Fairness, Equality, and Power in Algorithmic Decision-Making

Maximilian Kasy
University of Oxford
Department of Economics

Rediet Abebe
University of California, Berkeley
Department of Electrical Engineering & Computer Sciences

ABSTRACT

Much of the debate on the impact of algorithms is concerned with fairness, defined as the absence of discrimination for individuals with the same "merit." Drawing on the theory of justice, we argue that leading notions of fairness suffer from three key limitations: they legitimize inequalities justified by "merit;" they are narrowly bracketed, considering only differences of treatment within the algorithm; and they consider between-group and not within-group differences. We contrast this fairness-based perspective with two alternate perspectives: the first focuses on inequality and the causal impact of algorithms and the second on the distribution of power. We formalize these perspectives drawing on techniques from causal inference and empirical economics, and characterize when they give divergent evaluations. We present theoretical results and empirical examples which demonstrate this tension. We further use these insights to present a guide for algorithmic auditing and discuss the importance of inequality- and power-centered frameworks in algorithmic decision-making.

CCS CONCEPTS

• **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**; • **Social and professional topics** → *Computing / technology policy*.

KEYWORDS

Algorithmic fairness, inequality, power, auditing, empirical economics

ACM Reference Format:
Maximilian Kasy and Rediet Abebe. 2021. Fairness, Equality, and Power in Algorithmic Decision-Making. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages.

1 INTRODUCTION

A rich line of work within computer science examines the differential treatment by algorithms of historically disadvantaged and marginalized groups. Much of this work is concerned with fairness of algorithms, which is understood as the absence of discrimination. Many 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 the spirit of "reflective equilibrium" [55], in this work we discuss implications of these fairness definitions that may be deemed normatively undesirable. Leading notions of fairness take the objective of the algorithm's owner or designer 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 these leading notions of fairness, such as predictive parity or balance, suffer from the following three limitations.

(1) They 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.
(3) They focus on categories (protected groups) and ignore within-group inequalities, e.g., as emphasized by intersectional critiques [14]. Equal treatment across groups can be consistent with great inequality within groups.

Informed by insights from theories of justice and empirical economics, we discuss each of these limitations. We then compare this fairness-based perspective with two alternative perspectives. The first 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 Section 3, 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.

[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.

FAccT '21, March 3–10, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8309-7/21/03…$15.00
DOI: 10.1145/3442188.3445919

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."[2] Given actual decisions, what are the welfare weights that rationalize these decisions? We formalize this in Section 4, which builds on insights from Section 3 by 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. In Section 6, we present an empirical application of these insights. In Section 7, we present a step-by-step guide for algorithmic auditing, estimating the causal impact of algorithmic changes on measures of inequality or welfare. We close with a discussion on the importance of inequality- and power-based frameworks in algorithmic decision-making.

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 [61] to facial analysis [10] to search engine bias [53] and provision of public services [16]. Barocas and Selbst provide a framework for understanding the negative consequences of such automated decision-making systems [5]. For general overviews and discussions, see also [1, 7, 9, 20, 53, 54]. 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 politics. We direct the reader to [12, 19, 22, 38, 45, 50, 60, 64] for an overview and extensive discussions around various definitions of fairness as well as their relationship with other algorithmically-defined desiderata.

Our work draws on the economics literature 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 [6], and the notion of fairness defined in Equation (5) correspond to "hit-rate" based tests for taste-based discrimination as in [39]. Causal inference and the potential outcomes framework is reviewed in [32], social choice theory and welfare economics in [56]. Distributional decompositions are discussed in [18]; we draw in particular on the RIF regression approach of [17]. Understanding aggregation in social welfare functions in terms of welfare weights is common in optimal tax theory [58]. For a sociological perspective on discrimination, we direct the reader to an overview in [59].

Recent work has considered short-comings of fairness across a number of dimensions. For instance, there is a growing body of work examining the long-term impact of fairness-driven interventions [15, 28, 29, 41, 68]. Similarly, there is also a surge of work focused on understanding and improving fairness across subgroups [24, 35, 36] as well as in settings where group membership may not be known [21, 23, 33, 66]. Other work examines perceived trade-offs between fairness and other desiderata, such as accuracy [67].

