Fairness, equality, and power in algorithmic decisionmaking

Maximilian Kasy
Department of Economics
Oxford University
maximilian.kasy@economics.ox.ac.uk

Rediet Abebe
Society of Fellows
Harvard University
rabebe@fas.harvard.edu

Abstract

Public debate and the computer science literature worry about the fairness of algorithms, understood as the absence of discrimination. We argue that some leading definitions of fairness have three limitations. (1) They legitimize inequalities justified by "merit." (2) They are narrowly bracketed, considering only differences of treatment within the algorithm. (3) They consider only between-group differences.

We compare fairness to two alternative perspectives overcoming these limitations. The first asks what is the causal impact of the introduction of an algorithm on inequality? The second asks who gets to pick the objective function of an algorithm? We formalize these perspectives, characterize when they give divergent evaluations of algorithms, and provide empirical examples.

1 Introduction

Recent years have seen a lively debate on the evaluation of algorithmic decision making systems from a perspective of fairness. There are various definitions of fairness that have been proposed; we will review some of these. These definitions are influenced by legal definitions of discrimination, as well as by the corresponding protections against discrimination. The definitions that we focus on here are based on some variant of the question *are members of different groups who are of equal "merit" treated equally* by the algorithm (or the other way around)?[1] "Merit" here is defined as a measure of whatever promotes the decisionmaker's objective, such as for instance whatever maximizes profits. In the spirit of "reflective equlibrium" [32], we discuss some implications of such definitions of fairness that might be considered normatively undesirable.

We compare fairness-based perspectives to two alternative perspectives. The first of these alternatives asks *what is the causal impact of the introduction of an algorithm on inequality*? That is, what is the impact, relative to some counterfactual, on both within- and between-group inequality? The second of these alternatives asks *who gets to pick the objective function* of an algorithm?

Let us briefly discuss these three perspectives. Leading notions of fairness, such as predictive parity or balance, require prior definitions of merit M, and of protected categories A. Fairness-based perspectives of this form have the following three limitations: (1) They legitimize and perpetuate inequalities justified by "merit" (both within and between groups). (2) Fairness-based perspectives are narrowly bracketed. Fairness only requires equal treatment within the context of the algorithm or decision problem at hand. They do not consider the causal impact of the algorithm in the context of pre-existing inequalities. Unequal treatment by the algorithm might decrease or increase these inequalities, depending on how it correlates with them. (3) Fairness-based perspectives focus on categories (protected groups). This focus ignores within-group inequalities, as emphasized by

[1] Definitions that do not have this general form are notions of "disparate impact," which do not refer to merit, and notions of "individual fairness," which are based on merit but do not refer to group membership.

Preprint. Under review.

intersectional critiques. Equal treatment across groups is consistent with great inequality within groups.

In contrast to fairness, a perspective focused on equality is consequentialist. It depends on the distribution of outcomes Y affected by the algorithm. To talk about the causal impact on inequality of an algorithm, we need to define a counterfactual decision procedure. This counterfactual procedure could be either the status quo, or some alternative algorithm. This perspective encompasses both frameworks based on social welfare functions, and statistical measures of inequality. This perspective overcomes the limitations of a fairness based perspective, as listed above.

The third perspective we discuss is based on the notion of power. Algorithmic fairness is often defined as imposing a normative constraint on some given optimization problem, such as maximizing predictive accuracy or profits. But the objective function of the initial optimization problem is rarely questioned when discussing fairness. A political economy perspective suggests that we should ask who gets to pick this objective function, and why? Who has ownership and control rights over data and algorithms? Why is so much machine learning effort dedicated to maximizing ad clicks, to setting profit maximizing prices, or to enabling mass surveillance by intelligence agencies? We formalize one possible notion of power based on the idea of "inverse welfare weights:" Given actual decisions, what are the welfare weights that rationalize these decisions? Formally, this is the inverse of the social welfare maximization problem that corresponds to our second perspective.

1.1 Related Work

Many now-classic bodies of work study discrimination and harms caused by machine learning systems on historically disadvantaged and marginalized groups in settings ranging from ad delivery [37] to facial analysis [7] and word embedding [6]. [4] provide a framework for understanding the negative consequences of such automated decision-making systems. With a growing set of findings of algorithmic discrimination in the backdrop, researchers across computer science and economics have sought to formalize and define different notions of fairness as well as analyze their feasibility, incompatibility, and their politics. We direct the reader to [9, 13, 22, 36, 30, 39, 27, 3] for an overview and extensive discussions around various definitions of fairness as well as their relationship with other algorithmically-defined desiderata.

This paper also draws on the economics literatures on discrimination, causal inference, social choice, optimal taxation, and on inequality and distributional decompositions. Definitions of fairness correspond to notions of taste-based and statistical discrimination in economics [5], and the notion of fairness defined in Equation (5) correspond to "hit-rate" based tests for taste-based discrimination as in [24]. Causal inference and the potential outcomes framework is reviewed in [18], social choice theory and welfare economics in [33]. Distributional decompositions are discussed by [12]; we draw in particular on the RIF regression approach of [11]. Understanding aggregation in social welfare functions in terms of welfare weights is common in optimal tax theory, cf. [35].

