Fairness, Equality, and Power in Algorithmic Decision-Making

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 contrast fairness to two alternative perspectives, informed by the theory of justice and empirical economics, that do not share some of 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

There is a rich line of work in computer science concerned with the differential treatment by algorithms of historically disadvantaged and marginalized groups. Much of this work is concerned with fairness of algorithms, understood as the absence of discrimination. Many of the leading notions of fairness – such as predictive parity or balance – are based on some variant of the question *are members of different groups who are of equal "merit" treated equally* by the algorithm?"[1] Research in this space has ranged from translating these fairness notions to various domains to examining when and whether they are simultaneously achievable with other constraints.

In this work, in the spirit of "reflective equilibrium" [30], we discuss implications of such definitions of fairness that might be considered normatively undesirable. These notions of fairness take the objective of the algorithm's owner as a normative goal. In the context of hiring, for instance, if productivity is perfectly predictable and an employer's hiring algorithm is profit maximizing without constraints, then their hiring decisions are fair by definition; only deviations from profit maximization are considered discriminatory. Furthermore, we argue that leading notions of fairness, such as predictive parity or balance, suffer from the following three limitations:

1. Fairness-based perspectives legitimize and perpetuate inequalities justified by "merit" both within and between groups. The focus on "merit" – a measure promoting the decision-maker's objective – reinforces, rather than questions, the legitimacy of the status quo.
2. They are narrowly bracketed. Fairness only requires equal treatment within the context of the algorithm at hand, and does not consider the impact of the algorithm on inequality in the wider population. Unequal treatment that compensates pre-existing inequalities might reduce overall inequality.

[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.

3. They focus on categories (protected groups) and ignore within-group inequalities, e.g., as emphasized by intersectional critiques. Equal treatment across groups can be consistent with great inequality within groups.

Informed by insights from the theory of justice and empirical economics, we discuss each of these limitations in the context of algorithmic decision-making. We then compare this fairness-based perspective with two alternative perspectives. The first alternative perspective asks *what is the causal impact of the introduction of an algorithm on inequality*, both within and between groups? In contrast to fairness, this perspective is consequentialist. It depends on the distribution of *outcomes* affected by the algorithm rather than treatment, and it does so for the full population rather than only for individuals who are part of the algorithm. This perspective encompasses both frameworks based on social welfare functions and statistical measures of inequality. In Proposition 1, one of our main results, we provide a formal characterization of the impact of marginal policy changes on both fairness and inequality using influence functions, which allows us to elucidate the conflict between these two objectives.

The second alternative perspective focuses on the distribution of power and asks *who gets to pick the objective function* of an algorithm? The choice of objective functions is intimately connected with the political economy question of who has ownership and control rights over data and algorithms. To explore this question, 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? This is formalized in Corollary 2, which builds on Proposition 1, solving the inverse of a social welfare maximization problem.

The rest of this paper is structured as follows. Our setup is introduced in Section 2. We formalize the perspective based on the causal impact of algorithms in Section 3 and that based on distribution of power in Section 4. In doing so, we highlight limitations of a fairness-based perspective, which we expose further in Section 5 through examples. We close with a discussion. In the appendix, we discuss an empirical application of these insights in Section A, provide proofs in Section B, and provide results for alternative notions of fairness in Section C

1.1 Related Work

Many now-classic bodies of work study discrimination and harms caused by machine learning systems on historically disadvantaged groups in settings ranging from ad delivery [35] 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 numerous fields 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, 21, 34, 28, 37, 25, 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 [22]. Causal inference and the potential outcomes framework is reviewed in [18], social choice theory and welfare economics in [31]. 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. [33].

Recent work in economics and computation has sought to understand the intersection of fairness with social welfare and inequality. [15, 16] 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, [27] 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, [27] argues for a more holistic formalization of fairness notions.

