Oxford Review of Economic Policy, 2024, **40**, 530–546
https://doi.org/10.1093/oxrep/grae031
Article

OXFORD

Algorithmic bias and racial inequality: a critical review

Maximilian Kasy*

*Department of Economics, University of Oxford. maximilian.kasy@economics.ox.ac.uk.

I thank Johanna Barop for her research assistance, and Sandy Darity, Frank DiTraglia, Pirmin Fessler, Sukjin Han, Isabel Ruiz Olaya, and Alex Teytelboym, as well as an anonymous reviewer, for valuable feedback. This work was supported by the Alfred P. Sloan Foundation, under the grant 'Social foundations for statistics and machine learning.'

Abstract

Most definitions of algorithmic bias and fairness encode decision-maker interests, such as profits, rather than the interests of disadvantaged groups (e.g. racial minorities): bias is *defined* as a deviation from profit maximization. Future research should instead focus on the causal effect of automated decisions on the distribution of welfare, both across and within groups. The literature emphasizes some apparent contradictions between different notions of fairness, and between fairness and profits. These contradictions vanish, however, when profits are maximized. Existing work involves conceptual slippages between statistical notions of bias and misclassification errors, economic notions of profit, and normative notions of bias and fairness. Notions of bias nonetheless carry some interest within the welfare paradigm that I advocate for, if we understand bias and discrimination as mechanisms and potential points of intervention.

Keywords: Algorithmic bias, fairness, discrimination, AI, inequality.

JEL codes: D63, J15, J70, D81.

I. Introduction

Consider the following hypothetical scenario. The human resources department in some company has to decide which employees to promote. The company aims to promote the employees who will contribute most to profits, after promotion. They decide to employ an algorithm to predict future contribution to profits. They will make automated promotion decisions based on these predictions. After introducing the algorithm, several Black employees (let's call them Lakisha and Jamal) are not promoted, while their White colleagues (let's call them Emily and Greg) are. Concerns are raised about this system. An audit of the algorithm finds that, using historical data, it predicted that Lakisha was more likely to have children, and that Jamal was more likely to have care responsibilities for elder relatives, relative to their colleagues. This might reduce their availability for working outside regular hours. An economist called in to testify concludes that the system is *not* discriminatory (i.e. not biased), since the algorithm is maximizing profits, as intended.

Now consider another hypothetical scenario. Some university has to decide which students to admit to their undergraduate programme. They do so using an automated system that is based on high school grades, standardized tests, and demographic information. These variables help to predict future academic performance, as measured by the students' grade point average (GPA) in university exams. Because minority students are historically underrepresented among undergraduates, the admissions algorithm partially corrects for this underrepresentation. The algorithm applies a different cutoff in terms of predicted GPA for different demographic groups. After a complaint by a White applicant (let's call him John) who was not admitted, an auditor of the algorithm concludes that the algorithm *was* biased or discriminating, since it did not treat all applicants equally, given their predicted GPA.

How should we think about the conclusions of the economist and the auditor in these two scenarios? Do their assessments line up with our normative intuitions? Or is there something wrong with the notions of fairness employed? How should a policy-maker who wants to address racial inequality respond to these two scenarios? And what would the large and growing literature on fairness and bias in algorithmic decision making say about this? In this article, I provide an opinionated review of this literature. This review expands on arguments previously made in Kasy and Abebe (2021) and Kasy (2023). This review is motivated by debates around racial discrimination and racial inequality. Most of my discussion is more general, however, and applies equally to discrimination and inequality across other demographic dimensions, such as gender, ethnicity, or religion.

Two paradigms for evaluating decision functions

I contrast two different normative paradigms, the *fairness* paradigm and the *welfare* paradigm. The fairness paradigm, closely related to the notion of *taste based discrimination* in economics, undergirds much of the literature on algorithmic fairness. This paradigm is based on a notion of *bias*, which is defined as a deviation from some legitimate objective or decision-maker motive, such as profits in the case of hiring or promotion decisions. Put differently, fairness in this sense encodes the interests of *capital owners* (profits), rather than the interests of *disadvantaged groups* (e.g. racial minorities) affected by their decisions. Viewed in these terms, it appears as a remarkable, and remarkably successful, ideological sleight of hand to substitute one for the other. Correspondingly, this paradigm rationalizes unequal treatment based on some underlying *merit* of the treated, where the treated might be job candidates or university applicants, and where merit corresponds to their contribution to the legitimate objective. In our first scenario, unequal promotions are rationalized by differences in expected care responsibilities. In our second scenario, unequal admissions are rationalized by differences in expected GPA.

I contrast the fairness paradigm to the welfare paradigm. The welfare paradigm undergirds much of the literature on programme evaluation and welfare economics. This paradigm asks about the *consequences* of unequal treatment for the welfare of those who are treated. These welfare consequences are then traded off in some aggregate notion of social welfare. Aggregate social welfare typically puts a higher weight on improving the welfare of those who are worse off. In our first scenario, promotions have consequences for the income and status of the promoted employees. In our second scenario, admissions have consequences for the educational trajectory and future labour market prospects of the admitted students. If the members of a demographic group are worse off, on average, then improving their welfare should be a priority, according to this paradigm.[1]

Following up on the arguments of Kasy and Abebe (2021), I make the case that we should think about algorithmic decision-making more in terms of the welfare paradigm. This paradigm often leads to different conclusions relative to the dominant fairness paradigm, as the opening vignettes illustrate.

Two reasons to study (algorithmic) bias and discrimination

At this point, the reader might be inclined to disagree with my characterization of the notions of bias and fairness, since we might care about bias even if we subscribe to the welfare paradigm.

There are at least two reasons why we might be concerned about bias in decision-making, whether human or algorithmic. The first reason is *normative*. In line with the fairness paradigm described above, one might argue that inequalities due to bias or discrimination are less legitimate than inequalities due to other factors, and should therefore be eliminated–while other inequalities are allowed to persist. This is the normative interpretation of the notion of taste-based discrimination as introduced by Becker (1957). I draw on Pessach and Shmueli (2020) in reviewing definitions of algorithmic fairness from the computer science literature, which are in this tradition.[2]

The second reason to be concerned about bias and discrimination is *positive* and concerns the mechanisms which generate inequality of welfare. One might care about inequality of welfare across individuals, in line with the welfare paradigm described above, and might therefore aim to understand the mechanisms which generate it. One of these mechanisms might be bias or discrimination, in the narrow sense of a deviation from profit maximization. If, in particular, we are normatively committed to reducing inequality, then understanding the mechanisms which generate inequality is important to inform the choice of policy tools to achieve our goal. Much of the discussion about

[1] A related but distinct line of reasoning is pursued in Small and Pager (2020). Drawing on the literature in sociology, these authors criticize the narrowness of economists' notions of discrimination. They list additional forms of discrimination that should be considered, including institutional discrimination, structural discrimination, and institutional racism. Their expanded notion of discrimination arguably approximates a notion of inequality of welfare, as considered here.

[2] Other references providing a general overview and discussion of algorithmic fairness include Abebe et al. (2020); Benjamin (2019); Broussard (2018); Gebru (2019); Noble (2018); O'Neil (2016). For additional discussions of different notions of fairness as well as their feasibility, incompatibility, and politics, we refer to Chouldechova (2017); Friedler et al. (2016); Hardt et al. (2016); Suresh and Guttag (2019); Kleinberg et al. (2016); Mitchell et al. (2018); Narayanan (2018); Verma and Rubin (2018).

algorithmic discrimination is, explicitly or implicitly, based on the normative reason (i.e. the fairness paradigm). I will return to the positive reason to study bias (within the welfare pardigm) in the conclusion.

