u)} R(c) = \max_{w \in \mathbb{R}} \max_{\substack{c\colon \mathcal{X}\times A \to [0,1] \\ W_{u,c}(0) = W_{u,c}(1) = w}} \mathbb{P}(A = 0) R_0(c) + \mathbb{P}(A = 1) R_1(c).$$

First, the bank finds the revenue-maximizing classifier $c^*_{a,w} : \mathcal{X} \to [0, 1]$ that maximizes the revenue $R_a(c) = \mathbb{E}[r(X, A) \cdot c(X) \mid A = a]$ in the subgroup $\{A = a\}$ given some welfare-level $W_{u,c}(a) = w$. Then, the bank finds the optimal level w^* of w by maximizing the total revenue $\mathbb{P}(A = 0)R^*_0(w) + \mathbb{P}(A = 1)R^*_1(w)$, whereas in the statement of the proposition, $R^*_a(w)$ denotes $R_a(c^*_{a,w})$.

Thus, the optimal WE classifier $c^*_{\mathrm{WE}(u)}(x, a)$ equals $c^*_{a,w^*}(x)$ provided that the optimization problems for $c^*_{a,w}$ and w^* have a solution. The threshold representation in Equation (5) will follow from a similar representation for $c^*_{a,w}$.

Existence of $c^*_{a,w}$ and w^*, concavity of $R^*_a(w)$:
For a given welfare-level w of a group $\{A = a\}$, the set of feasible classifiers $F(w) = \{c : \mathcal{X} \to [0, 1] \mid W_{u,c}(a) = w\}$ is non-empty if and only if $w \in [0, \mathbb{E}[u \mid A = a]]$. Indeed, for any c we have $W_{u,c}(a) \in [0, \mathbb{E}[u \mid A = a]]$, therefore, the set of feasible marginal classifiers $F(w)$ is empty outside of this interval. For any w inside, the constant classifier $\frac{w}{\mathbb{E}[u|A=a]} \in F(w)$ and thus $F(w)$ is non-empty.

Therefore, for $w \in [0, \mathbb{E}[u_+ \mid A = a]]$, $R^*_a(w) = \sup_{c \in F(w)} R_a(c)$ is finite; outside this interval we assume $R^*_a(w) = -\infty$. Let us show that $R^*_a(w)$ is concave and continuous and that supremum is, in fact, maximum, i.e., that the optimal (marginal) classifier $c^*_{a,w}$ exists.

For any $c' \in F(w')$ and $c'' \in F(w'')$ the convex combination $c = \beta c' + (1 - \beta)c''$ for $\beta \in [0, 1]$ belongs to $F(w)$, where $w = \beta w' + (1 - \beta)w''$. Therefore, $R^*_a(w) \geq \beta R_a(c') + (1 - \beta)R_a(c'')$. Taking supremum over $c' \in F(w')$ and $c'' \in F(w'')$ we obtain $R^*_a(w) \geq \beta R^*_a(w') + (1 - \beta)R^*_a(w'')$. Thus $R^*_a(w)$ is concave in w.

Next, we prove that the maximum is attained. Consider $c^{(n)} \in F(w), n = 1, 2, ...,$ a sequence of classifiers such that $R_a(c^{(n)}) \to R^*_a(w)$. This sequence is a bounded set in the Hilbert space $L^2(\mathcal{X}, \mathbb{P}|_{(X|A=a)})$; thus, by the Banach–Alaoglu theorem (see Section 5 in [14]) contains a weakly-convergent subsequence $c^{(n_k)} \to c \in F(w)$. Since $R_a(c^{n_k})$ is a scalar product of c and $r(\cdot\,, a)$ in $L^2(\mathcal{X}, \mathbb{P}|_{(X|A=a)})$, we get $R_a(c) = \lim_{k\to\infty} R_a(c^{n_k}) = R^*_a(w)$. Thus the maximum is attained.

A similar argument proves continuity of $R^*_a(w)$. By concavity, continuity on $[0, \mathbb{E}[u \mid A = a]]$ follows from upper-semi-continuity: $R^*_a(w) \geq \lim_{n\to\infty} R^*_a(w^{(n)})$ for any sequence $w^{(n)} \to w,\ n \to \infty$. Consider the sequence of optimal classifiers $c^*_{a,w^{(n)}},\ n = 1, 2, \ldots$. There is a weakly convergent subsequence $c^*_{a,w^{(n_k)}}$ with some c as the weak limit. We have $R_a(c) = \lim_{k\to\infty} R^*_a(w^{(n_k)})$ and similarly $W_{u,c}(a) = \lim_{k\to\infty} w^{(n_k)} = w$. Thus $c \in F(w)$ and $R^*_a(w) \geq \lim_{n\to\infty} R^*_a(w^{(n)})$.

The existence of the optimal w^*, where

$$w^* \in \arg\max_{w} \{\mathbb{P}(A = 0)R^*_0(w) + \mathbb{P}(A = 0)R^*_1(w)\}$$

follows from continuity of $R_a^*(w)$ for $w \in [0, \mathbb{E}[u \mid A = a]]$.