2 Setup and Notation

A decision-maker $\mathscr{D}$, such as a firm, a court, or a school, makes repeated decisions on individuals i, who may be job applicants, defendants, or students. We omit the subscript i, when it is not necessary for clarity. For each individual i, a binary decision W – such as hiring, release from jail, college admission – is made. Individuals are characterized by some unobserved "merit" $M \in \mathbb{R}$, such as marginal productivity, recidivism, 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 individuals 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 college 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 has access to some 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 supervised machine learning algorithm. 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 $m(X) = c$

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

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

Fairness Definitions Numerous definitions of fairness have been proposed in the literature, see for instance [21], [17], and [29]. 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, [22]. A similar requirement could be imposed for $W = 0$. Another related requirement is "balance for the positive (or negative) class," $E[W|M = m, A] = E[W|M = m]$, which indicates equality of false positive (respectively 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 [9, 13, 21] for the binary M case, balance and

[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 [20] and [2].

predictive parity cannot hold at the same time, unless either prediction is perfect ($M = E[M|X]$), or base rates are equal ($M \perp A|W$). In our subsequent discussion, we focus on "predictive parity" as the leading measure of fairness; we provide parallel results for balance in Appendix C.

For $A \in \{0, 1\}$, the assignment rule $w(\cdot)$ satisfies predictive parity if and only if $\pi = 0$, where

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

Fairness as a constraint A leading approach in the recent literature is to consider fairness as a constraint to be imposed on the decision-maker's policy space. That is, $w^*(\cdot)$ is defined as above, but $\mathscr{W}$ is specified to be of the form

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

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

In the introduction we argued that fairness takes the objective of the algorithm's owner as a normative goal. This is formalized by the following observation.

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 and points to the limited critical potential of such a definition of fairness. It implies, for instance, that if M is perfectly predictable given the available features 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.

This observation throws the three limitations of a fairness-based perspective into sharp relief: under this perspective, inequality both between and within groups is acceptable if it is justified by merit M ($\mathscr{D}$'s objective), no matter where the inequality in M is coming from. Furthermore, given merit, fairness aims for equal treatment within the algorithm, rather than aiming for compensating pre-existing inequalities of welfare-relevant outcomes in the wider population. And, finally, predictive parity or balance do not consider inequality of treatments (or outcomes) within the protected groups, but rather only between them. Below, we provide examples where changes to an assignment algorithm $w(\cdot)$ decreases un-fairness, while at the same time also increasing inequality and decreasing welfare.

3 Inequality and the Causal Impact of Algorithms

In this section, drawing on theories of justice, we turn to a perspective focused on social welfare and inequality as well as the causal impact of algorithms [31, 33, 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

$$Y = W \cdot Y^1 + (1 - W) \cdot Y^0, \quad (8)$$

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

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

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

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

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 $A = 0$).

[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.

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 [36], 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)

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

By Equations (9), it then follows that

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

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 used in public economics and optimal tax theory, cf. [32, 33]. We present examples to give intuition of welfare weights and influence functions.

Examples 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 [8, 24], 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 $\nu = E[(\omega \cdot Z) \cdot W]$. The welfare weights ω now measure the value assigned to a marginal unit of money for a given person. Welfare weights reflect distributional preferences.

Tension between the decision-maker'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 decision-maker'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.

[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.

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

$$l(X) = E[M|X = x] - c, \tag{14}$$

$$p(X) = E\left[(M - E[M|W = 1, A = 1]) \cdot \frac{A}{E[WA]} - \quad (M - E[M|W = 1, A = 0]) \cdot \frac{(1-A)}{E[W(1-A)]}\Big| X = x\right], \tag{15}$$

$$n(x) = E\left[IF(Y^1, x) - IF(Y^0, x)|X = x\right]. \tag{16}$$

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*

$$w(x) = \mathbf{1}(l(x) > \lambda p(x)), \tag{17}$$

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 Distribution of Power

As discussed above, fairness provides a framework to critique the unequal treatment of individuals i with the same merit, where merit defined in terms of $\mathscr{D}$'s objective. The equality framework takes a broader perspective by requiring that we consider the causal impact of an algorithm on the distribution of relevant outcomes Y across individuals i more generally. Both of these perspectives, however, fail to address another key component: *who gets to set the objective function and why? i.e., who gets to be $\mathscr{D}$ and pick the objective function μ?*