Recent work at the intersection of economics and computation has explored the intersection of fairness with social welfare and inequality. [16, 15] present a welfare-based study of fair classification and study the relationship between fairness definitions and the long-standing notions of social welfare considered in this work. By translating a loss minimization program into a social welfare maximization problem, they show that more strict fairness criteria can lead to worse outcomes for both advantaged and disadvantaged groups. [14] similarly consider fairness and welfare, proposing welfare-based measures of fairness that can be incorporated into a loss minimization program. In a related discussion, [29] considers algorithmic fairness questions within a social welfare framework, comparing policies by machine learning systems with those set by a social planner that cares both about efficiency and equity. Discussing examples where fair algorithms can increase or decrease inequities, [29] argues for a more holistic formalization of fairness notions.

2 Setup and notation

Throughout this paper, we consider the following setting. A decision maker $\mathscr{D}$, such as a firm, a court, or a school, makes repeated decisions for a cross-section of units i, for instance job applicants, defendants, or students. We omit the subscript i where not necessary for clarity. For each unit i, a binary decision W is made (hiring, release from jail, admission). Units are characterized by some

unobserved "merit" $M \in \mathbb{R}$ (marginal productivity, potential crime commission, future educational success). In some settings, M is binary, but we do not make this assumption unless otherwise noted. The decision maker's objective is to maximize

$$\mu = E[W \cdot (M - c)], \tag{1}$$

where the expectation averages over the units i, and c is the unit cost of choosing $W = 1$.[2] In the hiring context, μ corresponds to profits, and c to the wage rate. In the student admissions context, μ corresponds to average student performance, and c might be the Lagrange multiplier (shadow cost) of some capacity constraint.

$\mathscr{D}$ does not observe M, but they do observe covariates (features) X. They can also form a predictive model for M given X based on past data,

$$m(x) = E[M|X = x]. \tag{2}$$

In practice, m needs to be estimated using some machine learning algorithm, such as a Lasso regression, a random forest, or a deep neural net. We will abstract from this estimation issue for now, and assume that $m(\cdot)$ is known to $\mathscr{D}$.

$\mathscr{D}$ can allocate W as a function of X, and possibly some randomization device. We assume throughout that W is chosen independently of all other variables conditional on X, and thus is conditionally exogenous.[3] Denote $w(x) = E[W|X = x]$ the conditional probability of $W = 1$. Given their available information, the optimal assignment policy for $\mathscr{D}$ satisfies

$$w^*(\cdot) = \underset{w(\cdot)\in\mathscr{W}}{\operatorname{argmax}}\ E[E[W \cdot (M - c)|X]] = \underset{w(\cdot)\in\mathscr{W}}{\operatorname{argmax}}\ E[w(X) \cdot (m(X) - c)], \tag{3}$$

where $\mathscr{W}$ is a set of admissible assignment policies.[4] The second equality holds because of conditional exogeneity of W and the law of iterated expectations. If $\mathscr{W}$ is unrestricted, then (up to arbitrary tie breaking for $m(X) = c$)

$$w^*(x) = \mathbf{1}(m(X) > c). \tag{4}$$

To define fairness below, we assume that individuals are additionally characterized by a binary variable A, corresponding to protected groups such as gender or race. A might or might not be part of the features X.

2.1 Fairness – review

Definitions of fairness Numerous definitions of fairness have been proposed in the literature, see for instance [23], [17], [20], and [31]. We will focus on the following definition of fairness, corresponding to the notion of "predictive parity" or calibration:

$$E[M|W = 1, A = a] = E[M|W = 1] \quad \forall\, a. \tag{5}$$

This equality is the basis of tests for preferential discrimination in empirical economics, see for instance [24]. A similar requirement could be imposed for $W = 0$. Another related requirement is "balance for the positive (negative) class," $E[W|M = m, A] = E[W|M = m]$ (equality of false positive/negative rates).

Predictive parity requires that expected merit, conditional on having received treatment 1 (or 0), is the same across the groups A. Balance requires that the probability of being treated, conditional on merit, is the same across the groups A. As noted by [23] for the binary M case, balance and predictive parity can not hold at the same time, unless either prediction is perfect ($M = E[M|X]$), or base rates are equal ($M \perp A|W$). Note that reversely, under our assumptions, if X perfectly predicts M, then the rule $w^*(x) = \mathbf{1}(m(x) > c) = \mathbf{1}(M > c)$ satisfies all of these fairness constraints.