Apparent contradictions, and conceptual slippages

Another possible objection to my argument, that fairness encodes decision-maker objectives (profits) rather than the interests of disadvantaged groups, could be based on the results of Kleinberg et al. (2016); Chouldechova (2017) and others: Different notions of fairness are mutually inconsistent, in general, and they impose constraints on profit maximization. There seem, therefore, to be real tradeoffs between profits and fairness. How can that be, if all these definitions of fairness are just measuring deviations from profit maximization, as I have claimed?

In section V, I provide a formal discussion of these questions. I show that leading fairness definitions are approximately satisfied whenever profit maximization is approximately satisfied. The tensions between them vanish in the limit. Relatedly, I also discuss the conceptual slippages that often occur between statistical notions of bias and predictive loss, economic notions of profit motives, and normative notions of fairness and discrimination.

Outline: The rest of this article is organized as follows. In section II, I introduce a simple formal model of hiring decisions, which will serve as a running example throughout the article. In section III, I contrast two paradigms for normatively evaluating decision functions, fairness and welfare. In section IV, I review a number of definitions of fairness which have been proposed in the literature. In section V, I discuss the formal relationship between profits, different notions of fairness, and statistical notions of bias and predictive errors. In section VI, I conclude and briefly discus how bias and discrimination can be mechanisms generating inequality of welfare, and how they can consitute potential points of intervention.

Notation: Any probabilities or expectations in this paper are across the distribution of job applicants. I use capital letters to denote random variables, that is, properties of job applicants. Let for example A denote gender, with $A = 1$ for women. Then $P(A = 1)$ is the share of women among job applicants. $P(B|A)$ denotes the conditional probability distribution of B given A. Let for example B denote age. Then $P(B|A = 1)$ denotes the age distribution among female job applicants. Similarly, $E[B|A]$ denotes the conditional expectation of B given A. In our example, $E[B|A = 1]$ then denotes the average age among women job applicants. I assume that all random variables have discrete support, to avoid technicalities. Lastly, 1 is the indicator function which equals 1 if its argument is true, and 0 otherwise. For example, $1(B > 40)$ equals 1 for everyone who is older than 40 years.

II. Setup

Consider an employer who selects whom to hire among a set of candidates for a job opening. We denote the decision to hire a candidate by D, where D may take the values 0 (don't hire) or 1 (hire). Hiring a candidate entails a wage cost w, where w does not vary across job candidates. The employer wants to hire candidates according to their marginal contribution to profits, also called *merit*, or *productivity*. We denote this contribution to profits by M. If the employer could directly observe M, they would want to hire all the candidates whose M exceeds the wage w.

Typically, however, the employer cannot observe M directly. Instead, they observe some other features X of the candidate, which might be helpful for prediciting M.[3] We denote one of the components of X by A. The variable A may take the values 0 or 1, and describes whether the candidate belongs to some group of interest; for instance whether the candidate belongs to some racial minority, is a woman, etc.

The employer determines the probability d of hiring a given candidate based on the features X, that is,

$$d(X) = P(D = 1|X).$$

When the literature on algorithmic decision-making talks about notions such as fairness, bias, or discrimation, then these are typically properties of the function d.

Profit maximizing hiring decisions

Suppose now that the employer makes hiring decisions that maximize expected profit, where expected profit is given by $E[D \cdot (M - w)]$. Denote the expected productivity of a candidate given their features X by

$$m(X) = E[M|X].$$

[3] In line with the literature, I use the terms 'features,' 'covariates,' and 'predictors' interchangeably.

With this notation, and using that D is by construction independent of M given X because X includes all the information available to the employer, expected profit for the hiring function d is given by

$$E[d(X) \cdot (m(X) - w)].$$

The employer does not observe M. They do, however, observe the features X. Suppose that they also know the population relationship m of productivity M to the observed features X. Under this assumption, the profit-maximizing decision function $d^*(\cdot)$ hires everyone whose expected productivity exceeds the wage w, that is,

$$d^*(X) = 1(m(X) > w).$$

Relatedly, suppose that there is a pair of values x, x' for the feature vector X such that

$$m(x) > m(x'), \qquad d(x) < 1, \qquad \text{and} \qquad d(x') > 0.$$

Then profits could be increased by hiring more candidates with features x and fewer candidates with features x', holding constant the total number of candidates who are hired. This observation is at the heart of tests for algorithmic bias, in the sense of deviations from profit maximization. We will return to this observation in section (i) below.

Alternative interpretations

Throughout this paper, I emphasize the example of hiring. There are many other contexts where similar considerations apply, however, including the following:

1. **Consumer credit**
   A bank decides which consumer credit applications to approve ($D = 1$). They might wish to do so based on the probability of repayment M. Since repayment is not observed ex-ante, they might try to predict it based on observable features X.

2. **Bail setting**
   A judge decides whether to grant bail to a defendant ($D = 1$). They might wish to do so based on the probability of recidivism, i.e. further police encounters M of the defendant. Since recidivism is not observed ex-ante, they might try to predict it based on observable features X.

3. **Student admissions**
   A university decides which students to admit ($D = 1$). They might wish to do so based on future student performance as measured by their grade point average M. Since future student performance is not observed ex-ante, they might try to predict it based on observable features X.

4. **Medical care**
   A medical provider decides which patients should receive preventative care ($D = 1$). The provider might wish to do so based on the probability of future chronic health conditions M. Since future chronic health conditions are not observed ex-ante, the provider might try to predict them based on observable features X.

III. Two paradigms

Having discussed profit-maximizing decisions of the employer, I next contrast two distinct paradigms for the normative assessment of such decisions, building on Kasy and Abebe (2021). The first paradigm, encapsulated in notions such as fairness, bias, and discrimination, asks whether unequal treatment across groups can be justified. Justification here typically requires that unequal treatment contributes to profit maximization. Only deviations from profit maximization are considered unjustified.

The second paradigm, encapsulated in causal inference, programme evaluation, and social welfare assessments, asks about the consequences of a particular decision procedure for the welfare of those impacted, and in particular for the inequality of welfare, both across and within groups.[4]

(i) Profits and fairness: rationalizing unequal treatment

Consider a hiring function d. How can we test whether this hiring function maximizes profits? As discussed above, if we can find a pair of values x, x' for the feature vector X such that $m(x) > m(x')$, $d(x) < 1$, and $d(x') > 0$, then we can conclude that d is not profit-maximizing. Increasing $d(x)$ and decreasing $d(x')$ would increase profits.

Write now $X = (A, Z)$, where Z includes all features except for group membership A. Suppose that $x = (a, z)$ and $x' = (a', z)$ for $a' \neq a$. If $m(a, z) > m(a', z)$, $d(a, z) < 1$, and $d(a', z) > 0$, then group a is treated worse than group a' in a way that hurts profits.

If we make the further strong assumption that there are no additional components Z observed by the employer, besides group membership, then this condition becomes $m(a) > m(a')$, $d(a) < 1$, and $d(a') > 0$. This observation provides a rationalization for the 'hit rate test' for discrimination of Knowles et al. (2001),[5] and leads to the condition

$$E[M|D = 1, A = 1] - E[M|D = 1, A = 0] = 0$$

as a test of discrimination. We will meet this condition again in section IV below, under the names *predictive parity* or *balance for positive class*.

