The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*

Sam Corbett-Davies
Stanford University

Sharad Goel
Stanford University

August 14, 2018

Abstract

The nascent field of fair machine learning aims to ensure that decisions guided by algorithms are equitable. Over the last several years, three formal definitions of fairness have gained prominence: (1) anti-classification, meaning that protected attributes—like race, gender, and their proxies—are not explicitly used to make decisions; (2) classification parity, meaning that common measures of predictive performance (e.g., false positive and false negative rates) are equal across groups defined by the protected attributes; and (3) calibration, meaning that conditional on risk estimates, outcomes are independent of protected attributes. Here we show that all three of these fairness definitions suffer from significant statistical limitations. Requiring anti-classification or classification parity can, perversely, harm the very groups they were designed to protect; and calibration, though generally desirable, provides little guarantee that decisions are equitable. In contrast to these formal fairness criteria, we argue that it is often preferable to treat similarly risky people similarly, based on the most statistically accurate estimates of risk that one can produce. Such a strategy, while not universally applicable, often aligns well with policy objectives; notably, this strategy will typically violate both anti-classification and classification parity. In practice, it requires significant effort to construct suitable risk estimates. One must carefully define and measure the targets of prediction to avoid retrenching biases in the data. But, importantly, one cannot generally address these difficulties by requiring that algorithms satisfy popular mathematical formalizations of fairness. By highlighting these challenges in the foundation of fair machine learning, we hope to help researchers and practitioners productively advance the area.

Keywords— Algorithms, anti-classification, bias, calibration, classification parity, decision analysis, measurement error

*We thank Alex Chohlas-Wood, Alexandra Chouldechova, Avi Feller, Aziz Huq, Moritz Hardt, Daniel E. Ho, Shira Mitchell, Jan Overgoor, Emma Pierson, and Ravi Shroff for their thoughtful comments. This paper synthesizes and expands upon material that we first presented in tutorials at the 19th Conference on Economics and Computation (EC 2018) and at the 35th International Conference on Machine Learning (ICML 2018). We thank the audiences at those presentations for their feedback.

1 Introduction

In banking, criminal justice, medicine, and beyond, consequential decisions are often informed by statistical risk assessments that quantify the likely consequences of potential courses of action (Barocas and Selbst, 2016; Berk, 2012; Chouldechova et al., 2018; Shroff, 2017). For example, a bank's lending decisions might be based on the probability that a prospective borrower will default if offered a loan. Similarly, a judge may decide to detain or release a defendant awaiting trial based on his or her estimated likelihood of committing a crime if released. As the influence and scope of these risk assessments increase, academics, policymakers, and journalists have raised concerns that the statistical models from which they are derived might inadvertently encode human biases (Angwin et al., 2016; O'Neil, 2016). Such concerns have sparked tremendous interest in developing *fair* machine-learning algorithms.

Over the last several years, the research community has proposed a multitude of formal, mathematical definitions of fairness to help practitioners design equitable risk assessment tools. In particular, three broad classes of fairness definitions have gained prominence. The first, which we call *anti-classification*, stipulates that risk assessment algorithms not consider protected characteristics—like race, gender, or their proxies—when deriving estimates.[1] The second class of definitions demand *classification parity*, requiring that certain common measures of predictive performance be equal across groups defined by the protected attributes. Under this definition, a risk assessment algorithm that predicts loan default might, for example, be required to produce similar false negative rates for white and black applicants. Finally, the third formal fairness definition, known as *calibration*, requires that outcomes are independent of protected attributes after controlling for estimated risk. For example, among loan applicants estimated to have a 10% chance of default, calibration requires that whites and blacks default at similar rates.

These formalizations of fairness have considerable intuitive appeal. It can feel natural to exclude protected characteristics in a drive for equity; and one might understandably interpret disparities in error rates as indicating problems with an algorithm's design or with the data on which it was trained. However, we show, perhaps surprisingly, that all three of these popular definitions of algorithmic fairness—anti-classification, classification parity, and calibration—suffer from deep statistical limitations. In particular, they are poor measures for detecting discriminatory algorithms and, even more importantly, designing algorithms to satisfy these definitions can, perversely, negatively impact the well-being of minority and majority communities alike.

In contrast to the principle of anti-classification, it is often necessary for equitable risk assessment algorithms to explicitly consider protected characteristics. In the criminal justice system, for example, women are typically less likely to commit a future violent crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions. Recognizing this problem, some jurisdictions, like Wisconsin, have turned to gender-specific risk assessment tools to ensure that estimates are not biased against women. Enforcing classification parity can likewise lead to discriminatory decision making. When the true underlying distribution of risk varies across groups, differences in group-level error rates are an expected consequence of algorithms that accurately capture each individual's risk. This general statistical phenomenon, which we discuss at length below, is known as the problem of *infra-marginality* (Ayres, 2002; Simoiu et al., 2017). Attempts to adjust for these differences often require implicitly or explicitly misclassifying low-risk members of one group as high-risk, and high-risk members of another as low-risk, potentially harming members of all groups in the process. Finally, we show that calibration, while generally desirable, provides only a weak guarantee of equity. In particular, it is often straightforward to satisfy calibration while strategically misclassifying individuals in order to discriminate. Indeed, the illegal practice of redlining in banking is closely related to such a discriminatory strategy. For example, to unfairly limit loans to minority applicants, a bank could base risk estimates only on coarse information, like one's neighborhood, and ignore individual-level factors, like income and credit history. The resulting risk scores would be calibrated—assuming majority and minority applicants default at similar rates within neighborhood—and could be used to deny loans to creditworthy minorities who live in relatively high-risk neighborhoods. Similar discriminatory effects can

[1] The term "anti-classification" is popular among legal scholars, but it is not commonly used by computer scientists. In general, given the interdisciplinarity and nacency of fair machine learning, a variety of terms are often used by different authors to describe the same underlying concept. We have attempted to avoid ambiguity by explicitly defining the key phrases we use.

arise from inexperience rather than malice, with algorithm designers inadvertently neglecting to include important predictors in risk models.

As opposed to following prevailing mathematical definitions of fairness, practitioners have long designed tools that adhere to an alternative fairness concept. Namely, after constructing risk scores that best capture individual-level risk—and potentially including protected traits to do so—similarly risky individuals are treated similarly, regardless of group membership. For example, when determining which defendants to release while they await trial, or which loan applicants to approve, decision makers would first select an acceptable risk level and then release or fund those individuals estimated to fall below that threshold. As we show below, this *threshold policy* follows widely accepted legal standards of fairness. Further, such a decision strategy—with an appropriately chosen decision threshold—maximizes a natural notion of social welfare for all groups. Importantly, however, this thresholding approach will in general violate classification parity, and may additionally violate anti-classification, as some risk assessments use protected characteristics. The underlying risk scores will typically satisfy calibration, but that is largely incidental; it is not the reason for the normative appeal of the approach.