The closest work to ours have sought to understand the intersection of fairness with social welfare and inequality [25, 26, 30, 48, 49]. Despite tackling a different set of questions than ours, there are several papers that consider welfare-based analyses of fairness notions. In a notable example, Hu and Chen 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 [30]. 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. Heidari et al. similarly consider fairness and welfare, proposing welfare-based measures that can be incorporated into a loss minimization program, and Heidari et al. connect fairness to notions of equality of opportunity [25, 26]. In a related discussion, Mullainathan 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 [49]. These lines of line of work argue for more holistic assessments of welfare and equity in examining the impact of algorithmic decision-making.

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. For clarity, we omit the subscript i when there is no ambiguity. 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, or future educational success. In some settings, M is binary, but we do not make this assumption unless otherwise noted. In this work, merit M refers to the variable that the decision-maker cares about; for instance a worker's productivity in the hiring context, or recidivism in the bail setting context. This is the variable that supervised learning methods typically aim to predict. By contrast, the outcome Y – which we introduce in Section 3 below – is the variable that the treated individuals care about; for instance income in the hiring context, or time spent in jail in the bail setting context.[3]

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 $W = 1$.[4] Upper-case letters denote random variables, lower-case letters values that these random variables might take. In the hiring context, μ corresponds to profits and c to the wage rate. In the college admissions context, μ corresponds

[2] Note, influence functions and welfare weights are commonly used in economics and statistics, but are less common in computer science. We present a self-contained introduction in the appendix as they are key tools in our analyses.

[3] Our terminology here deviates from familiar usage in that we use "*outcome*" to refer to Y rather than M. Note here, there are two possible "outcomes" – one for the individual concerned and one for the decision-maker – and this usage is intended to avoid ambiguity between them.

[4] 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 $\mathscr{A}$.

to average student performance among admitted students, and c might be the Lagrange multiplier (shadow cost) of some capacity constraint.

The decision-maker $\mathscr{D}$ does not observe M, but has access to some covariates (features) X. $\mathscr{D}$ 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 here 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.[5] 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,

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

where $\mathscr{W}$ is a set of admissible assignment policies.[6] 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 $m(X) = c$,

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

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

Fairness Definitions. Numerous definitions of fairness have been proposed in the literature [31, 50]. We will focus on the following popular 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 [39].) 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. For the binary M case, balance and predictive parity cannot hold at the same time, unless either prediction is perfect ($M = E[M|X]$), or base rates are equal ($M \perp A|W$) [12, 22, 38] . In our subsequent discussion, we focus on "predictive parity" as the leading measure of fairness; we provide parallel results for balance in Appendix A.

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

[6]This type of decision problem, with a focus on estimation in finite samples, has been considered for instance in [37] and [3].

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

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

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,

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

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

We argued in the introduction 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), (ii) $w^*(x) = 1(m(X) > c)$ (unconstrained maximization of $\mathscr{D}$'s objective μ), and (iii) $W, M \in \{0, 1\}$ (classification setting). 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. Similar observations could be stated for other notions of fairness, such as "balance for the positive class" (cf. Observation 2 in the appendix), and for other settings, such as regression, where $M, W \in \mathbb{R}$.

Absent perfect predictability, there may be a tension between the maximization of μ and predictive parity ($\pi = 0$). This tension has been discussed in economics as a failure of predictive parity (called "hit rate test" in this literature, [39]) to perfectly reflect taste-based discrimination [6]. This failure is due to the difference between average and marginal expected merit among the treated. Taste-based discrimination corresponds to differences between the merit of the marginally treated, while predictive parity corresponds to equality of average merit among the treated. profit-maximization is equivalent to the absence of taste-based discrimination, by definition.

Observation 1 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

Drawing on theories of justice, we turn to a perspective focused on social welfare and inequality as well as the causal impact of algorithms [34, 56, 58]. 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. [32]. 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[7]

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

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 $p_{Y,X}$.[8] Then, as discussed [63], as well as in [13], [17], and [34], we can locally approximate ν by

$$\nu(p_{Y,X}) - \nu(p^*_{Y,X}) = E[IF(Y,X)] + o\left(\|p_{Y,X} - p^*_{Y,X}\|\right), \tag{11}$$

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 $p_{Y,X}$.

For completeness, in Section B in the appendix, we provide an introduction and review, as well as a more formal definition of the influence function as dual representation of the Fréchet derivative of ν.

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:

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

Proposition 1 below 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. [57, 58]. We present examples to give intuition of welfare weights and influence functions.