In our subsequent discussion, we will focus on "predictive parity" as the leading measure of fairness, but similar arguments apply to other notions. For binary $A \in \{0, 1\}$, the assignment rule $w(\cdot)$ satisfies predictive parity if and only if $\pi = 0$, where

$$\pi = E[M|W = 1, A = 1] - E[M|W = 1, A = 0] = E\left[M \cdot \left(\tfrac{WA}{E[WA]} - \tfrac{W(1-A)}{E[W(1-A)]}\right)\right]. \tag{6}$$

[2] Formally, we consider a probability space $(\mathscr{I}, P, \mathscr{A})$, where all expectations integrate over $i \in \mathscr{I}$ with respect to the probability measure P, and all random variables are functions on $\mathscr{I}$ that are measurable with respect to $\mathscr{A}$.

[3] This assumption holds by construction, if X captures all individual-specific information available to $\mathscr{D}$.

[4] This type of decision problem, with a focus on estimation in finite samples, has been considered for instance in [21] and [2].

Observation 1 *Suppose that (i) $m(X) = M$ (perfect predictability) and (ii) $w^*(x) = \mathbf{1}(m(X) > c)$ (unconstrained maximization of $\mathscr{D}$'s objective μ). Then $w^*(x)$ satisfies predictive parity, i.e., $\pi = 0$.*

This observation is an immediate consequence of the definition of fairness as predictive parity. This observations points to the limited critical potential of such a definition of fairness. It implies for instance that if surveillance is complete and employers are profit maximizing without constraints, then their hiring decisions will be fair by definition. The algorithm $w(\cdot)$ violates fairness only if (i) $\mathscr{D}$ is not actually maximizing π (taste-based discrimination), (ii) outcomes are mismeasured, leading to biased predictions $m(\cdot)$, or (iii) predictability is imperfect, leading to statistical discrimination.

Fairness as a constraint A leading approach in the recent literature on fairness and machine learning [20] is to consider fairness as a constraint to be imposed on the decision-makers policy space. That is, as before $w^*(\cdot) = \text{argmax}_{w(\cdot)\in\mathscr{W}} E[w(X)\cdot(m(X)-c)]$, but $\mathscr{W}$ is specified to be of the form

$$\mathscr{W} = \{w(\cdot) : \pi = 0\}, \tag{7}$$

for predictive parity (and similarly for other notions of fairness). We characterize the solution to this optimization problem in Corollary 1 below.

3 Inequality and the causal impact of algorithms

We next turn to a perspective focused on notions of social welfare or inequality, and the causal impact of algorithms, drawing on [33], [35], and [19]. Suppose that we are interested in outcomes Y that might be affected by the treatment W, where the outcomes Y are determined by the potential outcome equation

$$Y = W\cdot Y^1 + (1-W)\cdot Y^0, \tag{8}$$

cf. [18]. Suppose that treatment is assigned randomly conditional on X with assignment probability $w(X)$. Then the joint density of X and Y[5] is given by

$$p_{Y,X}(y,x) = \left[p_{Y^0|X}(y,x) + w(x)\cdot\left(p_{Y^1|X}(y,x) - p_{Y^0|X}(y,x)\right)\right]\cdot p_X(x). \tag{9}$$

We are interested in the impact of $w(\cdot)$ on a general statistic ν of the joint distribution of outcomes Y and features X,

$$\nu = \nu(p_{Y,X}). \tag{10}$$

ν might be a measure of inequality, such as the variance of Y or the ratio between two quantiles of Y, a measure of welfare, such as the expectation of Y^γ (where γ parametrizes inequality aversion), or a measure of group-based inequality such as the difference in the conditional expectation of Y given $A = 1$ and $A = 0$.

The influence function and welfare weights In order to characterize the impact of changes to the assignment policy $w(x)$ on the statistic ν it is useful to introduce the following local approximation to ν. Assume that ν is differentiable as a function of the density $p_{Y,X}$.[6] Then, as discussed in chapter 20 of [38], as well as in [10], [11], and [19], we can locally approximate ν by

$$\nu(p_{Y,X}) - \nu(p^*_{Y,X}) \approx E[IF(Y,X)], \tag{11}$$

where $IF(Y,X)$ is the influence function of $\nu(p_{Y,X})$ at $p^*_{Y,X}$, evaluated at the realization Y, X, and the expectation averages over the distribution $p_{Y,X}$.

Suppose now that

$$w(x) = w^*(x) + \epsilon\cdot dw(x), \tag{12}$$

where w^0 is some baseline assignment rule, and $dw(x)$ is a local perturbation to w. Suppose that p and p^* are the outcome distributions corresponding to w and w^*. By Equation (11),

$$\nu(p_{Y,X}) - \nu(p^*_{Y,X}) \approx \int IF(y,x)(p_{Y,X}(y,x) - p^*_{Y,X}(y,x))dydx.$$

[5] The density is assumed to exist with respect to some dominating measure. For simplicity of notation, our expressions are for the case where the dominating measure is the Lebesgue measure, but they immediately generalize to general dominating measures.

[6] To be precise, we need Fréchet-differentability with respect to the L^∞ norm on the space of densities of Y, X with respect to some dominating measure.