Statistical risk assessment algorithms are often built on data that reflect human biases in past decisions and outcomes. As such, it is important to consider the ways in which problems with the training data can corrupt risk scores. In particular, we focus on measurement error and sample bias. First, the outcome of interest may be imperfectly observed, what we call *label bias*. For example, in the criminal justice system, whites and blacks who commit the same offense are often arrested and convicted for those offenses at different rates, particularly for low-level crimes, like minor drug use. Consequently, statistical models that predict future arrests or convictions can systematically overstate recidivism risk for minorities. Unfortunately, there is no easy solution to this measurement problem, since arrests and convictions are typically all that is observed. Practitioners often combat this issue by focusing on outcomes less likely to exhibit such bias. For example, instead of training models to predict arrests for minor crime, one can predict arrests for violent offenses, which are believed to be less susceptible to measurement error (D'Alessio and Stolzenberg, 2003; Skeem and Lowenkamp, 2016). Second, the predictive power of features can vary across groups, what Ayres (2002) calls the problem of *subgroup validity*. For example, if arrests are differentially recorded across race groups, they might also have differential predictive power. Even absent measurement error, it is in general possible for the relationship between a predictor and outcome to differ across groups, potentially skewing estimates that ignore such distinctions. However, when labels are accurately measured, this phenomenon can be countered by fitting group-specific risk models that learn such idiosyncratic patterns—violating anti-classification. Indeed, as mentioned above, this is precisely the rationale for employing gender-specific recidivism models. Finally, one must take care to ensure that training data are representative of the population to which algorithms are eventually applied. As a case in point, Buolamwini and Gebru (2018) found that commercial image analysis programs have difficulty classifying the gender of dark-skinned individuals, a shortcoming that is potentially due to the relative dearth of dark-skinned faces in popular facial analysis datasets.

In addition to addressing such potential problems with the data, it is important to consider the design and analysis of decision algorithms in more complex environments. While single-threshold policies are often a useful starting point, they do not work in all circumstances, particularly when externalities and equilibrium effects may dominate the immediate, localized costs and benefits of decisions. For example, in banking, one might improve the long-run distribution of wealth by setting different lending standards for different groups; and in education, a diverse student body may yield benefits for all students (Page, 2008), similarly justifying group-specific standards. Importantly, though, this complexity does not mean that the popular formalizations of fairness we study can help one achieve equitable outcomes. Indeed, requiring either anti-classification or classification parity can in fact exacerbate these problems.

The need to build fair algorithms will only grow over time, as automated decisions become even more widespread. As such, it is critical to address limitations in past formulations of fairness, to identify best practices moving forward, and to outline important open research questions. By reviewing and synthesizing recent developments in fair machine learning, we hope to help both researchers and practitioners advance this nascent yet increasingly influential field.

2 Background

To ground our discussion of fair machine learning, we begin by reviewing the leading notions of discrimination in economics and American law. We then formally define the concepts of algorithmic fairness described above, and discuss some recent applications of these definitions.

2.1 Discrimination in law and economics

There are two dominant economic categories of discrimination, statistical (Arrow et al., 1973; Phelps, 1972) and taste-based (Becker, 1957), both of which focus on utility. With statistical discrimination, decision makers explicitly consider protected attributes in order to optimally achieve some non-prejudicial goal. For example, profit-maximizing auto insurers may charge a premium to male drivers to account for gender differences in accident rates. In contrast, with taste-based discrimination, decision makers act as if they have a preference or "taste" for bias, sacrificing profit to avoid certain transactions. This includes, for example, an employer who forfeits financial gain by failing to hire exceptionally qualified minority applicants. In Becker's original formulation of the concept, he notes that taste-based discrimination is independent of intent, and covers situations in which a decision maker acts "not because he is prejudiced against them but because he is ignorant of their true efficiency" (Becker, 1957).[2]

As opposed to utility-based definitions, the dominant legal doctrine of discrimination focuses on a decision maker's motivations. Specifically, equal protection law—as established by the U.S. Constitution's Fourteenth Amendment—prohibits government agents from acting with "discriminatory purpose" (Washington v. Davis, 1976). It bars policies undertaken with animus (i.e., it bars a form of taste-based discrimination, since acting with animus typically means sacrificing utility); but it allows for the limited use of protected attributes to further a compelling government interest (i.e., it allows a form of statistical discrimination). As one example, certain race-conscious affirmative action programs for college admissions are legally permissible to further the government's interest in promoting diversity (Fisher v. University of Texas, 2016).

The equal protection doctrine has evolved over time, and reflects ongoing debates about the role of *classification* (use of protected traits) versus *subordination* (subjugation of disadvantaged groups) in discrimination cases (Fiss, 1976). By law, it is presumptively suspect for government entities to explicitly base decisions on race, gender, or other protected attributes, with such policies automatically triggering heightened judicial scrutiny (Winkler, 2006). In this sense, the principle of anti-classification is firmly encoded in current legal standards. Importantly, however, one can clear this hurdle by arguing that such classifications are necessary to achieve equitable ends—as with affirmative action. There is thus recognition under constitutional law that society's interests are not always served by a mechanical blindness to protected attributes. Further, several legal scholars have argued that courts, even when formally applying anti-classification criteria, are often sympathetic to the potential effects of judgments on social stratification, indicating tacit concern for anti-subordination (Balkin and Siegel, 2003; Colker, 1986; Siegel, 2003). Others, though, have noted that such judicial support for anti-subordination appears to be waning (Nurse, 2014).

In certain situations—particularly those concerning housing and employment practices—intent-free economic notions of discrimination are more closely aligned with legal precepts. Namely, under the statutory *disparate impact* standard, a practice may be deemed discriminatory if it has an unjustified adverse effect on protected groups, even in the absence of explicit categorization or animus (Barocas and Selbst, 2016).[3] The disparate impact doctrine was formalized in the landmark U.S. Supreme Court case *Griggs v. Duke Power Co.* (1971). In 1955, the Duke Power Company instituted a policy that mandated employees have a high school diploma to be considered for promotion, which had the effect of drastically limiting the eligibility of black employees. The Court found that this requirement had little relation to job performance, and thus

[2]Some economists have re-interpreted taste-based discrimination as requiring intent (cf. Bertrand et al., 2005), but we use the term as Becker originally defined it.

[3]The legal doctrine of disparate impact stems largely from federal statutes, not constitutional law, and applies only in certain contexts, such as employment (via Title VII of the 1964 Civil Rights Act) and housing (via the Fair Housing Act of 1968). Apart from federal statutes, some states have passed more expansive disparate impact laws, including Illinois and California. The distinction between statutory and constitutional rules is particularly relevant here, as there is debate among scholars over whether disparate impact laws violate the equal protection clause and are thus unconstitutional (Primus, 2003).

deemed it to have an unjustified—and illegal—disparate impact. Importantly, the employer's motivation for instituting the policy was irrelevant to the Court's decision; even if enacted without discriminatory purpose, the policy was deemed discriminatory in its effects and hence illegal. Note, however, that disparate impact law does not prohibit *all* group differences produced by a policy—the law only prohibits *unjustified* disparities. For example, if, hypothetically, the high-school diploma requirement in *Griggs* were shown to be necessary for job success, the resulting disparities would be legal.

In modern applications of statistical risk assessments, discriminatory intent is often of secondary concern—indeed, many policymakers adopt algorithms in part to reduce bias in unaided human decisions. Instead, the primary question is whether algorithms inadvertently lead to discriminatory decisions, either through inappropriate design or by implicitly encoding biases in the data on which they are built. As such, our discussion of fairness below draws heavily on the economic concept of taste-based discrimination and its counterpart in the law, unjustified disparate impact.