Example 3.1. 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 $IF = Y \cdot \left(\frac{A}{E[A]} - \frac{1-A}{1-E[A]}\right) - \nu$ 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 [11, 44], 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.

Then

$$\begin{aligned} d\mu &= E[dw(X) \cdot l(X)], \\ d\pi &= E\left[dw(X) \cdot p(X)\right], \\ d\nu &= E[dw(X) \cdot n(X)], \end{aligned}$$

[7] 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.
[8] 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.

where

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

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

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

Proof. 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 [63]), Lemma 1 in [34], 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. □

Proposition 1 has a number of important consequences: First, it provides the basis for analyzing the distributional impact of algorithms or changes to algorithms, as part of an algorithmic auditing process, along the lines we demonstrated in our empirical application in Section 6. We provide a step-by-step guide for such an algorithmic auditing procedure in Section 7. Suppose that in our data the treatment W is plausibly exogenous given the features X. Then, $n(x)$ can be estimated by regressing the influence function $IF(Y, X)$ (for some statistic ν) on W, controlling for X, using for instance a causal forest or some other supervised learning method. The impact of switching from assignment algorithm $w^*(x)$ to some other algorithm $w(x)$ is then (to first order) given by the average of $dw(X)$ times $n(X)$, over the distribution of features X. This allows us to estimate the impact of the change of the algorithm on inequality, welfare, or between group differences.

Second, Proposition 1 helps us elucidate the tension between conflicting objectives such as profits, fairness, and equality or welfare, and to connect these to the notion of welfare weights. Suppose, for instance, that π is negative and that for some feature value x we have that $n(x)$ (for some measure of welfare ν) is positive, while $p(x)$ is negative. This tells us that increasing the treatment probability $w(x)$ at x is good for welfare and bad for fairness. We can thus understand which parts of the feature space drive the tension between alternative objectives.

Third, Proposition 1 allows us to characterize the *optimal assignment* $x \to w(x)$ from the decision-maker's point of view, when constrained to fair allocations; this is done in Corollary 1, below.

Fourth, Proposition 1 allows us to understand *whose welfare a status-quo decision procedure implicitly values*, by deriving inverse welfare weights; this is done in Corollary 2 below. This insight is again relevant for the practice of algorithmic auditing. For a related approach, see for instance [8].

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*

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

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.

Proof. We are looking for a solution to

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

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

$$w^*(x) = 1 \Rightarrow dw(x) \le 0$$
$$w^*(x) = 0 \Rightarrow dw(x) \ge 0.$$

By Proposition 1,

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

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 then follows. □

4 DISTRIBUTION OF POWER

Fairness provides a framework to critique the unequal treatment of individuals i with the same merit, where merit is 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, do not address another key component: *who gets to set the objective function and why?*

Here, we take a political economy perspective on algorithmic decision-making to provide a framework for examining this question. Political economy is concerned with the ownership of the means of production, as this brings both income and control rights [43]. 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 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 [47, 69].

Implied welfare weights as a measure of power. In the present work, we propose the following framework for the political economy of 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 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

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

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 one measure 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*

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

if and only if

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

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

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

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

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 $\mathrm{Var}_{(i)}(Y) = 0$, $\mathrm{Var}_{(ii)}(Y) = E[M](1 - E[M]) > 0$. 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

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

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

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

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

where we use Cat to denote the categorical distribution on $\{0, \ldots, 2\}$ with probabilities specified in brackets. This implies that $\mathrm{Var}_{(i)}(Y) = 3/16$, $\mathrm{Var}_{(ii)}(Y) = 1/2$, 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