By Equations (9) it then follows that

$$\frac{\partial}{\partial \epsilon}\nu(p_{Y,X}) = \int IF(y,x)\cdot\left(p_{Y^1|X}(y,x) - p_{Y^0|X}(y,x)\right)\cdot p_X(x)dx = E\left[dw(X)\cdot n(X)\right]$$
$$\text{where} \quad n(x) = E\left[IF(Y^1,x) - IF(Y^0,x)|X=x\right]. \tag{13}$$

Proposition 1 below formally proves this claim. Defining ω as the average slope of $IF(y,x)$ between Y^0 and Y^1, we can rewrite $IF(Y^1,x) - IF(Y^0,x) = \omega\cdot(Y^1 - Y^0)$. We can think of ω as the "welfare weight" for each person, measuring how much the statistic ν "cares" about increasing the outcome Y for that person. This is analogous to the welfare weights of public economics and optimal tax theory, cf. [34, 35].

Examples of influence functions and welfare weights A few examples help shape our intuition about influence functions and welfare weights. For the mean outcome $\nu = E[Y]$, we get $IF = Y - E[Y]$ and $\omega = 1$. For the variance of outcomes $\nu = \text{Var}(Y)$, we get $IF = (Y-E[Y])^2 - \text{Var}(Y)$ and $\omega \approx 2(Y - E[Y])$. For the mean of some power of the outcome, $\nu = E[Y^\gamma/\gamma]$, we get $IF = Y^\gamma - E[Y^\gamma]$ and $\omega \approx Y^{\gamma-1}$. And lastly, for the between-group difference of average outcomes, $\nu = E[Y|A=1] - E[Y|A=0]$, we have $IF = Y\cdot\left(\frac{A}{E[A]} - \frac{1-A}{1-E[A]}\right)$ and $\omega = \frac{A}{E[A]} - \frac{1-A}{1-E[A]}$.

Utilitarian welfare Thus far, we have discussed welfare in terms of outcomes Y that are observable in principle. This contrasts with the typical approach in welfare economics [26, 8], where welfare is defined based on the unobserved utility of individuals. Unobserved utility can be operationalized in terms of equivalent variation, that is, willingness to pay: What is the amount of money Z that would leave an individual indifferent between receiving Z and no treatment ($W = 0$), or receiving $W = 1$ but no money. Based on this notion of equivalent variation, social welfare can then be defined as

$$\nu = E[(\omega\cdot Z)\cdot W], \tag{14}$$

The welfare weights ω now measure the value assigned to a marginal unit of money for a given person. Welfare weights reflect distributional preferences.

Theoretical characterization of the tension between the decisionmaker's objective, fairness, and equality In the following proposition, we characterize the effect of a marginal change $dw(\cdot)$ of the policy $w(\cdot)$ on the different objectives, the decisionmaker's objective μ, the measure of fairness π, and statistics ν that might measure inequality or social welfare. Conflicts between these three objectives can arise if $l(x)$, $p(X)$, and $n(x)$, as defined below, are not affine transformations of each other.

Proposition 1 (Marginal policy changes) *Consider a family of assignment policies*

$$w(x) = w^*(x) + \epsilon\cdot dw(x),$$

and denote by $d\mu$, $d\pi$ and $d\nu$ the derivatives of μ ($\mathscr{D}$'s objective), π (the measure of fairness), and ν (inequality or social welfare) with respect to ϵ. Suppose that ν is Fréchet-differentable with respect to the L^∞ norm on the space of densities of Y, X with respect to some dominating measure.

Then

$$d\mu = E[dw(X)\cdot l(X)], \qquad d\pi = E\left[dw(X)\cdot p(X)\right], \qquad d\nu = E[dw(X)\cdot n(X)],$$

where

$$l(X) = E[M|X=x] - c, \tag{15}$$
$$p(X) = E\left[(M - E[M|W=1,A=1])\cdot\frac{A}{E[WA]}\right.$$
$$\left. - \quad (M - E[M|W=1,A=0])\cdot\frac{(1-A)}{E[W(1-A)]}\Big|X=x\right], \tag{16}$$
$$n(x) = E\left[IF(Y^1,x) - IF(Y^0,x)|X=x\right]. \tag{17}$$

Let us now reconsider the problem of maximizing μ subject to the fairness constraint $\pi = 0$. The solution to this problem is characterized in Corollary 1, drawing on Proposition 1.

Corollary 1 (Optimal policy under the fairness constraint) *The solution to the problem of maximizing $\mathscr{D}$'s objective μ subject to the fairness constraint $\pi = 0$ by choice of $w(\cdot)$ is given by*

$$w(x) = \mathbf{1}(l(x) > \lambda p(x)), \tag{18}$$

for some constant λ, where we have chosen $w(x)$ arbitrarily for values of x such that $l(x) = \lambda p(x)$, and the equality holds with probability 1.

4 Power

Thus far, we have discussed fairness and equality as two ways to normatively evaluate algorithms. Fairness is a frame to critique the unequal treatment of individuals i who are of the same merit M, where merit is defined in terms of $\mathscr{D}$'s objective. The equality frame takes a broader perspective. It asks us to consider the causal impact of an algorithm on the distribution of relevant outcomes Y across individuals i more generally.