2.2 Defining algorithmic fairness

To formally define measures of algorithmic fairness, we first introduce the notion of decision rules. Suppose we have a vector $x_i \in \mathbb{R}^p$ that we interpret as the visible attributes of individual i. For example, x might represent a loan applicant's age, gender, race, and credit history. We consider the problem of *fairly* selecting between one of two possible actions, a_0 and a_1. In the context of banking, a_0 may correspond to granting a loan application and a_1 to denying it; in the pretrial domain, a_0 may correspond to releasing a defendant awaiting trial and a_1 to detaining that individual. A *decision algorithm*, or a *decision rule*, is any function $d : \mathbb{R}^p \mapsto \{0, 1\}$, where $d(x) = k$ means that action a_k is taken.

We next present several additional assumptions and notational conventions that are helpful in stating and investigating common fairness definitions. First, we assume x can be partitioned into *protected* and *unprotected* features: $x = (x_p, x_u)$. For ease of exposition, we often imagine the protected features indicate an individual's race or gender, but they might also represent other attributes. Second, for each individual, we suppose there is a quantity $y \in \{0, 1\}$ that specifies the target of prediction. For example, in the pretrial setting, we might set $y_i = 1$ for those defendants who would have committed a violent crime if released, and $y_i = 0$ otherwise. Importantly, y is not known to the decision maker, who at the time of the decision has access only to information encoded in the visible features x. Third, we define random variables X and Y that take on values $X = x$ and $Y = y$ for an individual drawn randomly from the population of interest (e.g., the population of defendants for whom pretrial decisions must be made). We use X_p and X_u to denote the projections of x onto its protected and unprotected components. Fourth, we define the *true* risk function $r(x) = \Pr(Y = 1 \mid X = x)$. Finally, we note that many risk assessment algorithms, instead of simply outputting a decision a_0 or a_1, produce a risk score $s(x)$ that may be viewed as an approximation of the true risk $r(x)$. In reality, $s(x)$ may only be loosely related to the true risk, and $s(x)$ may not even lie in the interval $[0, 1]$ (e.g., $s(x) \in \{1, 2, \ldots, 10\}$ may represent a risk decile). To go from risk scores to decisions, it is common to simply threshold the score, setting $d(x) = 1$ if and only if $s(x) \geq t$ for some fixed threshold $t \in \mathbb{R}$.

With this setup, we now describe three popular definitions of algorithmic fairness.

Anti-classification. The first definition we consider is anti-classification, meaning that decisions do not consider protected attributes. Formally, anti-classification requires that:

$$d(x) = d(x') \text{ for all } x, x' \text{ such that } x_u = x'_u. \tag{1}$$

Some authors have suggested stronger notions of anti-classification that aim to guard against the use of unprotected traits that are proxies for protected attributes (Bonchi et al., 2017; Grgic-Hlaca et al., 2016; Johnson et al., 2016; Qureshi et al., 2016). We will demonstrate, however, that the exclusion of *any* information—including features that are explicitly protected—can lead to discriminatory decisions. As a result, it is sufficient for our purposes to consider the weak version of anti-classification articulated in Eq. (1).

Classification parity. The second definition of fairness we consider is classification parity, meaning that some given measure of classification error is equal across groups defined by the protected attributes.

In particular, we include in this definition any measure that can be computed from the two-by-two confusion matrix tabulating the joint distribution of decisions $d(x)$ and outcomes y for a group. Berk et al. (2017) enumerate seven such statistics, including false positive rate, false negative rate, precision, recall, and the proportion of decisions that are positive. We also include the area under the ROC curve (AUC), a popular measure among criminologists and practitioners examining the fairness of algorithms (Skeem and Lowenkamp, 2016).[4]

Two of the above measures—false positive rate, and the proportion of decisions that are positive—have received considerable attention in the machine learning community (Agarwal et al., 2018; Calders and Verwer, 2010; Chouldechova, 2017; Edwards and Storkey, 2015; Feldman et al., 2015; Hardt et al., 2016; Kamiran et al., 2013; Pedreshi et al., 2008; Zafar et al., 2015, 2017; Zemel et al., 2013). Formally, parity in the proportion of positive decisions, also known as *demographic parity* (Feldman et al., 2015), means that

$$\Pr(d(X) = 1 \mid X_p) = \Pr(d(X) = 1), \tag{2}$$

and parity of false positive rates means that

$$\Pr(d(X) = 1 \mid Y = 0,\, X_p) = \Pr(d(X) = 1 \mid Y = 0). \tag{3}$$

In our running pretrial example, demographic parity means that detention rates are equal across race groups; and parity of false positive rates means that among defendants who would not have gone on to commit a violent crime if released, detention rates are equal across race groups. Demographic parity is not strictly speaking a measure of "error", but we nonetheless include it under classification parity since it can be computed from a confusion matrix. We note that demographic parity is also closely related to anti-classification, since it requires that a classifier's predictions $d(X)$ be independent of protected group membership X_p.

Calibration. Finally, the third definition of fairness we consider is calibration, meaning that outcomes should be independent of protected attributes conditional on risk score. In the pretrial context, calibration means that among defendants with a given risk score, the proportion who would reoffend if released is the same across race groups. Formally, given risk scores $s(x)$, calibration is satisfied when

$$\Pr(Y = 1 \mid s(X),\, X_p) = \Pr(Y = 1 \mid s(X)). \tag{4}$$

Note that if $s(x) = r(x)$, then the risk scores trivially satisfy calibration.

2.3 Utility functions and threshold rules

When developing risk assessment tools in practice, it is common to first approximate the true risk $r(x)$ with a score $s(x) = \hat{r}(x)$, and then set $d(x) = 1$ if and only if $\hat{r}(x) \geq t$ for some fixed threshold t—we call these *threshold rules*. For example, in the banking context, one may deny loans to all applicants considered a high risk of default; and in the pretrial context, one may detain all defendants considered a high risk of committing a violent crime if released.

This strategy, while not explicitly referencing fairness, satisfies a compelling notion of equity, with all individuals evaluated according to the same standard.[5] When the threshold is chosen appropriately, threshold rules also satisfy the economic and legal concepts of fairness described above. To see this, we follow Corbett-Davies et al. (2017) and start by partitioning the aggregate costs and benefits of decisions for each individual.

[4]The AUC of risk scores $s(x)$ is defined as follows. Suppose X_+ is the feature vector for a random individual with label $y = 1$; for example, in the pretrial setting, X_+ is the feature vector for a random individual who ultimately recidivates. Similarly define X_- to be the feature vector for a random individual with label $y = 0$. Then the AUC of $s(x)$ is $\Pr(s(X_+) > s(X_-))$. In particular, given a random individual who ultimately recidivates and one who ultimately does not, the AUC is the probability that the one who recidivates is rated higher risk. Perfect risk scores would thus have 100% AUC, and completely uninformative risk scores would have 50% AUC.

[5]Threshold rules have received relatively little attention in the recent literature on fair machine learning. For example, this notion of equity was not included in a popular list of fairness definitions by Berk et al. (2017). Dwork et al. (2012) allude to a related concept by considering a constraint in which "similar individuals are treated similarly"; that work, however, does not operationalize similarity, nor does it explicitly consider statistical risk.