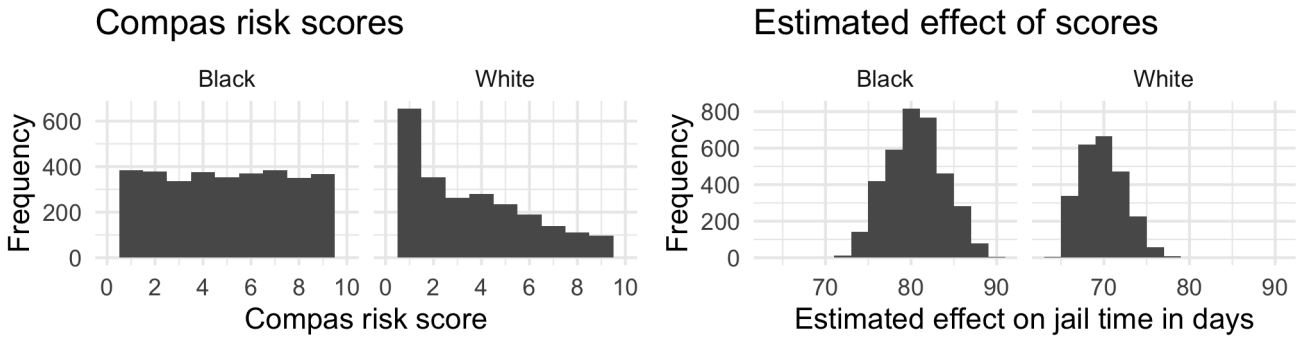

Figure 1: Distribution of Compas risk scores for Black and white defendants (left) and distribution of estimated average causal effects given observable features of the risk score on jail time, across defendants (right). Estimation of causal effects was based on a causal forest; see text for details.

Table 1: Counter-factual scenarios

	Black			White			All			
Scenario	(Score>4)	Recid|(Score>4)	Jail time	(Score>4)	Recid|(Score>4)	Jail time	Score>4	Mean JT	IQR JT	SD of log JT
Affirmative Action	0.49	0.67	49.12	0.47	0.55	36.90	0.48	44.23	23.8	1.81
Status quo	0.59	0.64	52.97	0.35	0.60	29.47	0.49	43.56	25.0	1.89
Perfect predictability	0.52	1.00	65.86	0.40	1.00	42.85	0.48	56.65	59.9	2.10

Table 2: Comparison of the consequences of two counterfactual scenarios (affirmative action, perfect predictability) to the status quo for Black, white, and all defendants.

the reform mandates predictive parity, $E[M|W^{(ii)} = 1, A = 1] = E[M|W^{(ii)} = 1, A = 0]$. Then

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

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

$$Y^{W^{(i)}} = M + 1(M \geq 1) \sim Cat(1/8, 0, 3/8, 3/8, 1/8),$$

$$Y^{W^{(ii)}} = M + 1(M + A \geq 2) \sim Cat(1/8, 2/8, 1/8, 3/8, 1/8),$$

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 EMPIRICAL STUDY

We illustrate our arguments using the Compas risk score data for recidivism. These data have received much attention following ProPublica's reporting on algorithmic discrimination in sentencing [2]. 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 ProPublica'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 [65] 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 Black and white defendants decreases as we go from one scenario to the next.

On the other hand, Table 1 also shows 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.

7 A GUIDE FOR ALGORITHMIC AUDITING USING DISTRIBUTIONAL DECOMPOSITIONS

In this section, we provide a step-by-step guide to algorithmic auditing for distributional impacts. The method we discuss here builds on our characterization of distributional impacts in Proposition 1. This method enables an auditor to estimate the causal impact of switching from treatment assignment $w^*(\cdot)$ (the baseline) to a counterfactual treatment assignment $w(\cdot)$ on some statistic ν of the distribution of outcomes Y. The approach is based on the framework introduced in Section 3.

Step 0: Normative choices. Before any analysis can begin, a number of important normative choices have to be made. First, we need to determine the *relevant outcomes* Y for individuals' welfare – income, education, jail time, and so on. Second, we need to decide on the relevant *measures of welfare or inequality* ν. A number of choices were discussed in the paper. One option, allowing to represent the entire distribution, is to report the impact on a series of quantiles; e.g. all percentiles. Third, the *population of interest* needs to be determined: Do we care about inequality only in our sample, or the population in the state, or the population in the entire country? In many cases, the population of interest will be large relative to the sample treated by the algorithm.

Step 1: Calculation of influence functions. The next step is to calculate the influence function for the measures of interest, at the appropriate baseline distribution of the population of interest. This influence function is then evaluated at each of the observed outcomes in our sample, and stored in a new variable. For example, for the variance of outcomes $\nu = \text{Var}(Y)$, we impute $IF(y_i) = (y_i - E[Y])^2 - \text{Var}(Y)$ for each observation in the sample, where $E[Y]$ and $\text{Var}(Y)$ are evaluated for the population of interest. For the between-group difference of average outcomes, $\nu = E[Y|A = 1] - E[Y|A = 0]$, we impute $IF(y_i, a_i) = y_i \cdot \left(\frac{a_i}{E[A]} - \frac{1-a_i}{1-E[A]}\right) - \nu$, where again $E[A]$ and ν are evaluated for the population of interest. See [13] for further examples.