There is an "elephant in the room" on the other side of the algorithm, though, which neither of these perspectives addresses. They both consider inequality between the different individuals i who are treated, but they don't consider who gets to be the decision maker $\mathscr{D}$ – who gets to pick the objective function μ?

We would like to suggest a political economy perspective on algorithmic decision making as a third approach, in addition to fairness and equality. Political economy, since [25], is concerned with the ownership of the means of production. Ownership of the means of production brings both income and control rights. In the setting of algorithmic decision making, this maps into two related questions. First, who owns and controls data, in particular data X about individuals? And second, who gets to pick the algorithms $\mathscr{W}$ and objective functions μ that use these data? Furthermore, what are the consequences of this structure of ownership and control? The answers to these questions depend on contingent historical developments and political choices, rather than natural necessity, as forcefully argued by [28] and [41], for instance.

Implied welfare weights as a measure of power In the present paper, we propose the following step toward a field that studies the political economy of artificial intelligence and algorithmic decision making. We propose to study actual decision procedures $w(\cdot)$ by considering the welfare weights ω that would rationalize these procedures as optimal. Put differently, we ask to consider the dual problem of the problem of finding the optimal policy for a given measure of social welfare.

Above, we discussed the effect of marginal policy changes on statistics ν that might measure welfare. We argued that this effect can be written as $E[dw(X) \cdot E[\omega \cdot (Y^1 - Y^0)|X]]$, where ω are "welfare weights," measuring how much we care about a marginal increase of Y for a given individual. The optimal policy problem of maximizing a linear (or linearized) objective $\nu = E[\omega \cdot Y]$ net of the costs of treatment $E[c \cdot W]$ defines a mapping

$$(\omega_i)_i \to w^*(\cdot) = \underset{w(\cdot)\in\mathscr{W}}{\operatorname{argmax}}\ E\left[\left(\omega \cdot (Y^1 - Y^0) - c\right) \cdot w(X)\right]. \tag{19}$$

We are now interested in the inverse mapping $w^*(\cdot) \to (\omega_i)_i$. This mapping gives the welfare weights ω which would rationalize a given assignment algorithm $w(\cdot)$ as optimal. These welfare weights can be thought of as measures of the effective social power of different individuals. Corollary 2 below characterizes this inverse mapping in the context of our binary treatment assignment setting.

In the following Corollary of Proposition 1, we characterize the implied welfare weights ω that would rationalize a given policy $w(\cdot)$.

Corollary 2 (Implied welfare weights) *Suppose that welfare weights are a function of the observable features X, and that there is again a cost of treatment c. A given assignment rule $w(\cdot)$ is a solution to the problem*

$$\underset{w(\cdot)}{\operatorname{argmax}}\ E[w(X) \cdot (\omega(X) \cdot E[Y^1 - Y^0|X] - c)] \tag{20}$$

if and only if

$$\begin{aligned} w(x) = 1 &\Rightarrow \omega(X) > c/E[Y^1 - Y^0|X]) \\ w(x) = 0 &\Rightarrow \omega(X) < c/E[Y^1 - Y^0|X]) \\ w(x) \in]0,1[&\Rightarrow \omega(X) = c/E[Y^1 - Y^0|X]). \end{aligned} \tag{21}$$

5 Examples for the three tensions between fairness and equality

In the introduction we noted three limitations of notions of fairness (in the sense of predictive parity / calibration or balance for the positive / negative class): (1) Fairness-based perspectives legitimize and perpetuate inequalities justified by "merit" (both within and between groups). (2) Fairness-based perspectives are narrowly bracketed, since fairness only requires equal treatment within the context of the algorithm, rather than equal outcomes across a wider population. (3) Fairness-based perspectives focus on categories (protected groups), and ignore within-group inequalities. We next illustrate each of these three limitations, by providing examples where some change to the assignment algorithm $w(\cdot)$ decreases un-fairness, while at the same time also increasing inequality and decreasing welfare.

In each of the examples, we consider the impact of an assignment rule $w^{(ii)}$, relative to some baseline rule $w^{(i)}$, on fairness, inequality and welfare. We contrast fairness as measured by "predictive parity" to inequality (welfare) as measured by either the variance of Y, or the average of Y^γ, where $\gamma < 1$ measures the degree of inequality aversion.

Legitimizing inequality based on merit In our first example, we consider an **improvement in** the **predictability** of merit. Suppose that initially (under scenario a), the decisionmaker $\mathscr{D}$ only observes A, while under scenario b they can perfectly predict (observe) M based on X. Assume that $Y = W$. Recall that c denotes the cost of treatment, and assume that M is binary with $P(M = 1|A = a) = p^a$, where $0 < c < p^1 < p^0$. Under these assumptions we get

$$W^{(i)} = \mathbf{1}(E[M|A] > c) = 1, \qquad W^{(ii)} = \mathbf{1}(E[M|X] > c) = M.$$

The policy a is unfair (in the sense of predictive parity), since for this policy $E[M|W^{(i)} = 1, A = 1] = p^1 < p^0 = E[M|W^{(i)} = 1, A = 0]$, while the policy b is fair, since $E[M|W^{(ii)} = 1, A = 1] = 1 = E[M|W^{(ii)} = 1, A = 0]$. The increase in predictability has thus improved fairness.