Let us recapitulate the logic of this argument. We started from the decision-maker's objective – profits, in our running example. We then derived a necessary condition for maximization of profits. This condition required that we could not improve profits by hiring more from one group and less from another group, conditional on other features Z observed by the employer. If this condition is violated, then profits can be improved, and the decision function is called unfair. This logic is at the heart of the definition of taste-based discrimination that was originally introduced by Becker (1957), who equated competitive market outcomes to fairness *by definition*, no matter how unequal these outcomes are between or within groups.

This same logic is reflected in other notions of fairness, as reviewed in section IV below. Fairness in this sense is a formalization of the decision-maker's interests, and *not* a formalization of the interests of disadvantaged groups. Those two might sometimes be aligned, but they are not in general, as the two vignettes at the start of this paper illustrate. Instead, in general these notions of fairness boil down to rationalizations of unequal treatment. They justify unequal treatment D on the basis of unequal merit M, where merit corresponds to the contribution to the decision-maker's objective.

It should be noted that these notions of fairness are also consistent with statistical discrimination (Phelps, 1972; Aigner and Cain, 1977): If A is predictive of M, conditional on Z, then it is profit-maximizing to take A into account when making hiring decisions, and thus also fair. Much recent work in information economics, learning theory, and mechanism design, as reviewed by Onuchic (2022), builds on the notion of statistical discrimination in constructing potential explanations of (racial) inequality. With the exception of biased beliefs, inequality based on such informational mechanisms would not be considered 'unfair,' in the sense considered here.

(ii) Causal effects and welfare: the consequences of unequal treatment

Let us now contrast the fairness paradigm with an alternative paradigm, which asks 'What is the impact of a treatment assignment algorithm on the distribution of welfare of those who are subject to this algorithm?' Denote individual welfare by Y. Welfare Y is, in general, distinct from the treatment D. Below, we briefly discuss different possible notions of welfare.

Causal effects

Following standard notation in causal inference (Imbens and Rubin, 2015), consider the potential outcomes Y^0, Y^1. Here Y^d, for $d \in \{0, 1\}$, denotes the welfare that an individual would experience if they were treated with $D = d$.

[4] These two paradigms bear some imperfect resemblance to the concepts of direct and indirect discrimination in EU and UK law, and the concepts of disparate treatment and disparate impact in US law; cf. Adams-Prassl et al. (2023). Direct discrimination and disparate treatment are present, roughly speaking, when group membership affected treatment in an undue way, while indirect discrimination and disparate impact are present when a system impacts groups differentially. The latter is permitted, however, if there is a 'business necessity.' In this review I focus on normative questions and sidestep the issues of legal strategy discussed in Adams-Prassl et al. (2023).

[5] Knowles et al. (2001) derive this condition in the context a more complicated equilibrium model, where M is endogenous to d.

Their realized welfare can then be written as

$$Y = D \cdot Y^1 + (1 - D) \cdot Y^0.$$

This equation presumes that outcomes are not affected by the treatment of other individuals; an assumption that I maintain for simplicity.

Assume that D is independent of potential outcomes given X. This holds by construction if X includes all the information available to the employer or algorithm when assigning D. Under this assumption of conditional independence, we can identify the distribution of potential outcomes[6] from

$$P(Y^d|X) = P(Y|D = d, X).$$

Counterfactual distributions of welfare

If the employer hires an individual characterized by features X with probability $d(X)$, then the conditional distribution of realized welfare Y is a mixture of the two potential outcome distributions $P(Y^1|X)$ and $P(Y^0|X)$,

$$P(Y = y|X) = P(Y^1 = y|X) \cdot d(X) + P(Y^0 = y|X) \cdot (1 - d(X)).$$

The overall distribution of welfare is then given by

$$\begin{aligned} P(Y = y) &= \sum_x P(Y = y|X = x) \cdot P(X = x) \\ &= \sum_x \Big[P(Y|D = 1, X = x) \cdot d(x) + P(Y|D = 0, X = x) \cdot (1 - d(x))\Big] \cdot P(X = x). \end{aligned}$$

This expression is at the heart of the distributional decompositions that have been studied in labour economics; see for instance the review in Firpo et al. (2011).

Social welfare

We have thus derived (i) the counterfactual distribution of welfare, which depends on the assignment algorithm $d(\cdot)$, and (ii) a way to identify this distribution, leveraging the conditional independence that automatically holds for algorithmic decision-making.

In order to make normative statements about the relative desirability of different algorithms, we need to aggregate the distribution of individual welfare into some notion of social welfare, trading off the welfare of different individuals. Such aggregation is at the heart of social choice theory and theories of distributive justice in political philosophy, cf. Roemer (1998). The following provides a brief summary; for a more detailed discussion along the same lines see Kasy (2016) and Kasy (2023).

We evaluate social welfare based on the welfare of a set of individuals $i = 1, \ldots, n$. This raises the question *Who is to be included* in this set of individuals – whose lives matter? Everybody of a certain citizenship, or everybody living in a certain territory? All living human beings? What about future generations? What about animals?

Given the set of individuals, we next need to decide *how to measure their welfare*. The goal is to assign a number Y_i to each individual i, where Y_i measures how well they are doing. A minimal notion would only consider the formal legal rights enjoyed by individuals. A broader notion might also take into account various resources that allow individuals to achieve their objectives, such as education, income, and health. A comprehensive notion of opportunities might aim to take into account all factors that influence individuals' options in life, and evaluate the options effectively available to them. And we might finally consider the outcomes actually achieved by individuals, evaluated either by some common criteria, or by their individual preferences. Utilitarianism, the most common perspective in welfare economics, evaluates individual welfare by the outcomes actually achieved, as evaluated by individual preferences.

Given the set of individuals i, and given evaluations Y_i of their welfare, we finally ask how well society as a whole is doing. Formally, we consider a *social welfare function*,

$$F(Y_1, \ldots, Y_n).$$

[6] Identification given X fails if $P(D = d|X) = 0$, which happens for deterministic assignment algorithms, where the 'common support' condition is not satisfied.

The function F determines how much we care about different individuals, i.e. how much weight we assign to the welfare of i relative to the welfare of j. F encodes how much we care about an additional dollar for a poor person versus a rich person, for a sick person versus a healthy person. It might also incorporate the normative assumption that variables such as race, gender, or parental status should not determine individual welfare.