Step 2: Causal effect estimation. The next step in the proposed analysis is to estimate the conditional average treatment effect of W on $IF(Y)$ given the observed features X. That is, to estimate

$$n(x) = E\left[IF(Y^1, x) - IF(Y^0, x)|X = x\right], \tag{21}$$

in the notation of Proposition 1.

Such causal estimation requires random variation of the treatment W conditional on X. i.e., conditional statistical independence

$$W \perp (Y^0, Y^1)|X. \tag{22}$$

Conditional independence is ensured in experimental settings (e.g., data coming from A/B tests or (contextual) bandits). More generally, independence might be a reasonable approximation if the space of features X is sufficiently rich, and X does not include any variables that are causally "downstream" from W or Y.

Many estimators are available to estimate $n(x)$ under the assumption of conditional independence; in our application we have used the causal forest approach of [65]. After estimating the function $n(\cdot)$, an estimated value of $n(x_i)$ is imputed for every observation i in the sample.

Step 3: Counterfactual assignment probabilities. In order to evaluate the impact of an algorithm $w(\cdot)$, we need to compare it to a baseline algorithm with assignment probabilities $w^*(\cdot)$. In Step 3, we need to evaluate these assignment probabilities $w(x_i)$ and $w^*(x_i)$ for all i, and impute

$$\Delta w(x_i) = w(x_i) - w^*(x_i) \tag{23}$$

for all i in the sample.

Step 4: Evaluation of distributional impact. The last step of our analysis then requires putting the pieces together, and evaluating

$$\hat{\Delta\nu} = \alpha \cdot \frac{1}{n} \sum_i \Delta w(x_i) \cdot n(x_i), \tag{24}$$

where α is the share of the population of interest that is assigned treatment by the algorithm. $\hat{\Delta\nu}$ is the estimated impact of switching from algorithm $w^*(\cdot)$ to algorithm $w(\cdot)$ on the measure ν of inequality or welfare.

8 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 status quo; that they are narrowly bracketed and do not adequately engage with the impact of algorithms on pre-existing inequalities; and that they do not consider within-group inequalities, leading to intersectional concerns [4, 14, 27, 40, 42, 46, 51, 52]. To help alleviate these limitations, we consider two alternative perspectives drawing on theories of justice and empirical economics.

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, any decision system can be viewed as a step in 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 a fixed fairness notion within the algorithm. The latter also risks normatively privileging between group equality to within group equality (cf. for instance Black feminist critiques of second wave feminism [14, 27, 42]).

A perspective focused on the distribution of power compels us to consider the design of the algorithms themselves: Don't just ask how the algorithm treats different people differently, but also who gets to do the treating. By taking a political economy perspective, we examine what implicit distribution of social power justifies

the current choice of objectives. Such a question foregrounds how power gets allocated, and what is the process that leads some groups to have more control over data in decision making processes.

These alternative perspectives focused on inequality and power are not intended to entirely solve the above fairness concerns, but rather to elucidate them and bring to the forefront concerns that haven't been adequately considered in the literature thus far. In doing so, we add to a recent line of work aiming to broaden discussions on the social impact of algorithmic decision-making.

ACKNOWLEDGEMENTS

We thank Stefano Caria, Zöe Hitzig, Jon Kleinberg, Joshua Loftus, Daniel Privitera, Ana-Andreea Stoica, Sam Taggart, Bryan Wilder, Angela Zhou, and other members of the MD4SG Working Group on Inequality for helpful feedback and comments.

A 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

$$\begin{aligned}\widetilde{\pi} &= E[W|M=1,A=1] - E[W|M=1,A=0] \\ &= E\left[W \cdot \left(\frac{MA}{E[MA]} - \frac{M(1-A)}{E[M(1-A)]}\right)\right]. \end{aligned} \tag{25}$$

In analogy to Observation 1, the following is immediate.

Observation 2. *Suppose that (i) $m(X) = M$ (perfect predictability) and (ii) $w^*(x) = 1(m(X) > c)$ (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

$$\begin{aligned}\widetilde{p}(x) &= E\left[\left(\frac{MA}{E[MA]} - \frac{M(1-A)}{E[M(1-A)]}\right)\middle| X = x\right] \\ &= \left(\frac{E[MA|X=x]}{E[MA]} - \frac{E[M(1-A)|X=x]}{E[M(1-A)]}\right). \end{aligned} \tag{26}$$

The proof is immediate.