Inequality of outcomes, however, has has also increased. By assumption $Y = W$, so that $\text{Var}_{(i)}(Y) = 0$, $\text{Var}_{(ii)}(Y) = E[M](1 - E[M]) > 0$. Furthermore, expected welfare $E[Y^\gamma]$ has decreased, since $E_{(i)}[Y^\gamma] = 1$, $E_{(ii)}[Y^\gamma] = E[M] < 1$.

Narrow bracketing In our second example, we consider a reform that **abolishes affirmative action**. Suppose that (M, A) is uniformly distributed on $\{0, 1\}^2$, that M is perfectly observable to the decision maker $\mathscr{D}$, and that $0 < c < 1$. Suppose that initially (under scenario a) the decision maker receives a reward (subsidy) of 1 for hiring members of the group $A = 1$. Suppose that this reward is abolished under scenario b. Under these assumptions we get

$$W^{(i)} = \mathbf{1}(M + A \geq 1), \qquad W^{(ii)} = M.$$

As before, the policy under scenario a is unfair, while the policy under scenario b is fair, since $E[M|W^{(i)} = 1, A = 1] = .5 < 1 = E[M|W^{(i)} = 1, A = 0]$, while $E[M|W^{(ii)} = 1, A = 1] = 1 = E[M|W^{(ii)} = 1, A = 0]$.

Suppose now that potential outcomes are given by $Y^w = (1 - A) + w$. Under the two scenarios, the outcome distributions are $Y^{W^{(i)}} = 1 + \mathbf{1}(A = 0, M = 1) \sim Cat(0, 3/4, 1/4)$, and $Y^{W^{(ii)}} = (1 - A) + M \sim Cat(1/4, 1/2, 1/4)$, where we use Cat to denote the categorical distribution on $\{0, \ldots, 2\}$ with probabilities specified in brackets. This implies $\text{Var}_{(i)}(Y) = 3/16$, $\text{Var}_{(ii)}(Y) = 1/2$, and $E_{(i)}[Y^\gamma] = .75 + .25 \cdot 2^\gamma$, $E_{(ii)}[Y^\gamma] = .5 + .25 \cdot 2^\gamma$. As before, the inequality of outcomes increases and welfare declines as we move from scenario a to scenario b.

Within-group inequality In our third example we consider a reform that **mandates fairness** to the decision maker. Suppose that $P(A = 1) = .5$, $c = .7$, and $M|A = 1 \sim Unif(\{0, 1, 2, 3\})$,

$M|A = 0 \sim Unif(\{1, 2\})$, Suppose that initially $\mathscr{D}$ is unconstrained, but the reform mandates predictive parity, $E[M|W^{(ii)} = 1, A = 1] = E[M|W^{(ii)} = 1, A = 0]$. Then

$$W^{(i)} = \mathbf{1}(M \geq 1), \qquad W^{(ii)} = \mathbf{1}(M + A \geq 2).$$

Once again, the policy under scenario a is unfair, while the policy under scenario b is fair, since $E[M|W^{(i)} = 1, A = 1] = 2 > 1.5 = E[M|W^{(i)} = 1, A = 0]$, while $E[M|W^{(ii)} = 1, A = 1] = 2 = E[M|W^{(ii)} = 1, A = 0]$.

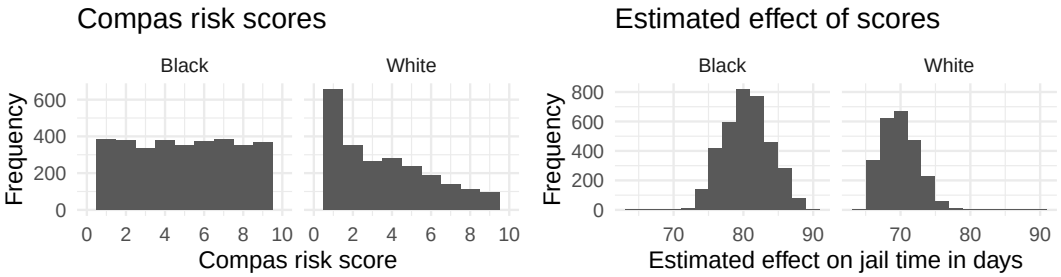
Suppose now that potential outcomes are given by $Y^w = M + w$. Under the two scenarios, the outcome distributions are $Y^{W^{(i)}} = M + \mathbf{1}(M \geq 1) \sim Cat(1/8, 0, 3/8, 3/8, 1/8)$, $Y^{W^{(ii)}} = M + \mathbf{1}(M + A \geq 2) \sim Cat(1/8, 2/8, 1/8, 3/8, 1/8)$, where we use Cat to denote the categorical distribution on $\{0, \ldots, 4\}$ with probabilities specified in brackets. This implies $\text{Var}_{(i)}(Y) = 1.24$, $\text{Var}_{(ii)}(Y) = 1.61$, and, choosing $\gamma = .5$, $E_{(i)}[Y^\gamma] = 1.43$, $E_{(ii)}[Y^\gamma] = 1.33$. Again, the inequality of outcomes increases and welfare declines as we move from scenario a to scenario b.