Let $b_{00} \geq 0$ be the benefit of taking action a_0 when $y = 0$, and let $b_{11} \geq 0$ be the benefit of taking action a_1 when $y = 1$. For example, in the pretrial domain, b_{00} is the value of releasing a defendant who would not recidivate if released, and b_{11} is the value of detaining a defendant who would recidivate if released. If we think of the decision as a binary prediction of the outcome, then b_{00} and b_{11} are the values of true negatives and true positives, respectively. Likewise denote by $c_{01} \geq 0$ the cost of taking action a_0 when $y = 1$, and denote by $c_{10} \geq 0$ the cost of taking action a_1 when $y = 0$. Then c_{01} and c_{10} can similarly be thought of as the costs of false negatives and false positives, respectively.

Now, since $r(x)$ is the probability that $y = 1$, the expected utility $u(0)$ of taking action a_0 is given by,

$$\begin{aligned} u(0) &= b_{00}(1 - r(x)) - c_{01}r(x) \\ &= b_{00} - (b_{00} + c_{01})r(x), \end{aligned}$$

and the expected utility $u(1)$ of taking action a_1 is given by,

$$\begin{aligned} u(1) &= -c_{10}(1 - r(x)) + b_{11}r(x) \\ &= -c_{10} + (c_{10} + b_{11})r(x). \end{aligned}$$

Note that $u(0)$ is a decreasing linear function of $r(x)$, and $u(1)$ is an increasing linear function of $r(x)$. Consequently, after rearranging terms, we have that $u(1) \geq u(0)$ if and only if

$$r(x) \geq \frac{b_{00} + c_{10}}{b_{00} + b_{11} + c_{01} + c_{10}}. \tag{5}$$

Thus, for a risk-neutral decision maker, the optimal action is a_1 when $r(x)$ is sufficiently large, and otherwise the optimal action is a_0. Further, if the costs and benefits above are the same for all individuals, and if $\hat{r}(x) = r(x)$, a threshold rule with $t = (b_{00} + c_{10})/(b_{00} + b_{11} + c_{01} + c_{10})$ produces optimal decisions for each person. As a result, threshold rules—under the utility framework we have laid out—ensure there is no taste-based discrimination or unjustified disparate impact.

In many domains, it is convenient to re-parameterize the utility above into more easily interpreted quantities. For example, in the pretrial context, the most salient costs are associated with the direct social and financial effects of detention, which we call c_{det}, and the most salient benefits result from crime prevented, which we call b_{crime}. Accordingly, the value b_{11} of a true positive is $b_{\text{crime}} - c_{\text{det}}$, and the cost c_{10} of a false positive is c_{det}. The value b_{00} of a true negative and the cost c_{01} of a false negative are both 0, since no one is detained—and hence no crime is prevented—in either case. By Eq. (5), $u(1) \geq u(0)$ when

$$\begin{aligned} r(x) &\geq \frac{b_{00} + c_{10}}{b_{00} + b_{11} + c_{01} + c_{10}} \\ &= c_{\text{det}}/b_{\text{crime}}. \end{aligned}$$

In particular, if the value of preventing a crime is twice the cost of detaining an individual, one would detain defendants who have at least a 50% chance of committing a crime. A threshold rule ensures that only the riskiest defendants are detained while optimally balancing the costs and benefits of incarceration.

Threshold rules are predicated on the assumption that errors are equally costly for all individuals. However, under a threshold rule, error *rates* generally differ across groups, violating classification parity. This disconnect highlights the counterintuitive nature of classification parity, an issue we discuss at length in Section 3. Threshold rules can also violate anti-classification, since the most statistically accurate risk scores $\hat{r}(x)$ may depend on group membership. These risk scores typically do satisfy calibration. However, we note that a risk score $s(x)$ that is calibrated need not approximate the true risk $r(x)$. To give an extreme example, $s(x) = 1 - r(x)$ is calibrated yet inversely related to risk, and would accordingly yield poor decisions.

In some cases, the costs and benefits of decisions are largely internalized within communities. For example, since the majority of violent crime is committed by someone known to the victim (Harrell, 2012), the benefits of detaining a high-risk defendant mostly accrue to members of the defendant's community.[6] Similarly, a borrower reaps the benefits of being issued a loan but also bears the costs of default—in terms of fees, penalties, and effects on future credit. Any spillover benefits (such as real-estate development raising

[6]There is likewise evidence that most violent crime is intra-racial (O'Brien, 1987).

the value of neighboring properties) or costs (such as foreclosure lowering neighboring property values) also mostly accrue to the borrower's community.[7] In these situations—and assuming the costs and benefits of decisions are approximately equal for all individuals—the threshold rule that maximizes aggregate social welfare also maximizes utility for each community considered in isolation.

The general approach outlined above is quite flexible, and can be used to approximate costs and benefits in a variety of situations. There are, however, several subtleties that one must consider in practice, a point we return to in Section 4. First, the costs and benefits of decisions might vary across individuals, complicating the analysis. For example, it may, hypothetically, be more socially costly to detain black defendants than white defendants, potentially justifying group-specific decision thresholds (Huq, 2019). Second, decision makers may not be risk neutral. A risk adverse decision maker might prefer the certainty of a lower expected payoff over one that is greater in expectation but uncertain. When uncertainty differs across individuals or groups, one might again reasonably deviate from a simple threshold rule. Third, resource constraints might mean we cannot take the higher-utility action for all individuals. For example, there may be more qualified borrowers than a lender has the capital to finance. In such circumstances, policymakers might decide to prioritize the well-being of certain groups over others. Finally, decisions might involve significant externalities, where the value of taking some action for an individual depends on what actions are taken for others. In this case, decisions cannot be considered in isolation, limiting the value of threshold rules. Despite these complications, threshold rules are often a natural starting point. More generally, a utility-based framework, in which one explicitly details the costs and benefits of actions, is a useful paradigm for reasoning about policy choices.

2.4 Applications of formal fairness criteria

Several authors have applied formal fairness criteria to evaluate existing decision algorithms. Perhaps most prominently, a team of investigative journalists at ProPublica reported that the popular COMPAS algorithm for recidivism prediction had higher false positive rates for black defendants than for whites—a finding widely interpreted as meaning the tool was biased against blacks (Angwin et al., 2016) (cf. Chouldechova, 2017, for discussion). The false positive rate metric has likewise been applied to assess algorithms for credit scoring (Hardt et al., 2016) and for child welfare services (Chouldechova et al., 2018). Others have examined whether algorithmic predictions in criminal justice have similar AUC across protected groups, a form of classification parity (Skeem and Lowenkamp, 2016), and whether such predictions are calibrated (Skeem et al., 2016).