B AN INTRODUCTION TO INFLUENCE FUNCTIONS AND WELFARE WEIGHTS

In this section we provide a self-contained introduction to and review of influence functions and welfare weights. Influence functions are useful in various subfields of statistics. Welfare weights are useful, in particular, in public finance and optimal tax theory. Both welfare weights and influence functions correspond to definitions of derivatives, that is, local linear approximations to (non-linear) functionals of interest.

Let P denote a probability measure corresponding to the distribution of some real valued random variable Y, possibly jointly with additional features X. We are interested in a statistic ν of P, where ν might for instance correspond to some measure of inequality or of social welfare. Our goal in the following is to provide a local linear approximation to ν in a vicinity of some baseline distribution P^*.

Discrete finite distributions. We first consider the case of discrete distributions with finite support, where the definition of influence functions and welfare weights is elementary. Let δ_y denote the measure with unit mass at the point y, and assume that

$$P = \sum_{i=1}^{k} p_i \delta_{y_i}. \tag{27}$$

Denote $\boldsymbol{p} = (p_1, \ldots, p_k)$ as well as $\boldsymbol{y} = (y_1, \ldots, y_k)$. Then we can write, with a slight abuse of notation

$$\nu(P) = \nu(\boldsymbol{p}, \boldsymbol{y}). \tag{28}$$

We assume that ν is a differentiable function of all of its arguments.

To give an example motivating the following definitions, suppose hat P describes a distribution of income Y across different types of people i. Then the welfare weights ω_i measure how much ν would change if the income of people of type i was marginally increased, while the influence function IF_i measures how much ν would change if the share of people of type i was marginally increased.

Formally, the welfare weights ω_i are given by the derivative of ν with respect to the vector of outcome values $\boldsymbol{y}$, evaluated at the baseline distribution P^*,

$$\omega_i = \frac{\partial}{\partial y_i} \nu(\boldsymbol{p}^*, \boldsymbol{y})\Big|_{\boldsymbol{y}=\boldsymbol{y}^*}. \tag{29}$$

The influence function IF_i is given by the derivative of ν with respect to the vector of probabilities $\boldsymbol{p}$, evaluated at the baseline distribution P^*. A minor complication here is that the statistic ν is, in general, only defined for probability vectors $\boldsymbol{p}$, which satisfy $\sum_i p_i = 1$. The influence function at y_i is therefore defined as the derivative with respect to ϵ of ν evaluated at $P^* + \epsilon \cdot (\boldsymbol{\delta}_{y_i} - P^*)$, that is,

$$IF_i = \frac{\partial}{\partial \epsilon} \nu(\boldsymbol{p}^* + \epsilon \cdot (e_i - \boldsymbol{p}^*);\ \boldsymbol{y})\Big|_{\epsilon=0}, \tag{30}$$

where e_i is the ith unit vector.

By the definition of differentiability of ν, we can now locally approximate ν in the following two ways:

$$\nu(\boldsymbol{p}^*;\ \boldsymbol{y}) = \nu(\boldsymbol{p}^*;\ \boldsymbol{y}^*) + \sum_i \omega_i \cdot (y_i - y_i^*) + o(\|\boldsymbol{y} - \boldsymbol{y}^*\|), \tag{31}$$

$$\nu(\boldsymbol{p};\ \boldsymbol{y}^*) = \nu(\boldsymbol{p}^*;\ \boldsymbol{y}^*) + \sum_i IF_i \cdot (p_i - p_i^*) + o(\|\boldsymbol{p} - \boldsymbol{p}^*\|). \tag{32}$$

Influence functions for general distributions. We have introduced influence functions and welfare weights for discrete distributions. These concepts extend to functionals ν of general probability measures P that describe some joint distribution of Y and X. In particular, suppose that ν is *Fréchet differentiable* at P^*, in the sense that

$$\lim_{P \to P^*} \frac{\|(\nu(P) - \nu(P^*)) - D\nu(P - P^*)\|}{\|P - P^*\|} = 0, \tag{33}$$

where $D\nu$ is a continuous linear functional with respect to the L^2 norm on the space of densities which is defined by $\|P\| = \sqrt{\int (dP/dP^*)^2 dP^*}$. The limit has to equal 0 for all sequences of probability measures P converging to P^*.