6 Case study

We next present an illustration of our arguments using the Compas risk score data for recidivism. These data have received much attention following Pro-Publica's reporting on algorithmic discrimination in sentencing [1]. We map our setup to the Compas data as follows. A denotes race (Black or White), W denotes a risk score exceeding 4 (as in Pro-Publica's analysis, based on the Compas classification as medium or high risk), M denotes recidivism within two years, and Y denotes jail time. The predictive features X that we consider include race, sex, age, juvenile counts of misdemeanors, fellonies, and other infractions, general prior counts, as well as charge degree.

We compare three counterfactual scenarios. (1) A counterfactual "affirmative action" scenario, where race-specific adjustments are applied to the risk scores. We decrease the scores generated by Compas by one unit for Black defendants, and increase them one unit for Whites. (2) The status-quo scenario, taking the original Compas scores as given. (3) A counterfactual "perfect predictability" scenario, where scores are set to 10 (the maximum value) for those who actually recidivated within 2 years. Scores are set to 1 (the minimum value) for all others.

For each of these scenarios, we impute corresponding values of W (i.e., a counterfactual score bigger than 4), and counterfactual jail time Y. The latter is calculated based on a causal-forest estimate [40] of the impact on Y of risk scores, conditional on the covariates in X. This relies on the (strong) assumption of conditional exogeneity of risk-scores given X.



As can be seen in Table 1, fairness as measured by predictive parity improves when moving from the affirmative action scenario to the status-quo, and is fully achieved in the perfect predictability scenario. This follows because the difference in expected recidivism, conditional on having a score bigger than 4, between blacks and whites decreases as we go from one scenario to the next.

On the other hand, both Table 1 and Table 2 show that inequality both between and within racial groups increases as we go from one scenario to the next. The difference in mean jail time between blacks and whites increases from about 12 days to about 23 days. The interquartile range in the distribution of counterfactual jail time increases from about 24 days to 60 days. And the standard deviation of log jail time increases from 1.8 to 2.1.

Acknowledgements

We thank Daniel Privitera for helpful feedback and comments.

Table 1: Counterfactual scenarios, by group

	Black			White		
Scenario	(Score>4)	Recidl(Score>4)	Jail time	(Score>4)	Recidl(Score>4)	Jail time
Aff. Action	0.49	0.67	49.12	0.47	0.55	36.90
Status quo	0.59	0.64	52.97	0.35	0.60	29.47
Perfect predict.	0.52	1.00	65.86	0.40	1.00	42.85

Table 2: Counterfactual scenarios, outcomes for all

Scenario	Score>4	Jail time	IQR jail time	SD log jail time
Aff. Action	0.48	44.23	23.8	1.81
Status quo	0.49	43.56	25.0	1.89
Perfect predict.	0.48	56.65	59.9	2.10

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Propublica*, May 2016.

[2] Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

[3] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *NIPS Tutorial*, 2017.

[4] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

[5] Gary S Becker. *The economics of discrimination*. 1957.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[8] Raj Chetty. Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics*, 1(1):451–488, 2009.

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

[10] F.A. Cowell and M.P. Victoria-Feser. Robustness properties of inequality measures. *Econometrica: Journal of the Econometric Society*, pages 77–101, 1996.

[11] S. Firpo, N. Fortin, and T. Lemieux. Unconditional quantile regressions. *Econometrica*, 77:953–973, 2009.

[12] S. Firpo, N. Fortin, and T. Lemieux. Decomposition methods in economics. *Handbook of Labor Economics*, 4:1–102, 2011.

[13] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[14] Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas 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.

[15] Lily Hu and Yiling Chen. Welfare and distributional impacts of fair classification. *arXiv preprint arXiv:1807.01134*, 2018.

[16] Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 535–545, 2020.

[17] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019.

[18] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[19] Maximilian Kasy. Partial identification, distributional preferences, and the welfare ranking of policies. *Review of Economics and Statistics*, 2015.

[20] Michael Kearns and Aaron Roth. *The Ethical Algorithm: The Science of Socially Aware Algorithm Design*. Oxford University Press, 2019.

[21] Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

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

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

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

[25] K. Marx. *Das Kapital: Kritik der politischen Ökonomie*, volume 1. 1867.

[26] Andreu Mas-Colell, Michael Dennis Whinston, and Jerry R. Green. *Microeconomic theory*. Oxford University Press, 1995.

[27] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[28] Evgeny Morozov. Socialize the data centers! *New Left Review*, 91, 2015.