Here, we take a political economy perspective on algorithmic decision-making to provide a framework for examining this question. Since [23], political economy is concerned with the ownership of the means of production, as this 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, and in particular data X about individuals? And second, who gets to pick the algorithms $\mathscr{W}$ and objective functions μ that use this data? We are further concerned with what the consequences of this structure of ownership and control are. The answers to these questions depend on contingent historical developments and political choices, rather than natural necessity, as argued by [26] and [39], for instance.

Implied welfare weights as a measure of power In the present paper, we propose the following framework that studies the political economy of artificial intelligence and algorithmic decision-making: we study actual decision procedures $w(\cdot)$ by considering the welfare weights ω that would rationalize these procedures as optimal. Put differently, we 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

$$(\omega_i)_i \to w^*(\cdot) = \underset{w(\cdot)\in\mathscr{W}}{\operatorname{argmax}}\ E\left[\big(\omega \cdot (Y^1 - Y^0) - c\big) \cdot w(X)\right]. \tag{18}$$

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. The following corollary of Proposition 1 characterizes this inverse mapping in the context of our binary treatment setting. 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*

$$\underset{w(\cdot)}{\operatorname{argmax}} \; E[w(X) \cdot (\omega(X) \cdot E[Y^1 - Y^0|X] - c)] \tag{19}$$

if and only if

$$\begin{aligned} w(x) = 1 &\Rightarrow \omega(X) > c/E[Y^1 - Y^0|X]) \\ w(x) = 0 &\Rightarrow \omega(X) < c/E[Y^1 - Y^0|X]) \\ w(x) \in]0,1[&\Rightarrow \omega(X) = c/E[Y^1 - Y^0|X]). \end{aligned} \tag{20}$$

5 Examples for the Tensions between Fairness and Equality

We return to the limitations of a fairness-based perspective formulated at the outset. We 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 $w^{(i)}$. We contrast fairness as measured by "predictive parity" to inequality (and 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 We consider an improvement in the predictability of merit. Suppose that initially (under scenario a), the decision-maker $\mathscr{D}$ 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.

On the other hand, inequality of outcomes has also increased and welfare has decreased. 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 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 further that under scenario a the decision-maker receives a reward (subsidy) of 1 for hiring members of the group $A = 1$, but that this reward is removed 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

$$Y^{W^{(i)}} = 1 + \mathbf{1}(A = 0, M = 1) \sim Cat(0, 3/4, 1/4), \text{and}$$

$$Y^{W^{(ii)}} = (1 - A) + M \sim Cat(1/4, 1/2, 1/4),$$

where we use Cat to denote the categorical distribution on $\{0, \ldots, 2\}$ with probabilities specified in brackets. This implies that $\text{Var}_{(i)}(Y) = 3/16$, $\text{Var}_{(ii)}(Y) = 1/2$, and $E_{(i)}[Y^\gamma] = .75 + .25 \cdot 2^\gamma$, $E_{(ii)}[Y^\gamma] = .5 + .25 \cdot 2^\gamma$. Thus, as before, the inequality of outcomes has increased and welfare has decreased when we move from scenario a to scenario b.

Within-group inequality We finally consider a reform that mandates fairness to the decision-maker. Suppose that $P(A = 1) = .5$, $c = .7$, and further that $M|A = 1 \sim Unif(\{0, 1, 2, 3\})$, $M|A = 0 \sim Unif(\{1, 2\})$, We assume 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

$$E[M|W^{(i)} = 1, A = 1] = 2 > 1.5 = E[M|W^{(i)} = 1, A = 0], \text{ and}$$
$$E[M|W^{(ii)} = 1, A = 1] = 2 = E[M|W^{(ii)} = 1, A = 0].$$

Assume 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 Conclusion

In this work, we articulate and discuss three limitations of fairness-based perspectives under leading notions of fairness: namely, that they legitimize inequalities justified by merit, rather than questioning the carceral state or capitalist labor markets; they are narrowly bracketed and do not engage with the impact of algorithms on pre-existing inequalities; and they do not consider within-group inequalities, leading to intersectional concerns.