Recall that L^2 is a Hilbert space, i.e., a vector space equipped with an inner product $\langle ., . \rangle$. By the *Riesz representation theorem*, we know that for any continuous linear functional $D : L^2 \to \mathbb{R}$ there exists an element $IF \in L^2$, such that

$$D(\boldsymbol{p}) = \langle IF, \boldsymbol{p} \rangle = E_{\boldsymbol{p}}[IF] \tag{34}$$

for all $\boldsymbol{p} \in L^2$. The vector IF is the dual representation of the linear functional D.[9]

Combining equations (33) and (34), and noting that by construction $E_{P^*}[IF] = 0$, we see that, again

$$\nu(P) = \nu(P^*) + E_P[IF] + o(\|P - P^*\|). \tag{35}$$

Uses of influence functions. Influence functions play a role in a number of different contexts in econometrics and statistics. In asymptotic distribution theory for estimators, and the derivation of efficiency bounds, one can show that *any* "regular" estimator $\hat{\nu}$ of the just-identified parameter $\nu(P)$ is asymptotically equivalent to a linearized plug-in estimator, $\hat{\nu} \approx \nu(P^*) + E_n[IF(X)]$. This implies the asymptotic efficiency bound $n\text{Var}(\hat{\nu}) \to \text{Var}(IF(X))$. See for instance [62], in particular chapter 3, and [63], chapter 20.

In robust statistics, since estimators can be approximated by $\hat{\nu} \approx \nu(P^*) + E_n[IF(X)]$, we get that the value of $\hat{\nu}$ can be dominated by a single outlier, even in large samples, *unless IF* is bounded. See for instance [**?**].

In the econometrics of partial identification, nonparametric models with endogeneity tend to lead to representations of potential outcome distributions of the form $P(Y^d) = \alpha P^1(Y) + (1-\alpha)P^2(Y^d)$, where draws from $P^1(Y)$ are observable, while the data are uninformative about $P^2(Y^d)$. A linear approximation to $\nu(P(Y^d))$ then implies $\nu(P(Y^d)) - \nu(P^*) \approx \alpha D\nu(P^1(Y) - P^*) + (1-\alpha)D\nu(P^2(Y^d) - P^*)$. The first term here is identified, the second term can be bounded if and only if $D\nu$ is bounded on the admissible counterfactual distributions $P^2(Y^d)$ - the same condition as in robust statistics. See for instance [**?**], section 3.

Lastly, the literature on distributional decompositions, considers the context closest to the present paper. In labor economics, we are often interested in counterfactual distributions of the form $P(Y) = \int P^1(Y|X)dP^2(X)$, where we observe samples from the distributions 1 and 2. In order to estimate $\nu(P)$, we can again use the approximation $\nu(P) \approx \nu(P^2) + \int IF(Y)dP(Y) = \nu(P^2) + \int IF(Y)dP^1(Y|X)dP^2(X) = \nu(P^2) + E^2[E^1[IF(Y)|X]]$. The conditional expectation $E^1[IF(Y)|X]$ can be estimated using regression methods, the expectation w.r.t. X can be estimated using the P^2 sample average of predicted values from the regression. See for instance [17].

Uses of welfare weights. Welfare weights are a commonly used tool in public finance and the theory of optimal taxation; see for instance [57], [11], and [58]. The typical problem of optimal taxation is to find a maximizer of

$$\nu = \sum_i \omega_i \cdot u_i, \tag{36}$$

where u_i is a money-metric measure of individual utility ("equivalent variation"), and ω_i is the social value assigned to a marginal unit of money for individual i. As noted above, this functional form for ν can be thought of as a local linear approximation to a more general differentiable social welfare function ν.

Typically, ω_i is smaller for those with higher income, reflecting a preference for redistribution. The problem of maximizing ν is subject to a number of constraints, in particular informational constraints, which depend on causal parameters (behavioral responses) that have to be determined empirically. A key insight that simplifies this optimization problem is the envelope theorem, which leverages the assumptions that individuals maximize their own welfare.

[9] This argument could be generalized further by considering weaker notions of differentiability, such as Gâteaux differentiability, and more general L^p spaces, which have dual space L^q where $1/p + 1/q = 1$.

REFERENCES

[1] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. Roles for computing in social change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 252–260, 2020.