One fairly general class of welfare functions rescales Y by some function ν, and then averages $\nu(Y)$ across individuals, so that

$$F(Y_1, \ldots, Y_n) = E\left[\nu(Y)\right],$$

where the expectation is again an average across $Y_1, \ldots, Y_n$. If ν is concave, then this aggregation encapsulates inequality aversion, by giving a higher weight to marginal increases of Y_i for those with a lower baseline value of Y_i. Counterfactual social welfare $E\left[\nu(Y)\right]$ for such an aggregation is then given by

$$\sum_x \Big[E\left[\nu(Y)|D=1, X=x\right] \cdot d(x) + E\left[\nu(Y)|D=1, X=x\right] \cdot (1-d(x))\Big] \cdot P(X=x).$$

It is also easily possible to form hybrid notions of welfare, which interpolate between the approach just described and standard notions of fairness. While the former only considers the welfare of those being treated, fairness only considers the welfare (profits) of the decision-maker. In general, one might consider a weighted average of the two, which assigns some normative weight to decision-maker profits.[7]

Practical implementation

How would one conduct an evaluation of the distributional or welfare impact of some algorithm in practice? Assume that social welfare is of the form $E\left[\nu(Y, X)\right]$. Assume further that an analyst has at their disposition data where treatment D was algorithmically assigned on the basis of X, so that conditional exogeneity is ensured. Under these conditions, and assuming sufficient support, one might estimate the conditional expectation $E\left[\nu(Y, X)|D, X\right]$ using a suitable flexible parametric or nonparametric regression method. For any counterfactual algorithm which assigns treatment with conditional probability $\tilde{d}(X)$, one can then estimate counterfactual welfare using the sample analogue of the above expression of social welfare,

$$\sum_{X_i} \Big[\widehat{E}\left[\nu(Y, X)|D=1, X=X_i\right] \cdot \tilde{d}(X_i) + \widehat{E}\left[\nu(Y, X)|D=1, X=X_i\right] \cdot (1-\tilde{d}(X_i))\Big],$$

where $\widehat{E}$ denotes the flexible regression estimates of the corresponding conditional expectations. For further detail on this and related estimation approaches, see Firpo et al. (2011, 2009).

IV. A zoo of fairness definitions

Many different definitions of fairness have been proposed in the literature. The following provides a quick overview; readers interested in a more complete list are referred to Pessach and Shmueli (2020). The majority of these definitions formalize the following intuition, which corresponds to the paradigm described in section (i) (see also Bohren et al. 2022): Consider a decision-maker, such as an employer, who makes a hiring decision D. The employer would like to hire candidates with a high contribution M to their profits, and not hire candidates with low M. This is considered legitimate. Differences in hiring probabilities across groups A are therefore justified if they can be rationalized by differences in productivity M. Differences in hiring probabilities that are not rationalizable by differences in M, however, are not justified. This intuition can be formalized in a number of ways, as the following list of criteria shows. In reviewing these criteria from the algorithmic fairness literature, I also relate them to analogous measures of discrimination from the economics literature.

Balance for the positive and negative class

These two criteria require that the distribution of M given $D=1$, or given $D=0$, does not vary across groups A. Balance for the positive class, also known as *predictive parity* (Chouldechova, 2017), requires that

$$E[M|D=1, A=1] - E[M|D=1, A=0] = 0.$$

[7] I thank an anonymous reviewer for this suggestion.

Balance for the negative class similarly requires that

$$E[M|D = 0, A = 1] - E[M|D = 0, A = 0] = 0.$$

Balance for the positive class is a version of the *hit rate test for taste-based discrimination* of Knowles et al. (2001).[8]

Equality of true and false positive rates

True positive rates and false positive rates correspond to the notions of *power and size in statistical testing*, cf. Casella and Berger (2001). Equality of true and false positive rates is analogous to balance for the positive and negative class, while switching the roles of D and M.

Assume that M is binary, $M \in \{0, 1\}$. *Equality of true positive* rates, which has also been called *equality of opportunity* (Hardt et al., 2016), requires that

$$E[D|M = 1, A = 1] - E[D|M = 1, A = 0] = 0.$$

Equality of false positive rates similarly requires that

$$E[D|M = 0, A = 1] - E[D|M = 0, A = 0] = 0.$$

If both conditions are imposed, then this corresponds to the criterion of *equalized odds*.

Conditional statistical parity

Equality of true and false positive rates requires that the probability of treatment does not vary across groups, conditional on M. A closely related criterion replaces M itself by a set of proxy variables X' for M. This is known as *conditional statistical parity*,

$$E[D|A = 1, X' = x'] - E[D|A = 0, X' = x'] = 0.$$

The feature vector X', which might include only a subset of the variables available to the employer, here takes the role of measuring legitimate sources of inequality. Conditional statistical parity corresponds to the *Oaxaca-Blinder decompositions*, which economists have been estimating for many decades, cf. Oaxaca (1973). Considerable controversy has surrounded the question of which controls to include in X', that is, which variables are legitimate sources of inequality.

Conditional statistical parity is also closely related to the *thin veil of ignorance* of Dworkin (1981*a*,*b*), and to the notion of *equality of opportunity* defended in Roemer (2009).

Causal notions of fairness

All definitions that we considered thus far are purely statistical, and do not refer to any causal counterfactuals. The causal analogue of conditional statistical parity would consider the causal effect of A on D, holding constant X', and would require this causal effect to be zero (Kusner et al., 2017). Experimental manipulation of group membership, such as race or gender, is typically not possible, so it is not clear whether this idea is well-defined (Hu and Kohler-Hausmann, 2020). To put it succinctly, one might argue that 'race' does not cause anything, but racism does.

What *can* be done is to consider the causal effect of manipulating *perceived* group membership A on D, holding constant other observable attributes X'. This is the idea that motivates the large literature on audit studies, starting with Bertrand and Mullainathan (2004). In these studies, job applications are submitted, where names (or other ancillary information) that indicate group membership are experimentally manipulated. Arguably, this yields the causal analogue of statistical parity. Consider the potential outcome (structural function) notation $D(a, x')$ for the counterfactual hiring decision for a job application of group a and with features x' other than group membership. Using this notation, we can define the corresponding fairness criterion as

$$E[D(1, X')|X' = x'] - E[D(0, X')|X' = x'] = 0.$$

[8] They compare the probability of finding illegal drugs M among cars searched by police ($M = 1$) on a highway, across racial groups A. A higher probability of finding drugs in the cars of White drivers relative to Black drivers is taken as evidence of discrimination against Black drivers.

Individual fairness

Yet another variation on the notions of equality of true and false positive rates and of statistical parity is the notion of individual fairness (Dwork et al., 2012), which dispenses with the idea of groups. The former notions require that similar individuals should be treated similarly, independent of group membership, but unequal treatment of similar individuals within groups is allowed. Individual fairness, by contrast, requires that all similar individuals should be treated similarly, whether or not they are members of different groups. We can think of this as a requirement of *horizontal equity*. As such, individual fairness does not impose any restrictions on the distribution of D across dissimilar individuals.

Formally, suppose that $d(X_i, X_j)$ is a measure of distance (dissimilarity) between individuals i and j. Then individual fairness requires that

$$E[D|X = x_i] - E[D|X = x_j] \approx 0 \text{ for } d(X_i, X_j) \approx 0.$$

Disparate impact and Demographic parity

We conclude with two measures of fairness that are *not* based on the intuition of rationalizing unequal treatment in terms of profit maximization. These notions of fairness are instead based on the idea that the distribution of D should be the same across groups A, without any additional conditioning variables such as M or X' (Calders et al., 2009).

The first of these notions is *disparate impact*, which requires that

$$\frac{E[D|A = 1]}{E[D|A = 0]} = 1.$$

The second of these notions is *demographic parity*, which imposes

$$E[D|A = 1] - E[D|A = 0] = 0.$$

These two requirements are of course the same, even if the corresponding measures of bias are not. The notion of disparate impact (or adverse impact) plays an important role in the (US) legal context (Vinik, 2023), where a value of the ratio defining disparate impact that is below 80 per cent is often taken as an indicator of discrimination.