Moving beyond diagnostics, a large body of work in computer science has proposed procedures for satisfying various fairness definitions. For example, Hardt et al. (2016) developed a method for constructing randomized decision rules that ensure true positive and false positive rates are equal across protected groups, a form of classification parity that they call *equalized odds*; they further study the case in which only false positive rates must be equal, which they call *equal opportunity*. Agarwal et al. (2018) similarly propose a technique for constructing algorithms that satisfy various forms of classification parity, including equality of false positive rates. Many papers, particularly early work in fair machine learning, proposed algorithms to achieve demographic parity via pre-processing, post-processing, and regularization techniques (Agarwal et al., 2018; Calders and Verwer, 2010; Edwards and Storkey, 2015; Feldman et al., 2015; Kamiran et al., 2013; Pedreshi et al., 2008; Zemel et al., 2013). Corbett-Davies et al. (2017) show that among all algorithms that satisfy demographic parity, the utility-maximizing rule uses different decision thresholds for each protected group; they similarly show that such multiple-threshold rules maximize utilty among algorithms that satisfy parity of false positive rates. Finally, several papers have suggested algorithms that enforce a broad notion of anti-classification, which prohibits not only the explicit use of protected traits but also the use of potentially suspect "proxy" variables (Grgic-Hlaca et al., 2016; Johnson et al., 2016; Qureshi et al., 2016). Recently, researchers have further expanded the idea of anti-classification to create algorithms that avoid potentially suspect *causal paths* (Bonchi et al., 2017; Datta et al., 2017; Kilbertus et al., 2017; Kusner et al., 2017; Nabi and Shpitser, 2018).

[7] A bank's optimal lending threshold might not coincide with the socially optimal threshold, necessitating a trade off between private and public utility (Liu et al., 2018). However, once one determines the appropriate trade-off, a threshold rule still maximizes utility.

Complicating these efforts to develop fair algorithms, several researchers have pointed out that many formal fairness definitions are incompatible (cf. Berk et al., 2017, who survey fairness measures and their incompatibilities). For example, Chouldechova (2017) shows that various classification parity constraints (specifically, equal positive/negative predictive values, and equal false positive/negative rates) are incompatible if base rates differ across groups. Kleinberg et al. (2017b) prove that except in degenerate cases, no algorithm can simultaneously satisfy calibration and a particular form of classification parity.[8] Corbett-Davies et al. (2016, 2017) and Pleiss et al. (2017) similarly consider the tension between calibration and alternative definitions of fairness.

This fast growing literature indicates the importance that many researchers place on formal, mathematical definitions of fairness, both as diagnostics to evaluate existing systems and as constraints when engineering new algorithms. Further, the impossibility results have been widely viewed as representing an unavoidable and unfortunate trade-off, with one desirable notion of fairness sacrificed to satisfy another, equally desirable one. However, as we discuss below, prevailing definitions of fairness typically do not map on to traditional social, economic or legal understandings of the concept. As a result, these formalizations are often ill-suited either as diagnostics or as design constraints. In particular, one can view the impossibility theorems as primarily identifying incompatibilities between various problematic fairness criteria, rather than as establishing more fundamental limits on fair machine learning.

3 Limitations of prevailing mathematical definitions of fairness

We now argue that the dominant mathematical formalizations of fairness—anti-classification, classification parity, and calibration—all suffer from significant limitations which, if not addressed, threaten to exacerbate the very problems of equity they seek to mitigate. We illustrate our points by drawing on a variety of statistical, legal, and economic principles, coupling our theoretical analysis with empirical examples.

3.1 Limitations of anti-classification

Perhaps the simplest approach to designing an ostensibly fair algorithm is to exclude protected characteristics from the statistical model. This strategy ensures that decisions have no explicit dependence on group membership. However, the history of the United States shows that clearly discriminatory behavior is possible even without using protected characteristics. For example, literacy tests were employed up until the 1960s as a facially race-neutral means of disenfranchising African Americans and others. This possibility has prompted many to argue that one should not only exclude protected attributes but also their "proxies". We note, however, that it is difficult to operationalize the definition of a "proxy", leading to a panoply of competing approaches. In part, this is because nearly every covariate commonly used in predictive models is at least partially correlated with protected group status; and in many situations, even strongly correlated covariates may be considered legitimate factors on which to base decisions (e.g., education in the case of hiring).[9] We circumvent this debate over proxies by arguing that there are important cases where even protected group membership itself should be explicitly taken into account to make equitable decisions. Once we establish that there is value to including protected traits themselves in risk models, the role of proxies largely becomes moot.

Consider our recurring example of pretrial recidivism predictions. After controlling for the typical "legitimate" risk factors, including criminal history, age, and substance use, women reoffend less often than men in many jurisdictions (DeMichele et al., 2018; Skeem et al., 2016). Consequently, gender-neutral risk assessments tend to overstate the recidivism risk of women. Figure 1 illustrates this phenomenon, plotting

[8] Specifically, Kleinberg et al. (2017b) show that, except in trivial cases, the following three properties cannot be simultaneously achieved: (1) calibration; (2) balance for the negative class, meaning that $\mathbb{E}[s(X) \mid Y = 0, X_p] = \mathbb{E}[s(X) \mid Y = 0]$; and (3) balance for the positive class, meaning that $\mathbb{E}[s(X) \mid Y = 1, X_p] = \mathbb{E}[s(X) \mid Y = 1]$.

[9] Legal definitions of proxies tend to focus on intent, where an unprotected trait x' is considered a proxy for a protected trait x_p if the decision maker intends for x' to be a replacement for x_p. Though such an intent-based definition can be useful in discrimination cases involving human decision makers, it is not as well-suited for evaluating algorithmic systems.

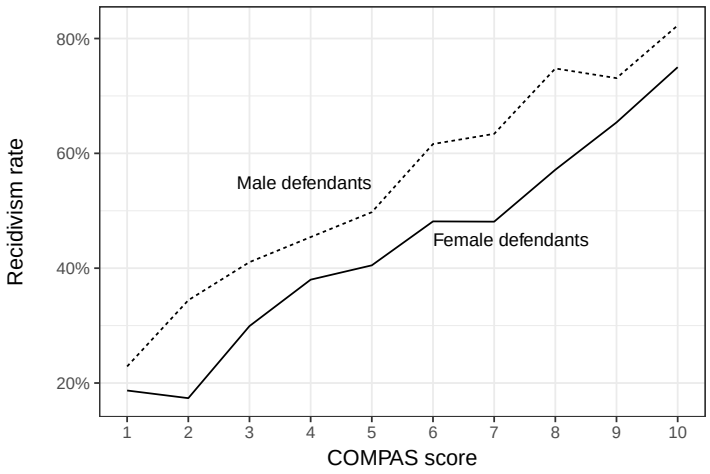

Figure 1: *The observed recidivism rate for men and women as a function of their (gender-neutral) COMPAS score, based on data from Broward County, Florida. Because women reoffend at lower rates than men with similar criminal histories, the gender-neutral COMPAS score overstates recidivism risk for women. If these risk estimates were used to make decisions, relatively low-risk women would be detained while higher-risk men were released. In this case, adhering to anti-classification would produce taste-based discrimination against female defendants.*

the observed recidivism rate for men and women in Broward County, Florida as a function of their gender-neutral COMPAS risk scores. In particular, women with a COMPAS score of seven recidivate less than 50% of the time, whereas men with the same score recidivate more than 60% of the time. Said differently, women with a score of seven recidivate about as often as men with a score of five, and this two-point differential persists across the range of scores. By acknowledging the predictive value of gender in this setting, one could create a decision rule that detains fewer people (particularly women) while achieving the same public safety benefits. Conversely, by ignoring this information and basing decisions solely on the gender-neutral risk assessments, one would be engaging in taste-based discrimination against women.