[2] 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.

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

[4] Ada Uzoamaka Azodo. Issues in African feminism: A syllabus. *Women's Studies quarterly*, 25(3/4):201–207, 1997.

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

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

[7] Ruha Benjamin. *Race after technology: Abolitionist tools for the new jim code*. John Wiley & Sons, 2019.

[8] Daniel Björkegren, Joshua Blumenstock, and Samsun Knight. (machine) learning what policymakers value. *Working Paper*.

[9] Meredith Broussard. *Artificial unintelligence: How computers misunderstand the world*. MIT Press, 2018.

[10] 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.

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

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

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

[14] Kimberle Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, 43:1241, 1990.

[15] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 525–534, 2020.

[16] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

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

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

[19] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

[20] Timnit Gebru. Oxford handbook on AI ethics book chapter on race and gender. *arXiv preprint arXiv:1908.06165*, 2019.

[21] Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.

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

[23] Tatsunori B Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*, 2018.

[24] Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

[25] 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.

[26] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of

opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 181–190, 2019.
[27] bell hooks. Yearning: Race, gender, and cultural politics. 1992.
[28] Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pages 1389–1398, 2018.
[29] Lily Hu and Yiling Chen. Welfare and distributional impacts of fair classification. *arXiv preprint arXiv:1807.01134*, 2018.
[30] 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.
[31] 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.
[32] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
[33] Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *arXiv preprint arXiv:1906.00285*, 2019.
[34] Maximilian Kasy. Partial identification, distributional preferences, and the welfare ranking of policies. *Review of Economics and Statistics*, 2015.
[35] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572, 2018.
[36] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.
[37] Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.
[38] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
[39] 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.
[40] Mary Modupe Kolawole. Transcending incongruities: Rethinking feminism and the dynamics of identity in Africa. *Agenda*, 17(54):92–98, 2002.
[41] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.
[42] Audre Lorde. Age, race, class, and sex: Women redefining difference. *Women in Culture: An intersectional anthology for gender and women's studies*, pages 16–22, 1980.
[43] K. Marx. *Das Kapital: Kritik der politischen Ökonomie*, volume 1. 1867.
[44] Andreu Mas-Colell, Michael Dennis Whinston, and Jerry R. Green. *Microeconomic theory*. Oxford University Press, 1995.
[45] 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.
[46] Ellis P Monk Jr. The cost of color: Skin color, discrimination, and health among african-americans. *American Journal of Sociology*, 121(2):396–444, 2015.
[47] Evgeny Morozov. Socialize the data centers! *New Left Review*, 91, 2015.
[48] Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 359–368, 2019.
[49] Sendhil Mullainathan. Algorithmic fairness and the social welfare function. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pages 1–1, 2018.
[50] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 2018.
[51] Naomi Nkealah. (West) African feminisms and their challenges. *Journal of literary Studies*, 32(2):61–74, 2016.
[52] Naomi N Nkealah. Conceptualizing feminism(s) in Africa: The challenges facing African women writers and critics. *The English Academy Review*, 23(1):133–141, 2006.
[53] Safiya Umoja Noble. *Algorithms of oppression: How search engines reinforce racism*. nyu Press, 2018.
[54] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.
[55] John Rawls. *A theory of justice*. Harvard University Press, Cambridge, 1973.
[56] John E Roemer. *Theories of distributive justice*. Harvard University Press, Cambridge, 1998.
[57] Emmanuel Saez. Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229, 2001.
[58] Emmanuel Saez and Stefanie Stantcheva. Generalized social welfare weights for optimal tax theory. *American Economic Review*, 106(1):24–45, 2016.
[59] Mario L Small and Devah Pager. Sociological perspectives on racial discrimination. *Journal of Economic Perspectives*, 34(2):49–67, 2020.
[60] Harini Suresh and John V Guttag. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002*, 2019.
[61] Latanya Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10–29, 2013.
[62] A.A. Tsiatis. *Semiparametric theory and missing data*. Springer Verlag, 2006.
[63] Aad W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.
[64] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.
[65] 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.
[66] Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020.
[67] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems*, pages 8783–8792, 2019.
[68] Xueru Zhang, Mohammad Mahdi Khalili, and Mingyan Liu. Long-term impacts of fair machine learning. *Ergonomics in Design*, 28(3):7–11, 2020.
[69] Shoshana Zuboff. *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books, 2019.