V. Profits, fairness, and statistical bias

In the preceding sections I have argued that most standard definitions of fairness encode decision-maker objectives, such as profits. These definitions require that inequalities in treatment status D are justified by inequalities in M, where M is individual contribution to profits. This is exactly what profit maximization requires.

How does this argument square with the emphasis in the literature that (i) different definitions of fairness are, in general, mutually inconsistent, and that (ii) imposing fairness comes at a cost for profit maximization? This section provides some clarification. I first discuss a simplified version of the argument in Kleinberg et al. (2016), who proved that mutual fairness definitions are, in general, inconsistent. This is stated as Proposition 1 below. I will then prove a simple result that demonstrates that (i) maximization of profits and minimization of classification loss are (almost) equivalent, and (ii) if profits or classification loss are close to optimal, then various definitions of fairness are almost satisfied. This is stated as Proposition 2 below. Expected classification loss is the probability that D is not equal to M. This is a natural performance criterion if we think of $d(X)$ as a predictor for M. The upshot of Proposition 2 is that profit maximization implies fairness, confirming the main argument of the present review. The apparent contradictions between different definitions of fairness vanish in the limit of profit-maximizing decisions. What continues to exist in this limit is the tension between profits and fairness on the one hand, and social welfare on the other hand, as we show by relating the notion of demographic parity to both profit maximization and to equality of welfare.

I then discuss various reasons why profits might not be close to optimal, and why correspondingly definitions of fairness might not be satisfied, including mismeasured outcomes, selection bias, partial observability, and finite sample uncertainty. Lastly, I discuss the various conceptual slippages that can occur between statistical notions of bias and misclassification errors, economic notions of profit, and normative notions of bias and fairness.

Table 1: Distribution of M and D given $A = a$

		D	
		0	1
M	0	p^a_{00}	p^a_{01}
	1	p^a_{10}	p^a_{11}

Notation for the case of binary *M*

To facilitate our discussion, I focus on the case where M is binary, $M \in \{0, 1\}$, throughout this section. I denote $p_{md} = P(M = m, D = d)$ and $p^a_{md} = P(M = m, D = d|A = a)$. We can thus represent the joint distribution of M and D given $A = a$ as follows:

(i) Apparent contradictions

In an influential article, Kleinberg et al. (2016) have shown that several definitions of fairness are mutually incompatible, except in special cases. Similar results were obtained independently by Chouldechova (2017). The following is a simplified version of the main result of Kleinberg et al. (2016).

Proposition 1. *Suppose that both* $P(M|D, A)$ *and* $P(D|M, A)$ *do not depend on* A. *Suppose further that* $p_{md} \neq 0$ *for at least* 3 *of the* 4 *possible values of* (m, d). *Then* $p^a_{md} = p_{md}$ *for all* a, m, d.

In words, we can state Proposition 1 as follows:

- Suppose that D is not a deterministic function of M. This is the meaning of $p_{md} \neq 0$ for at least 3 values. This holds whenever the employer cannot perfectly infer productivity M from observables X.
- Suppose further that the base rate $E[M|A]$ varies with A, that is, the probability that $M = 1$ is not the same for all groups A. This condition implies that we cannot have $p^a_{md} = p_{md}$ for all a, m, d. This will, in general, be the case in the presence of pre-existing inequality across groups.
- Then we cannot simultaneously satisfy (i) balance for the positive and negative class (independence of $P(M|D, A)$ of A), and (ii) equality of true positive and false positive rates (independence of $P(D|M, A)$ of A).

The proof of Proposition 1 is given in Appendix VI. It is based on the following argument: Balance for the positive and negative class pins down the ratio of the entries of Table 1 within columns. Equality of false positive and false negative rates pins down the ratio of the entries of Table 1 within rows. Having thus pinned down the ratio of all entries of Table 1, the fact that probabilities sum to 1 pins down the values of p^a_{md} for all m, d, which implies $p^a_{md} = p_{md}$ for all a, m, d.

Interpretation

One way to interpret this result is that there is a fundamental normative tension between different conceptions of fairness, such as balance for the positive and negative class, and equality of false positive and false negative rates. Such a tension requires us to make judgement calls about their relative importance. An even stronger interpretation would be that 'fairness is impossible,' and we might as well give up. Such an interpretation echoes earlier – and, in my opinion, equally misguided – interpretations of Arrow's impossibility theorem (Arrow, 1951), which some have interpreted as implying that 'democracy is impossible.'

Instead, Proposition 1 shows that balance for the positive and negative class, and equality of false positive and negative rates, are imperfect approximations of profit maximization. We next prove the following counterpart to this claim: The closer we are to an assignment of D that maximizes profits, which means that the assignment aligns D with M, the closer we are to satisfying all of these definitions of fairness.

In a nutshell, equality of false positive and of false negative rates, as well as balance for the positive and negative class, require us to *equate* error rates across demographic groups. Maximization of profits, and minimization of

misclassification loss, require us to *minimize* these error rates. If error rates are close to the theoretical minimum of 0 for all groups, they are also close to being equal.

Equivalence

For the binary case, if the decision-maker could observe M, they would choose $D = M$ as the profit-maximizing allocation. We can ask how much lower profits are relative to this hypothetical optimum. This difference is called regret:

$$\mathcal{R} = E[(M - D) \cdot (M - w)] = p_{10} \cdot (1 - w) + p_{01} \cdot w.$$

It is closely related to the misclassification probability,

$$P(M \neq D) = p_{10} + p_{01}.$$

The following proposition implies that, if the regret $\mathcal{R}$ is small, then we are close to achieving fairness in the sense of balance for the positive and negative class, as well as equality of false positive and false negative rates.

Proposition 2. *We can bound the criteria for balance for the positive and negative class, and for equality of true and false positive rates, in terms of profit regret as follows:*

$$\begin{aligned} & |E[D|M = m, A = 1] - E[D|M = m, A = 0]| \\ \leq \; & \frac{\mathcal{R}}{\min(w, 1 - w)} \cdot \left(\frac{1}{P(A = 1, M = m)} + \frac{1}{P(A = 0, M = m)} \right), \end{aligned}$$

and

$$\begin{aligned} & |E[M|D = d, A = 1] - E[M|D = d, A = 0]| \\ \leq \; & \frac{\mathcal{R}}{\min(w, 1 - w)} \cdot \left(\frac{1}{P(A = 1, D = d)} + \frac{1}{P(A = 0, D = d)} \right). \end{aligned}$$

The proof of Proposition 2 is again given in Appendix VI.

Profit maximization, demographic parity, and the inequality of welfare

Our review of fairness definitions in section IV concluded with the definitions of *disparate impact* and *demographic parity*, which require that $E[D|A = 1] = E[D|A = 0]$. In contrast to the other definitions that we reviewed, this requirement is an *unconditional* equality of treatment probabilities across groups, rather than *conditional* equality given M, or given some proxies X' for M.

Unconditional and conditional equality are logically independent. To see this, consider the following two scenarios.

1. **Balance without demographic parity**
   First, assume that we are in the profit-maximizing limit, where $D = M$, but there is inequality between the groups $A = 0, 1$ in terms of M, $E[M|A = 1] < E[M|A = 0]$. This might be the consequence of any number of prior historical inequalities. In this scenario, clearly $E[D|M, A] = M$, independently of A, and balance for the positive and negative class, as well as equality of false positive and false negative rates, are satisfied. Demographic parity, however, is violated.