At least anecdotally, some judges recognize the inequities associated with gender-neutral risk scores, and mentally discount the stated risk for women when making pretrial decisions. Further, some jurisdictions, including the state of Wisconsin, combat this problem by using gender-*specific* risk assessment tools, ensuring that all judges have access to accurate risk estimates, regardless of an individual's gender. The Wisconsin Supreme Court recently affirmed this use of gender-specific risk assessment tools as "[promoting] accuracy that ultimately inures to the benefit of the justice system including defendants" (State v. Loomis, 2016).[10]

When gender or other protected traits add predictive value, excluding these attributes will in general lead to unjustified disparate impacts; when protected traits do not add predictive power, they can be safely removed from the algorithm. But we note one curiosity in the latter case. If protected attributes are not predictive, one could in theory build an accurate risk model using only examples from one particular group (e.g., white men). Given enough examples of white men, the model would learn the relationship between features and risk, which by our assumption would generalize to the entire population. This phenomenon highlights a tension in many informal discussions of fairness, with scholars advocating both for representative training data and for the exclusion of protected attributes. In reality, representative data are often most important precisely when protected attributes add information, in which case their use is arguably justified.[11]

[10]The case examined the use of gender-specific risk assessment tools for sentencing decisions, though a similar logic arguably applies in the pretrial setting. We also note that Loomis challenged the use of gender on due process grounds. The court did not rule on the possibility of an equal protection violation, although it did write that "if the inclusion of gender promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose."

[11]Representative data can help in two additional ways. First, a representative sample ensures that the full support

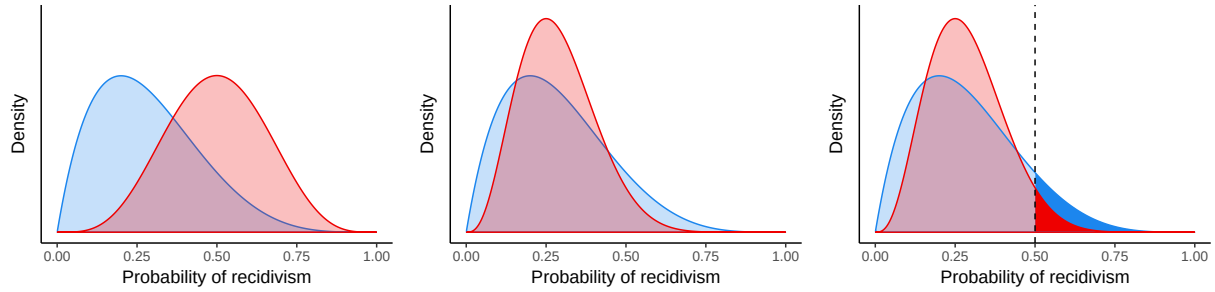

Figure 2: *Hypothetical risk distributions and a decision threshold (in the right-most plot). When risk distributions differ, infra-marginal statistics—like the precision and the false positive rate of a decision algorithm—also differ, illustrating the problem with requiring classification parity.*

3.2 Limitations of classification parity

Classification parity is not strictly a single fairness criterion, but rather a family of criteria that includes many of the most popular mathematical definitions of fairness. Starting with some aggregate measure of algorithmic performance, classification parity requires that measure be equal across groups. Commonly, the measures of performance are computed from an algorithm's group-specific confusion matrix. For example, as noted above, Angwin et al. (2016) found that the COMPAS risk assessment tool had a higher false positive rate for black defendants in Broward County, and had a higher false negative rate for white defendants. These differences led the authors to argue that COMPAS was biased against black defendants. In response, the developers of COMPAS countered that false positive and negative rates were flawed measures of fairness; they instead advocated for ensuring equality in positive and negative predictive values across race, which COMPAS achieved (Dieterich et al., 2016).[12] All four metrics are derived from COMPAS's confusion matrices for black and white defendants, so arguments that some combination of them should be equalized are all arguments for some type of classification parity. Here we demonstrate that in fact all such statistics computed from a confusion matrix are problematic measures of fairness.

Risk distributions. To understand the problems with classification parity, we must first understand risk distributions. Recall that the (true) risk function is defined by $r(x) = \Pr(Y = 1 \mid X = x)$. In the pretrial context, it is the probability that a defendant with observable features x will recidiviate if released. Let $\mathcal{D}_A = \mathcal{D}(r(X) \mid X \in A)$ denote the distribution of $r(X)$ over some subpopulation A. For example, $\mathcal{D}_A$ may be the distribution of true risk among white defendants, or alternatively, the distribution of true risk among black defendants. Figure 2 shows hypothetical risk distributions for two different groups. In the left-most plot, the two distributions have similar variances but different means; in the center plot, the distributions have the same mean but different variances. In general, we would expect any two groups—defined by race, gender, or any other attributes—to have risk distributions with different means and with different variances.

By the law of iterated expectations, the mean of $\mathcal{D}_A$ equals $\Pr(Y = 1 \mid X \in A)$. In the pretrial case, the

of features is present at training time. We note, though, that one might have adequate support even without a representative sample in many real-world settings, including criminal justice applications, particularly when models are trained on large datasets and the feature space is relatively low dimensional. Second, a representative sample can help with model validation, allowing one to assess the potential effects of group imbalance on model fit. In particular, without a representative sample, it can be difficult to determine whether a model trained on a single group generalizes to the entire population.

[12] Dieterich et al. (2016) also consider equality of AUC across race groups, a form of classification parity that they call "accuracy equity", which is approximately satisfied by COMPAS. We note that some authors use "calibrated" to describe *decisions* $d(X) \in \{0, 1\}$ that produce equal positive and negative predictive values (i.e., where $\Pr(Y = 1 \mid d(X), X_p) = \Pr(Y = 1 \mid d(X))$). Our definition of calibration—which is more common in the literature—applies to *scores*, not decisions, although the two definitions are equivalent if the score is binary and $s(X) = d(X)$. Thresholding a calibrated, multi-valued score will not, in general, result in a decision that produces equal positive and negative predictive values (see Section 3.3).

mean of a group's risk distribution is the group's overall rate of recidivism if all members of the group were released. As a result of this property, if two groups have different base rates, their risk distributions will necessarily differ, regardless of which features x are used to compute risk. The shape of the risk distribution about its mean depends entirely on the selection of features x; specifically, it depends on how well these features predict the outcome. Roughly, if risk is strongly related to the outcome, $\mathcal{D}_A$ will have most of its mass near zero and one (i.e., it will have high variance), while if the features are not particularly informative then most of the mass will lie near its mean.

As discussed in Section 2.3, the utility-maximizing decision rule is one that applies a threshold to the risk scores. In the pretrial setting, for example, one might detain all defendants above a certain risk level. Starting from any other decision algorithm, one could switch to a threshold rule and detain fewer defendants while achieving the same public safety benefits. In this sense, threshold rules are the unique rules that do not lead to unjustified disparate impacts, and hence satisfy strong legal and economic notions of fairness.

Given a risk distribution and a threshold, nearly every quantity of interest in algorithmic fairness can be computed, including all measures based on a confusion matrix, the area under the ROC curve (AUC), and the average risk of individuals for whom $Y = 0$ (i.e., the "generalized false positive rate" proposed by Kleinberg et al., 2017b).[13] For example, as illustrated in the right-most plot of Figure 2, the proportion of people labeled positive is the fraction of the distribution to the right of the threshold; and the precision of a classifier is the conditional mean of the distribution to the right of the threshold. Furthermore, since the risk distribution is fully defined by the choice of features and outcomes, these quantities are determined as soon as X, Y, and t are chosen. Importantly, Y and t are generally constrained by policy objectives, and X is typically constrained by the available data, so there is often little room for algorithm designers to alter the risk distribution.