Threshold structure of the optimal margianl classifiers c^*_{a,w^*}
By concavity and continuity of $R_a^*(w)$, the super-gradient at w^* is non-empty. Pick an element λ_a from it.

By the definition of super-gradient, for any $w \in \mathbb{R}$ it holds that $R_a^*(w) \leq R_a^*(w^*) + \lambda_a(w - w^*)$. Equivalently, for the optimal classifier $c^*_{a,w^*} \in F(w^*)$ and an arbitrary classifier $c : \mathcal{X} \to [0,1]$ we have

$$R_a(c) - \lambda_a W_{u,c}(a) \leq R_a(c^*_{a,w^*}) - \lambda_a W_{u,c^*_{a,w^*}}(a).$$

In other words, the optimal classifier c^*_{a,w^*} maximizes $R_a(c) - \lambda_a W_{u,c}(a)$ over all $c : \mathcal{X} \to [0,1]$. The converse is also true: any maximizer c gives an optimal classifier c_{a,w^*} provided that it belongs to $F(w^*)$.

Since

$$R_a(c) - \lambda_a W_{u,c}(a) = \mathbb{E}\left[\left(r(X,A) - \lambda_a \overline{u}(X,A)\right) c(X,A) \mid A = a\right],$$

the maximizer equals 1 when $r(x,a) - \lambda_a \overline{u}(x,a) > 0$ and equals 0 if the inequality has the strict opposite sign. The condition $c \in F(w^*)$ imposes the constraint on otherwise arbitrary values of $c(x,a) = \tau_a(x)$ for x with $r(x,a) - \lambda_a \overline{u}(x,a) = 0$. □

7 Worst Cases for Popular Fairness Concepts

In this section we show that Unawareness, DP and EQ can harm the disadvantaged groups. Each subsection is devoted to another fairness concept.

7.1 Unawareness may harm both groups and the bank.

See Section 4.

7.2 DP may harm the disadvantaged protected group

The DP improves upon *Unawareness*. It never hurts both groups but can harm the disadvantaged one as illustrated by the following Example 8. We consider the benchmark case where the welfare is given by the fraction of good borrowers who get loans, so $u(x,a,y) = y$ (i.e., we just ignore bad borrowers).

Example 8. The non-protected attribute is ternary, $\mathcal{X} = \{0,1,2\}$ and all combinations of (x,a) have the same probability of $\frac{1}{6}$. The fraction $p(x,a)$ of good borrowers is given by the following matrix

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

We see that the fraction of good borrowers is higher in the group $\{A=0\}$: $\frac{7}{12}$ against $\frac{1}{3}$. However, for $\{A=1\}$, good borrowers and bad borrowers are perfectly separated by X.

We assume that $\alpha_+ = 2$ and $\alpha_- = 3$, i.e., costs of false-positives are $\frac{3}{2}$ times higher than the benefit from true-positives. The threshold $t = \frac{\alpha_-}{\alpha_+ + \alpha_-}$ thus equals $\frac{3}{5}$.

Under optimal unconstrained classifier c^*_{unc}, all the good borrowers from the second group receive loans (for any t). In $\{A=0\}$, for $t \in \left(\frac{1}{4}, \frac{3}{4}\right)$, loans are given to agents with $x=0$ and $x=1$ only.

As a result, for $t = \frac{3}{5}$, the group $\{A = 0\}$ is disadvantaged: only $\frac{6}{7} = W_{u,c^*_{\text{unc}}}(0)$ of good borrowers there get loans compared to all good borrowers in $\{A = 1\}$.

Next, consider an optimal classifier c^*_{parity} satisfying DP. Since $\alpha_- > \alpha_+$, costs of giving loans to $(x, a) \in \{(1, 1), (2, 1)\}$ cannot be compensated by the revenue derived from allocating the same amount of loans to good borrowers; hence, in the group $\{A = 1\}$, c^*_{parity} gives money to all agents with $x = 0$ while no loans to $\{(1, 1), (2, 1)\}$. In order to satisfy parity constraint in the group $\{A = 0\}$ we have $\sum_x c^*_{\text{parity}}(x, 0) = 1$. This amount of 1 can be arbitrarily divided among the two equally profitable subgroups $x = 0$ and $x = 1$: for example, $c^*_{\text{parity}}(0, 0) = 1$ and $c^*_{\text{parity}}(1, 0) = 0$ (thus no agent from $(1, 0)$ gets money) or $c^*_{\text{parity}}(0, 0) = 1/2$ and $c^*_{\text{parity}}(1, 0) = 1/2$ (agents from $(0, 0)$ and $(1, 0)$ receive loans with probability $1/2$).

We see that under the constraint of DP, loan applications of good borrowers from the disadvantaged group are approved less likely than before imposing the fairness constraint! Agents from the other group, $\{A = 1\}$, feel no impact after imposing the fairness constraint.