2. **Demographic parity without balance**
   Second, assume now again that $E[M|A = 1] < E[M|A = 0]$. Assume that $D = 1$ whenever $M = 1$. Assume that additionally $D = 1$ for a subset of individuals in the $A = 1, M = 0$ group, where $P(D = 1, M = 0|A = 1) = E[M|A = 0] - E[M|A = 1]$. Such a scenario might for instance arise due to affirmative action, compensating historical injustices that led to inequalities in M. In this scenario, demographic parity holds, $E[D|A = 1] = E[D|A = 0]$, but conditional independence is violated, i.e. $E[D|M = 0, A = 1] > E[D|M = 0, A = 0]$, leading to inequality of *false positives*. The possibility that independence holds even if conditional independence is violated is sometimes called *Simpson's paradox*.

Demographic parity is in some sense intermediate between fairness notions that condition on M, on the one hand, and true equality of welfare. Demographic parity is not the same as equality of welfare because of (i) inequality within the demographic groups, and (ii) the difference between treatment and welfare; cf. the discussion in Kasy and Abebe (2021). To see the latter, suppose for example that welfare Y is determined by the potential outcomes $Y^d = d - A$. Demographic parity, setting $E[D|A = 1] = E[D|A = 0]$, would imply inequality of welfare across groups, $E[Y|A = 1] < E[Y|A = 0]$, in this example. Reversely, equality of welfare $E[Y|A = 1] = E[Y|A = 0]$, would imply $E[D|A = 1] > E[D|A = 0]$, thus violating demographic parity. Put differently, treatment assignment that compensates historical inequalities which impact welfare is not compatible with demographic parity *or*, typically, with fairness notions that require equality of treatment conditional on M.

(ii) Possible sources of suboptimal profits and of bias

Proposition 2 shows that, if profits are close to their maximum, then fairness criteria are close to being satisfied. There are, however, a number of reasons why profits might not be close to their maximum, why fairness criteria might therefore be violated, and why there are apparent contradictions between different notions of fairness. The following provides an overview of such reasons.

Mismeasured outcomes

One reason is mismeasurement of M. Often, the exact outcome of interest, from the perspective of the decision-maker, is hard or impossible to observe. Worker productivity, as in our motivating example, is a case in point. When proxies are used instead of M itself, this can lead to systematic distortions.

A number of examples have been documented in the literature. Healthcare providers might use algorithms to target healthcare resources to those with the greatest health risks. Obermeyer et al. (2019) document bias in one such algorithm, which assigns lower risk scores D to Black patients, relative to White patients of the same health M. Bias occurs because the algorithm uses health costs as a proxy for health needs, where health costs are also driven by income and insurance status. Another example is the use of recorded police encounters as a proxy for criminality or recidivism. Any bias of the police will be reproduced by algorithms that are based on predicted police encounters, cf. Knox et al. (2020).

Selection bias

Another reason why profit maximization might not be achieved is selection bias. Often, M is only observed when $D = 1$. As discussed in section III, treatment D is exogenous conditional on the features X used for treatment assignment. If we have access to all these features, and if M is observed whenever $D = 1$, then we can identify

$$m(X) = E[M|X] = E[M|X, D = 1],$$

at least for values of X such that $d(X) > 0$. The profit-maximizing decision rule is then given by $d(X) = 1(m(X) > w)$.

Consider now the following situation, instead. Suppose that the designer of the algorithm is using historical data, where decisions were made by humans, or by algorithms using a larger set of predictive featues $\widetilde{X}$. Then, in general, conditional exogeneity of D does not hold, and therefore $E[M|X] \neq E[M|X, D = 1]$. Such selection bias, if it is not taken into account, might lead to violations both of profit maximization and of fairness criteria, cf. Arnold et al. (2022), who also propose a quasi-experimental approach for overcoming selection bias.

Partial observability

Selection bias can occur when there is incomplete observability *by the algorithm designer* of the features which entered historical treatment assignment. A related issue arises when there is incomplete observability *by an auditor or researcher* of the features used by the algorithm under consideration.

To illustrate, consider the hit rate test for taste-based discrimination in Knowles et al. (2001). Suppose that car searches on the highway are indeed determined by $d^*(X) = 1(E[M|X] > w)$, where M indicates whether there are illegal drugs in the car. As discussed above, this implies that, whenever $d^*(x) < 1$ and $d^*(x') > 0$, then $m(x) < m(x')$. Suppose that $x = (a, z)$, that $E[M|D = 1, A = 1] - E[M|D = 1, A = 0] > 0$, $E[D|A = 1] < 1$, and $E[D|A = 0] > 0$. Does this imply a violation of profit maximization? Not necessarily, as the following example illustrates.

Suppose that $w = .5$, $Z \in \{0, 1\}$, $E[Z|A] = E[M|A] = .5$ (so that both M and Z are independent of A), but $E[M|Z = 1, A = 1] = 1$, while $E[M|Z = 1, A = 0] = 3/4$. Then $d(X) = 1(E[M|X] > w) = Z$ is profit-maximizing,

but $E[M|D=1,A=1]-E[M|D=1,A=0]=1/4$. The hit rate test thus indicates taste-based discrimination, that is, balance for the positive class does not hold, even though D is chosen optimally by the decision-maker.

Finite samples

Yet another reason for deviations from profit maximization is the fact that typically m needs to be estimated based on finite samples. This is where machine learning, or more traditional regression methods, come into play.

In particular, a prediction $\widehat{m}(x)$ of M given $X=x$ is typically estimated based on a sample of historical data (X_i,M_i). The expected squared prediction error $E\left[(M-\widehat{m}(x))^2|X=x\right]$ can be decomposed as

$$\text{Var}\,(M|X=x)+\text{Var}\,(\widehat{m}(x))+(E[\widehat{m}(x)]-E[M|X=x])^2.$$

Generally, both the second term (the variance of the prediction) and the third term (the squared bias of the prediction) will be non-zero. At the heart of most supervised learning methods are algorithms that trade off these two components of prediction errors, to minimize overall mispredictions. Both estimator variance and bias will be non-zero for finite training samples.

(iii) Statistical accuracy, profits, and fairness

We have encountered definitions from three different domains: statistical (prediction and testing), economic (profit maximization), and normative (fairness and discrimination). These domains are related but distinct, and it is easy to slip from one to the other. One might for instance slip from statistical notions of unbiasedness (correct predictions on average, conditional on predictive features) to normative notions of unbiasedness (no *unjustified* inequality in treatment between groups). To conclude this section, let us briefly remind ourselves of the definitions of these notions, and how they relate.

Statistical accuracy

The goal in supervised machine learning (prediction, regression, classification) is to find a function d of X which minimizes the expected loss $E[l(d(X),M)]$ for predicting M. Typical loss functions are the misclassification loss $l(d(X),M)=1(d(X)\neq M)$, and the squared error loss $l(d(X),M)=(d(X)-M)^2$. For binary M and $d(X)$ these two are equivalent, and given by

$$E\left[(d(X)-M)^2\right]=P(M\neq D)=p_{10}+p_{01}.$$

More generally, without imposing binary predictions, we can decompose the *mean squared prediction error* $E\left[(d(X)-M)^2\right]$ as

$$E\left[E\left[(d(X)-M)^2|X\right]\right]=E\left[\text{Var}\,(M|X)+(d(X)-E[M|X])^2\right].$$

The variance $\text{Var}\,(M|X)$ does not depend on d. The difference $d(X)-E[M|X]$ can be thought of as a *bias* of the prediction, conditional on d and X.[9] Optimal predictions thus minimize bias, in this sense.