[29] Sendhil Mullainathan. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 1–1, 2018.

[30] Arvind Narayanan. fairness definitions and their politics. In *Tutorial presented at the Conference on Fairness, Accountability, and Transparency*, 2018.

[31] Dana Pessach and Erez Shmueli. Algorithmic fairness. *arXiv preprint arXiv:2001.09784*, 2020.

[32] John Rawls. *A theory of justice*. Harvard University Press, Cambridge, 1973.

[33] John E Roemer. *Theories of distributive justice*. Harvard University Press, Cambridge, 1998.

[34] Emmanuel Saez. Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229, 2001.

[35] Emmanuel Saez and Stefanie Stantcheva. Generalized social welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45, 2016.

[36] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.

[37] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.

[38] Aad W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

[39] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[40] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[41] Shoshana Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books, 2019.

A Proofs

Proof of Proposition 1:
The case of μ is immediate from the definition of μ. The case of ν follows from the definition of Fréchet differentiability (cf. Section 20.2 in **(author?)** 38), Lemma 1 in [19], and the arguments in Section 3 of this paper. This leaves the case of π. Let us consider the first component of π,

$$E[M|W=1,A=1] = E\left[\frac{WMA}{E[WA]}\right],$$

and thus

$$\begin{aligned} dE[M|W=1,A=1] &= E\left[dw(X)\cdot\left(\frac{MA}{E[WA]} - \frac{E[WMA]}{E[WA]^2}\cdot A\right)\right] \\ &= E\left[dw(X)\cdot(M - E[M|W=1,A=1])\cdot\frac{A}{E[WA]}\right] \end{aligned}$$

The derivative of $E[M|W=1,A=0]$ can be calculated similarly, and the claim follows. □

Proof of Corollary 1:
We are looking for a solution to

$$\begin{aligned} &\max_{w(\cdot)} \mu = \int (m(x)-c)p_X(x)dx && \text{subject to} \\ &\pi = E\left[\frac{MWA}{E[WA]} - \frac{MW(1-A)}{E[W(1-A)]}\right] = 0 && \text{and} \\ &0 \le w(x) \le 1 \quad \forall\, x. \end{aligned}$$

The Lagrangian for the objective and the fairness constraint is given by $\mathscr{L} = \mu + \lambda\pi$. Consider a family of policies indexed by ϵ, $w(x) = w^*(x) + \epsilon\cdot dw(x)$, as in Proposition 1. The solution to our optimization problem has to satisfy the condition

$$\frac{\partial\mathscr{L}}{\epsilon} \le 0$$

for all feasible changes dw, that is, for all dw such that

$$\begin{aligned} w^*(x) = 1 &\Rightarrow dw(x) \le 0 \\ w^*(x) = 0 &\Rightarrow dw(x) \ge 0. \end{aligned}$$

By Proposition 1,

$$\frac{\partial\mathscr{L}}{\partial\epsilon} = \int dw(x)\,(l(x) + \lambda p(x))\,p_X(x)dx.$$

Suppose there is some set of values x of non-zero probability such that $w^*(x) < 1$ and $l(x) + \lambda p(x) > 0$. Setting $dw(x) = 1$ on this set would yield a contradiction. The claim follows. □

Proof of Corollary 2:
This follows immediately from the Karush–Kuhn–Tucker conditions for the constrained optimization problem defining $w^*(\cdot)$. □

B Balance for the positive class

In Section 2.1, we introduced predictive parity as a definition of fairness. In Proposition 1 we then characterized the impact of marginal policy changes on the measure π of predictive parity. An

alternative, related, notion of fairness is balance for the positive class, which requires that $\widetilde{\pi} = 0$, where

$$\begin{aligned}\widetilde{\pi} &= E[W|M=1, A=1] - E[W|M=1, A=0] \\ &= E\left[W \cdot \left(\frac{MA}{E[MA]} - \frac{M(1-A)}{E[M(1-A)]}\right)\right]. \end{aligned} \tag{22}$$

In analogy to Observation 1, the following is immediate.

Observation 2 *Suppose that (i) $m(X) = M$ (perfect predictability) and (ii) $w^*(x) = \mathbf{1}(m(X) > c)$ (unconstrained maximization of $\mathscr{D}$'s objective μ). Then $w^*(x)$ satisfies balance for the positive class, i.e., $\widetilde{\pi} = 0$.*

As in Proposition 1, we can also characterize the impact of marginal policy changes on $\widetilde{\pi}$ as $d\pi = E\left[dw(X) \cdot \widetilde{p}(X)\right]$, where

$$\begin{aligned}\widetilde{p}(x) &= E\left[\left(\frac{MA}{E[MA]} - \frac{M(1-A)}{E[M(1-A)]}\right)\Big| X = x\right] \\ &= \left(\frac{E[MA|X=x]}{E[MA]} - \frac{E[M(1-A)|X=x]}{E[M(1-A)]}\right). \end{aligned} \tag{23}$$

The proof is immediate.