The problem of infra-marginality. With this background, we now describe the limitations of classification parity. Because the risk distributions of protected groups will in general differ, threshold-based decisions will typically yield error metrics that also differ by group. This fact is visually apparent in Figure 2 (right-most plot) for one such metric, the precision of the classifier. Because the tails of the distributions differ across groups, one can visually see that the precision also differs across groups. The same discrepancy occurs for false positive rate—and for every other common error metric—though not all of these are as easily visualized.

This general phenomenon is known as the problem of *infra-marginality* in the economics and statistics literatures, and has long been known to plague tests of discrimination in human decisions (Ayres, 2002; Simoiu et al., 2017). In short, the most common legal and economic understandings of fairness are concerned with what happens at the *margin* (e.g., whether the same standard is applied to all individuals). What happens at the margin also determines whether decisions maximize social welfare, with the optimal threshold set at the point where the marginal benefits equal the marginal costs. However, popular error metrics assess behavior away from the margin, hence they are called *infra*-marginal statistics. As a result, when risk distributions differ, standard error metrics are often poor proxies for individual equity or social well-being.

To the extent that error metrics differ across groups, that tells us more about the shapes of the risk distributions than about the quality of decisions. In particular, it is hard to determine whether differences in error rates are due to discrimination or to differences in the risk distributions. This phenomenon is shown in Figure 3, which plots precision and false positive rate as a function of a group's base rate and the AUC of the group's risk scores. The figure was generated by assuming risk scores follow a discriminant distribution (Pierson et al., 2018), a type of logit-normal mixture that is well-suited to modeling risk distributions because it is naturally parameterized in terms of its mean and AUC. At a given threshold, both the precision and false positive rate increase with the base rate, with the false positive rate particularly sensitive to changes in the base rate. Further, for a fixed AUC and base rate, the false positive rate decreases in the threshold. As a result, a higher false positive rate for one group relative to another may either mean that the higher false positive rate group faces a lower threshold, indicating discrimination, or, alternatively, that the group has a higher base rate. Precision, in contrast, increases with the threshold, holding the AUC and base rate fixed. Accordingly, a lower precision for one group may either mean that the group faces a lower threshold

[13]The latter quantities—AUC and generalized false positive rate—depend only on the risk distribution and do not require a threshold.

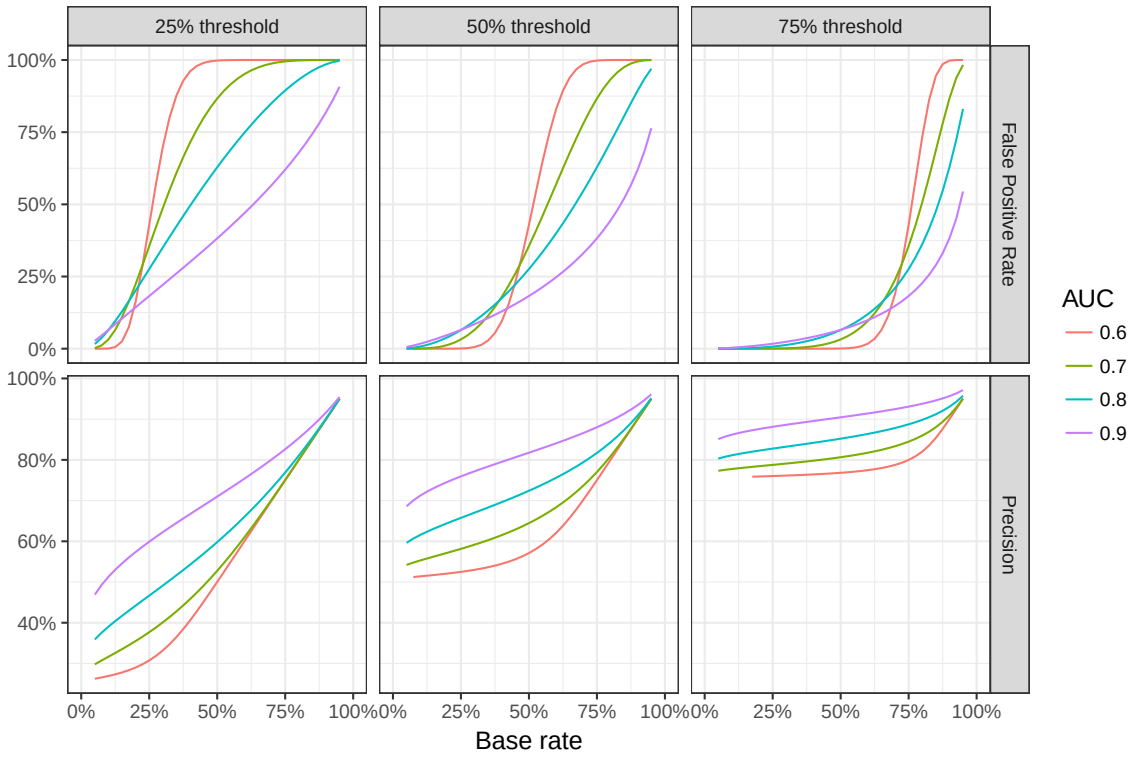

Figure 3: *The effect of different risk distributions on precision and false positive rate. At a given decision threshold, the metrics can differ dramatically depending on a group's risk distribution, parameterized here by base rate and AUC. This phenomenon illustrates the problem of infra-marginality, in which common error metrics are poor proxies for taste-based discrimination.*

or that the group has a lower base rate.

Infra-marginality in practice. The problem of infra-marginality is not merely a hypothetical possibility. Figure 4 shows estimated risk of violent recidivism for white and black defendants based on the Broward County COMPAS data (Larson et al., 2016). These distributions were generated by fitting an elastic net that predicts future arrests for violent offenses using all features in the dataset, including the original COMPAS risk scores designed to predict violent criminal activity. As in the stylized distributions of Figure 2, the empirical distributions plotted in Figure 4 differ considerably across groups. Consequently, threshold rules—which hold all individuals to the same standard—violate classification parity.

One might attribute differences in the estimated risk distributions in Figure 4 to problems with the statistical risk algorithm or with the dataset on which it was trained. But we caution against that conclusion. As noted above, when base rates of violent recidivism differ across groups, the true risk distributions will necessarily differ as well—and this difference persists regardless of which features are used in the prediction. The empirical data suggest that 21% of black defendants in Broward County are rearrested for violent offences, compared to 12% of white defendants. Though these rearrest rates are only approximations of the true, unobserved recidivism rate (a point we discuss in Section 4), they do suggest that there are indeed real differences between the two subpopulations. Of course, differences in recidivism rates are themselves a product of past social and economic discrimination. That fact, however, does not mean that statistical estimates of current, individual-level risk are inaccurate, or that better policy outcomes could be achieved by altering risk scores to satisfy classification parity. Policymakers may strive to reduce group differences, and they may debate the appropriate course of action to accomplish that goal, but we believe it would be misleading to characterize an algorithm or its training data as unfair for accurately identifying existing statistical patterns.