Closely related to these notions of prediction errors are (frequentist) notions of testing. Consider the null hypothesis that $M=0$, for a particular individual, and interpret D as the outcome of a statistical test. Then $E[D|M=0]=\frac{p_{01}}{p_{00}+p_{01}}$ is the *false positive rate*.

Analogously, $E[D|M=0,A=a]=\frac{p^a_{01}}{p^a_{00}+p^a_{01}}$ is the false positive rate in group a, that is, the probability of wrong rejections of the null. This corresponds to the *size* of statistial tests. Furthermore, $E[D|M=1,A=a]=\frac{p^a_{11}}{p^a_{10}+p^a_{11}}$ is the *true positive* rate for group a, which corresponds to the *power* of statistical tests. Importantly, these notions of false positives and false negatives are defined *conditional* on M, rather than averaging over M.

Profit maximization

We have already discussed at length the profit objective, $E[D\cdot(M-w)]$. Maximizing profit is equivalent to minimizing profit regret $\mathcal{R}=E[(M-D)\cdot(M-w)]$, which for binary M and D equals $\mathcal{R}=p_{10}\cdot(1-w)+p_{01}\cdot w$. Profit regret is almost the same as the misclassification loss $P(M\neq D)=p_{10}+p_{01}$, up to the weighting by w and $1-w$.

[9] If we take into account the randomness of d itself, which comes from the fact that it was trained on random data, then $E\left[(d(X)-E[M|X])^2\right]$ can in turn be decomposed into variance and bias.

Minimizing misclassification loss $p_{10} + p_{01}$ is equivalent to minimizing both false positive rates $\frac{p_{01}}{p_{00}+p_{01}}$ and false negative rates $\frac{p_{10}}{p_{10}+p_{11}}$, and equivalent to minimizing profit regret $p_{10} \cdot (1-w) + p_{01} \cdot w$, as shown by Proposition 2 above. This, of course, hinges on predicting the actual contribution M of individuals to profits, for the population of interest. Mismeasured M or selection bias lead to reduced profits, even if in the data prediction error rates are small.

Fairness

Equality of false positive and of false negative rates requires to *equate* the error rates $p^a_{m,1-m}/P(M = m|A = a)$ across groups A, for $m = 0, 1$. Similarly, balance for the positive and negative class requires to *equate* the error rates $p^a_{1-d,d}/P(D = d|A = a)$ across groups A, for $d = 0, 1$.

By contrast, profit maximization requires to *minimize* the error rates $p^a_{1-d,d}$. The theoretical optimum for any of these rates is 0. If profits are close to optimal, then error rates are close to 0 within in each group, and therefore also close to being equal across groups. This is reflected in the result of Proposition 2 above.

VI. Conclusion

In this review, I have argued that leading definitions of (algorithmic) fairness encode the objective of profit maximization, rather than the welfare of disadvantaged groups. I have contrasted such definitions of fairness with an alternative paradigm that emphasizes the causal impact of algorithmic decision-making systems on the distribution of welfare, both across and within groups.

Within this latter normative paradigm, we might still very much care about questions of discrimination. Understanding the mechanisms that lead to inequality of welfare, including racial inequality, is a pre-condition for being able to intervene on them, and to reduce the inequality of welfare.

One possible mechanism might indeed be algorithmic bias, in the sense of systematically distorted predictions that disadvantage certain groups. To what extent this mechanism is an important contributor to the inequality of welfare is an empirical question. Two points are worth recalling here. First, a profit-maximizing firm which is able to predict well will be close to satisfying most standard notions of algorithmic fairness, as shown by Proposition 2. To the extent that profit maximization and the ability to predict well are empirically plausible, we might expect that algorithmic bias is not a central contributor to racial inequality.

Second, however, even if it turns out that algorithms are not systematically biased, that does not mean that these algorithms do not create or amplify racial inequality, and decrease social welfare. Predictive bias and the impact on inequality are two very different objects. In Kasy and Abebe (2021), this is discussed in the context of affirmative action or redistribution (which are generally good for equality of welfare but bad for 'fairness'), and improved predictive capacity (which is generally good for fairness, but increases the inequality of treatment).

There are many mechanisms through which algorithms might create or amplify racial inequality. To list but some examples, this includes any form of 'statistical discrimination,' the many variations of learning-based inequality, multiple equilibria, self-enforcing norms, etc. that the recent economic theory literature has discussed (Onuchic, 2022). This also includes notions of merit M that might systematically disadvantage some groups, as a consequence of historic disadvantage (Small and Pager, 2020), due to systematically distorted proxies (Knox et al., 2020; Obermeyer et al., 2019), or because they are adjusted in an ad hoc manner (Uhlmann and Cohen, 2005). This further includes recommender systems on social networks which use existing network connections to make recommendations, thereby systematically privileging more centrally located individuals, and increasing racial segregation of networks (Stoica et al., 2018, 2020). This finally includes algorithmic pricing mechanisms. Geographic mobility might for example be lower for women and disadvantaged minorities (Manning and Petrongolo, 2017). Algorithmic pricing mechanisms might learn implied differences in price elasticities or wage elasticities, and charge these groups higher prices, or offer lower wages.

There remains much theoretical and empirical work to be done to understand these manifold mechanisms through which algorithms reproduce and amplify existing inequalities in society. In order to get there, we need to stop asking whether unequal treatments are justifiable by profit maximization, and start asking what the impact of automated decisions is on the distribution of welfare.

Appendix – Proofs

Proof of Proposition 1

Suppose that $p_{md} \neq 0$ for $(m,d) \neq (1,1)$ (the other cases follow analogously). Independence of $P(M|D,A)$ and $P(D|M,A)$ of A then implies the following four eqalities:

$$\frac{p^a_{10}}{p^a_{00}} = \frac{p_{10}}{p_{00}}, \qquad \frac{p^a_{11}}{p^a_{01}} = \frac{p_{11}}{p_{01}},$$
$$\frac{p^a_{01}}{p^a_{00}} = \frac{p_{01}}{p_{00}}, \qquad \frac{p^a_{11}}{p^a_{10}} = \frac{p_{11}}{p_{10}}.$$

The top two equalities correspond to balance for the positive and negative class (and to the columns of Table 1); the bottom two correspond to equality of false positive and false negative rates (and to the rows of Table 1).

Note that probabilities sum to 1, so that $p^a_{00}+p^a_{10}+p^a_{11}+p^a_{01}=1$. We can substitute the equalities into this sum to get

$$p^a_{00} \cdot \left(1 + \frac{p_{10}}{p_{00}} + \frac{p_{11}}{p_{00}} + \frac{p_{01}}{p_{00}}\right) = 1,$$

and thus, after multiplying this equation by p_{00}, $p^a_{00} = p_{00}$. A similar argument applies to p^a_{10}, p^a_{00}, and p^a_{01}.