7.3 EO helps the disadvantaged group

In a case the utility function is the same as the one given in the previous subsection, i.e., $u(x, a, y) = y$, EO always makes good borrowers from the disadvantaged group weakly better off. This follows from Theorem 6 – because *our choice of a utility function is aligned with the statistical constraint of the fairness concept* (here EO).

However, EO is not robust to other selections of a utility function. In particular, assume that $u(x, a, y) \equiv 1$, $\mathcal{X} = \{0, 1, 2\}$ with the same distribution of good/bad borrowers as in Example 8, and $\frac{5}{8} < t < \frac{3}{4}$. The group $\{A = 1\}$ is disadvantaged: $W_{u,c^*_{\text{unc}}}(0) = \frac{2}{3}$ and $W_{u,c^*_{\text{unc}}}(1) = \frac{1}{3}$.

EO classifier, in order to equalize the number of loans given to good borrowers, must either increase the amount of loans given to $\{A = 0\}$ by approving some applications from $x = 2$ or decrease the number of loans given to $\{A = 1\}$. For $t > \frac{5}{8}$ giving loans to $(x, a) = (2, 0)$ 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 $(0, 1)$. Therefore, the optimal EO classifier coincides with c^*_{unc} in $\{A = 0\}$ and gives less loans to $(0, 1)$: $c^*_{\text{opportunity}}(1, 0) = \frac{6}{7}$. Hence $W_{u,c^*_{\text{opportunity}}}(1) = \frac{2}{7} < W_{u,c^*_{\text{unc}}}(1)$ and the disadvantaged group is harmed by EO.

8 Popular Fairness Concepts

In this subsection give an introductory to the most influential mathematical concepts of fairness from the literature, which aim to prevent discrimination. Importantly, observe that these fairness concepts impose constraints on the policy of the bank without explicitly considering the welfare of the borrowers. We shall revisit these concepts in Section 3 under our utilitarian approach.

Unawareness One of the most intuitive concepts is *Unawareness*: to deliver a "fair" outcome to both groups $A = 0$ and $A = 1$, the classifier c must ignore the protected attribute A. Formally, a classifier c satisfies unawareness if $c(x, 0) = c(x, 1)$ for all $x \in \mathcal{X}$.

This, at first glance, natural idea is not innocuous: X and A can be dependent and thus information contained in X can be used as a proxy for A, driving the disadvantaged group even worse off, see [8]. Despite its flaws, unawareness is promoted by layers in labor market, insurance market[10], and also by GDPR.[11]

The optimal unaware classifier c^*_{unaware} can be constructed by the same logic as the optimal unconstrained, [6]: an applicant receives money if the probability $\mathbb{P}(Y=1 \mid X)$ of being a good borrower given the observed attribute X is above the threshold $t(X) = \frac{\alpha_-(X)}{\alpha_+(X)+\alpha_-(X)}$. Formally, $c^*_{\text{unaware}}(x,a) = c^*_{\text{unaware}}(x) = 1$ for borrowers x such that $\overline{p}(x) = \mathbb{P}(A=0)p(x,0) + \mathbb{P}(A=1)p(x,1) \geq t(x)$ and $c^*_{\text{unaware}}(x) = 0$ if $\overline{p}(x) < t(x)$.

Demographic Parity (DP) The fairness concept of DP, instead of reducing the information available to the decision-maker, imposes constraints on the outcomes of decisions within the two groups: it requires that the fraction of those who receive loans in the two groups must be the same. Formally, a classifier c satisfies DP if

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

Although DP circumvents the "proxy" flaw of *Unawareness*, it is criticized [11]: when the two groups are different in their average creditworthiness, DP forbids the perfect classifier $C(X,A) = Y$ even if such classifier is feasible, i.e., quality Y is a function of attributes X and A.

Equal Opportunity (EO) Motivated by drawbacks of DP, Hardt et al. [11] concluded that good and bad borrowers within protected groups must be treated separately and introduced the concept of EO. Under this fairness concept, the fraction of good borrowers who get loans must be the same in the two subgroups, i.e., DP is applied to smaller subgroups $\{Y=1, A=0\}$ and $\{Y=1, A=1\}$. Formally, a classifier c satisfies EO if

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

In contrast to DP, it allows for perfect classifier $C(X,A) = Y$ if Y is a function of observables, and does not artificially force a bank to give an equal amount of loans if the two groups significantly differ in the fraction of good borrowers [12]; unlike *Unawareness*, it does not lead to information losses and has no "proxy" issues since it constrains the decision, not the data.

[10] By the decision of the Court of Justice of the European Union, insurance premiums must be determined in a gender-blind way starting from December 2012. This initiative eventually increased the gap between premiums paid by females and males [5].

[11] The General Data Protection Regulation is a law adopted by European Union in 2016 that regulates data protection and privacy.

[12] Think of rich and poor neighborhoods. We would like to be fair to their residents, but there is no reason why equalizing the number of loans given is "fair" if the performances of the two subgroups are different. Moreover this might be considered discriminatory to the group with higher performance. However, the condition that for a good borrower, it must be equally easy to get a loan no matter what neighborhood she is from, is pretty natural.