To help alleviate these limitations, we consider two alternative perspectives drawing on theories of justice, empirical economics, and optimal tax theory. These perspectives focused on inequality and distribution of power are not intended to solve the above concerns but rather to bring to the surface frameworks that have been largely neglected within the algorithmic fairness literature. In doing so, we hope to broaden the scope of discussions around the impact of algorithms.

An inequality-centered perspective is pertinent in settings where we presume that inequalities of social outcomes are socially created, and the same holds for various forms of "merit" (marginal productivity, recidivism, etc.). Here, we can see any decision system as just another step on the causal pathway of reproducing or reducing these inequalities. An approach intending to minimize harm on disadvantaged groups therefore does better to consider the effect of any particular decision system (whether algorithmic or human) on inequality as a whole, rather than aiming to solely optimize for fairness within the algorithm.

7 Statement of Broader Impact

Acknowledgements

We thank Stefano Caria, Zöe Hitzig, Joshua Loftus, Daniel Privitera, Ana-Andreea Stoica, Sam Taggart, Bryan Wilder, and Angela Zhou for helpful feedback and comments.

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

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

[22] 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.

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

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

[25] 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.

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

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

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

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

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

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

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

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

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

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

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

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

[38] 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.

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

A Empirical Study

We illustrate 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, felonies, and other infractions, general prior counts, as well as charge degree.

We compare three counter-factual scenarios. (1) A counter-factual "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 white defendants. (2) The status-quo scenario, taking the original Compas scores as given. (3) A counter-factual "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 counter-factual score bigger than 4), and counter-factual jail time Y. The latter is calculated based on a causal-forest estimate [38] 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.

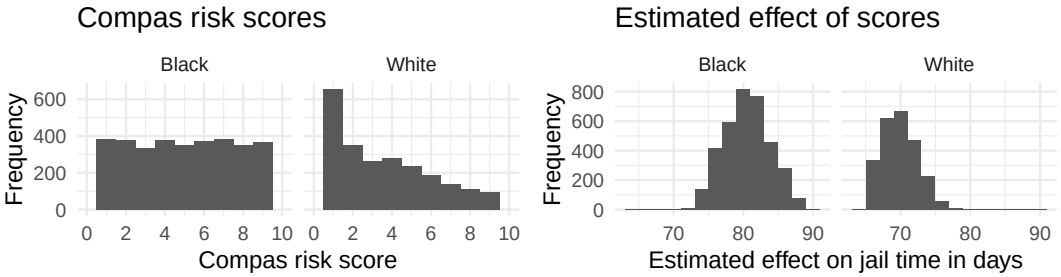

Table 1: Counter-factual scenarios, by group

	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: Counter-factual 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

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 Black and white defendants 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 Black and white defendants increases from about 12 days to about 23 days. The interquartile range in the distribution of counter-factual 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.

B 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 [36]), 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 $w^*(\cdot)$. □

C Balance for the positive class

We introduced predictive parity as a definition of fairness above. 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

$$\begin{aligned} \widetilde{\pi} &= E[W|M=1,A=1] - E[W|M=1,A=0] \\ &= E\left[W\cdot\left(\frac{MA}{E[MA]} - \frac{M(1-A)}{E[M(1-A)]}\right)\right]. \end{aligned} \tag{21}$$

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

$$\begin{aligned}\widetilde{p}(x) &= E\left[\left(\frac{MA}{E[MA]} - \frac{M(1-A)}{E[M(1-A)]}\right)\Big| X = x\right] \\ &= \left(\frac{E[MA|X=x]}{E[MA]} - \frac{E[M(1-A)|X=x]}{E[M(1-A)]}\right). \end{aligned} \tag{22}$$

The proof is immediate.