Proof of Proposition 2

We first note that

$$\mathcal{R} = p_{10} \cdot (1-w) + p_{01} \cdot w,$$

and thus

$$(p_{10}+p_{01}) \cdot \min(w, 1-w) \leq \mathcal{R} \leq (p_{10}+p_{01}) \cdot \max(w, 1-w)$$

In words, regret is small if and only if the misclassification probability $P(M \neq D) = p_{10}+p_{01}$ is small. Second, by the law of total probability,

$$P(M \neq D) = \sum_{m,a} P(M \neq D | A=a, M=m) \cdot P(A=a, M=m),$$

and thus

$$P(M \neq D | A=a, M=m) \leq \frac{P(M \neq D)}{P(A=a, M=m)}.$$

Third,

$$\begin{aligned} &|E[D|M=m, A=1] - E[D|M=m, A=0]| \\ =&|P(M \neq D|M=m, A=1) - P(M \neq D|M=m, A=0)| \\ \leq& P(M \neq D|M=m, A=1) + P(M \neq D|M=m, A=0). \end{aligned}$$

Putting these three observations together, we get

$$\begin{aligned} &|E[D|M=m, A=1] - E[D|M=m, A=0]| \\ \leq& \frac{\mathcal{R}}{\min(w, 1-w)} \cdot \left(\frac{1}{P(A=1, M=m)} + \frac{1}{P(A=0, M=m)}\right). \end{aligned}$$

The first claim follows. The second claim is proven analogously.

References

Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D. G. (2020), 'Roles for Computing in Social Change', in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–60.

Adams-Prassl, J., Binns, R., and Kelly-Lyth, A. (2023), 'Directly Discriminatory Algorithms', *The Modern Law Review*, **86**(1), 144–75.

Aigner, D. J., and Cain, G. G. (1977), 'Statistical Theories of Discrimination in Labor Markets', *Industrial and Labor Relations Review*, 175–87.

Arnold, D., Dobbie, W., and Hull, P. (2022), 'Measuring Racial Discrimination in Bail Decisions', *American Economic Review*, 112(9), 2992–3038.

Arrow, K. J. (1951), 'Social Choice and Individual Values', in *Social Choice and Individual Values*, New Haven, CT, Yale University Press.

Becker, G. S. (1957), *The Economics of Discrimination*, Chicago, IL, University of Chicago Press.

Benjamin, R. (2019), *Race after Technology: Abolitionist Tools for the New Jim Code*, John Wiley & Sons.

Bertrand, M., and Mullainathan, S. (2004), 'Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination', *American Economic Review*, **94**(4), 991–1013.

Bohren, J. A., Hull, P., and Imas, A. (2022), 'Systemic Discrimination: Theory and Measurement', Technical report, National Bureau of Economic Research.

Broussard, M. (2018), *Artificial Unintelligence: How Computers Misunderstand the World*, Cambridge, MA, MIT Press.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009), 'Building Classifiers with Independency Constraints', in 2009 IEEE International Conference on Data Mining Workshops, 13–18.

Casella, G., and Berger, R. L. (2001), *Statistical Inference*, Duxbury Press.

Chouldechova, A. (2017), 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments', *Big Data*, **5**(2), 153–63.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012), 'Fairness through Awareness, in Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, Cambridge, MA, ACM, 214–26.

Dworkin, R. (1981*a*), 'What is Equality? Part 1: Equality of Welfare', *Philosophy & Public Affairs*, 185–246.

— (1981*b*), 'What is Equality? Part 2: Equality of Resources', *Philosophy & Public Affairs*, 283–345.

Firpo, S., Fortin, N., and Lemieux, T. (2009), 'Unconditional Quantile Regressions', *Econometrica*, 77, 953–73.

— — — (2011), 'Decomposition Methods in Economics', *Handbook of Labor Economics*, **4**, 1–102.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016), 'On the (Im) Possibility of Fairness', *arXiv preprint arXiv:1609.07236*.

Gebru, T. (2019), 'Race and Gender', in *Oxford Handbook on AI Ethics*, Oxford, Oxford University Press.

Hardt, M., Price, E., and Srebro, N. (2016), 'Equality of Opportunity in Supervised Learning, in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, 3315–23.

Hu, L., and Kohler-Hausmann, I. (2020), 'What's Sex Got to Do with Fair Machine Learning?', in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, ACM.

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge, Cambridge University Press.

Kasy, M. (2016), *Empirical Research on Economic Inequality*.

— (2023), 'The Political Economy of AI: Towards Democratic Control of the Means of Prediction', Working Paper.

— Abebe, R. (2021), 'Fairness, Equality, and Power in Algorithmic Decision Making', *ACM Conference on Fairness, Accountability, and Transparency*.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016), 'Inherent Trade-offs in the Fair Determination of Risk Scores', *arXiv preprint arXiv:1609.05807*.

Knowles, J., Persico, N., and Todd, P. (2001), 'Racial Bias in Motor Vehicle Searches: Theory and Evidence', *Journal of Political Economy*, **109**(1), 203–29.

Knox, D., Lowe, W., and Mummolo, J. (2020), 'Administrative Records Mask Racially Biased Policing', *American Political Science Review*, **114**(3), 619–37.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017), 'Counterfactual Fairness', in I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds), *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc.

Manning, A., and Petrongolo, B. (2017), 'How Local are Labor Markets? Evidence from a Spatial Job Search Model', *American Economic Review*, **107**(10), 2877–907.

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., and Lum, K. (2018), 'Prediction-based Decisions and Fairness: A Catalogue of Choices, Assumptions, and Definitions', *arXiv preprint arXiv:1811.07867*.

Narayanan, A. (2018), 'Translation Tutorial: 21 Fairness Definitions and their Politics', in *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*.

Noble, S. U. (2018), *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York, NYU Press.

Oaxaca, R. (1973), 'Male-Female Wage Differentials in Urban Labor Markets', *International Economic Review*, **14**(3), 693–709.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019), 'Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations, *Science*, **366**(6464), 447–53.
O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Broadway Books.
Onuchic, P. (2022), 'Recent Contributions to Theories of Discrimination', *arXiv preprint arXiv:2205.05994*.
Pessach, D., and Shmueli, E. (2020), 'Algorithmic Fairness', *arXiv preprint arXiv:2001.09784*.
Phelps, E. S. (1972), 'The Statistical Theory of Racism and Sexism', *The American Economic Review*, **62**(4), 659–61.
Roemer, J. E. (1998), *Theories of Distributive Justice*, Cambridge, MA, Harvard University Press.
—(2009), *Equality of Opportunity*, Cambridge, MA, Harvard University Press.
Small, M. L., and Pager, D. (2020), 'Sociological Perspectives on Racial Discrimination', *Journal of Economic Perspectives*, **34**(2), 49–67.
Stoica, A.-A., Han, J. X., and Chaintreau, A. (2020), 'Seeding Network Influence in Biased Networks and the Benefits of Diversity', in *Proceedings of The Web Conference 2020*, 2089–98.
—Riederer, C., and Chaintreau, A. (2018), 'Algorithmic Glass Ceiling in Social Networks: The Effects of Social Recommendations on Network Diversity', in *Proceedings of the 2018 World Wide Web Conference*, 923–32.
Suresh, H., and Guttag, J. V. (2019), 'A Framework for Understanding Unintended Consequences of Machine Learning', *arXiv preprint arXiv:1901.10002*.
Uhlmann, E. L., and Cohen, G. L. (2005), 'Constructed Criteria: Redefining Merit to Justify Discrimination', *Psychological Science*, **16**(6), 474–80.
Verma, S., and Rubin, J. (2018), 'Fairness Definitions Explained', in *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, IEEE, 1–7.
Vinik, D. F. (2023), 'Disparate Impact', in *Encyclopedia Britannica*.