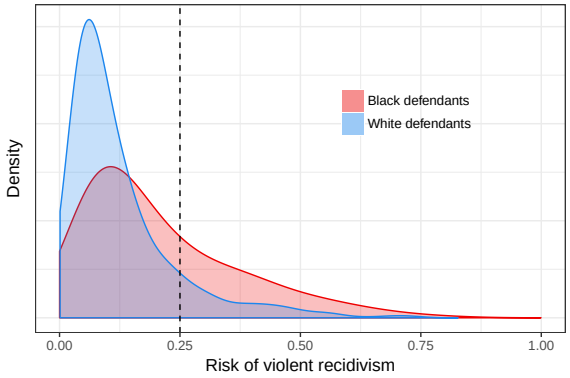

Figure 4: *Estimated distributions of violent recidivism risk for white and black defendants in the COMPAS dataset (Larson et al., 2016). Because the shapes of the risk distributions differ, threshold decisions necessarily mean that metrics like false positive rate also differ, illustrating the inherent problems with those metrics as measures of fairness.*

The effect of classification parity on group well-being. When risk distributions differ, enforcing classification parity can often decrease utility for all groups. Under relatively mild assumptions, Corbett-Davies et al. (2017) show that the optimal way to achieve classification parity is by setting different decision thresholds for different groups. Specifically, they show that among all decision rules that satisfy parity of false positive rates, utility (as defined in Section 2.3) is maximized by implementing group-specific decision thresholds. They similarly show that such multiple-threshold rules maximize utility among all algorithms satisfying demographic parity, though the optimal thresholds for satisfying demographic parity will in general differ from those necessary to optimally satisfy parity of false positive rates. Most importantly, the thresholds required to optimally satisfy these classification parity constraints will typically differ from the optimal thresholds for any community. Thus, requiring classification parity (or even approximate parity) can hurt majority and minority groups alike.

Consider, for example, the risk distributions from Broward County depicted in Figure 4, where we suppose that the vertical line at 25% is the utility-maximizing detention threshold (i.e., we suppose that $c_{\text{det}}/b_{\text{crime}} = 0.25$, meaning that society is willing to detain four individuals to prevent one additional violent crime). Then the utility-maximizing way to equalize false positive rates is by setting a 17% threshold for black defendants and a 31% threshold for white defendants. Likewise, the optimal way to achieve demographic parity is by setting a 16% threshold for black defendants and a 31% threshold for white defendants. In either case, whites face an overly strict detention threshold. Moreover, if the costs of releasing a high-risk defendant mostly fall on members of that defendant's community (e.g., when violent crime in a community is mostly committed by members of that community), then black communities experience harms due to an overly lenient detention threshold. In this example, members of both groups could be made better off by relaxing the requirement that the decisions satisfy classification parity. Importantly, this scenario is not a corner case designed to highlight the limitations of classification parity, nor is it a result of our assumption that the optimal threshold is 25%. When risk distributions differ, classification parity is typically costly to all groups, regardless of how society balances the relative costs of crime and detention.

Additional misconceptions about false positive rates. We conclude our discussion of classification parity by highlighting two popular misconceptions specific to equalizing false positive rates. First, one might believe that a difference in group-level false positive rates indicates an informational disparity. In our pretrial example, this view suggests that the higher observed false positive rate for black defendants relative to whites results from using less predictive features for blacks. Accordingly, some have argued that requiring parity of false positive rates creates an incentive for algorithm designers to collect better information on the higher error-rate group—black defendants in this case (Hardt et al., 2016).

While this argument is intuitively appealing, it is again important to consider the shape of risk distributions. As discussed above, risk distributions differ whenever base rates differ, regardless of the features

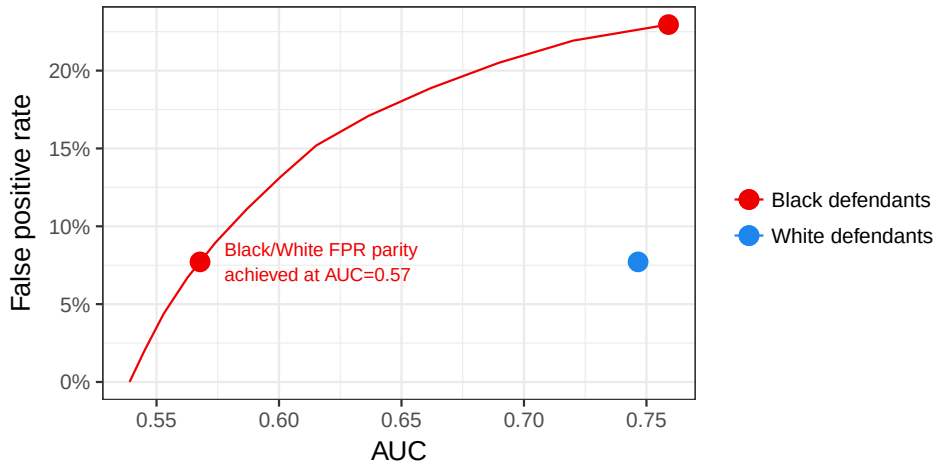

Figure 5: *False positive rates are a poor proxy of the predictive power of features, as measured by AUC. By ignoring observable differences between low and high-risk black defendants, one simultaneously lowers AUC and false positive rates. In this case, parity of false positive rates is achieved when the risk scores for black defendants are barely better than random, with an AUC of 0.57.*

collected, and so one would expect false positive rates to differ even when predictions are based on high-quality information.[14] Indeed, in the pretrial example in Figure 4, the AUC of predictions for black defendants (0.76) is slightly higher than the AUC for whites (0.75), despite the much higher false positive rate for black defendants.[15] Further, degrading predictions can, in certain circumstances, *lower* a group's false positive rate. This phenomenon is illustrated in Figure 5, which shows how false positive rate changes as we reduce the predictive quality of risk estimates for black defendants. Specifically, we degrade predictions by combining progressively larger subgroups of low-risk and high-risk black defendants, and then assigning them new risk scores equal to the average risk of the combined subgroups. This transformation is tantamount to ignoring information that could be used to tease apart the defendants. As a result, the risk distribution of black defendants becomes increasingly concentrated around its mean, lowering the proportion of defendants—both those who would have reoffended and those who would not have—above the decision threshold. (As in Figure 4, we fix the decision threshold at 25%.) We find that false positive rates are equalized when the black risk scores are barely better than random, with an AUC of 0.57. Importantly, those black defendants who are ultimately detained when we achieve false positive rate parity are not the riskiest ones—they are simply unlucky.

A second misconception is that the false positive rate is a reasonable proxy of a group's aggregate well-being, loosely defined. As above, however, this belief ignores the close relationship between false positive rates and risk distributions. Suppose, hypothetically, that prosecutors start enforcing low-level drug crimes that disproportionately involve black individuals, a policy that arguably hurts the black community. Further suppose that the newly arrested individuals have low risk of violent recidivism, and thus are released pending trial. This stylized policy change alters the risk distribution of black defendants, adding mass to the left-hand tail. As a result, the false positive rate for blacks would *decrease*. To see this, recall that the numerator of the false positive rate (the number of detained defendants who do not reoffend) remains unchanged while the denominator (the number of defendants who do not reoffend) increases. Without considering the distribution of risk—and in particular, the process that gave rise to that distribution—false positive rates can be a misleading measure of fairness.

[14] A classifier that perfectly predicts outcomes would have a false positive rate of zero, and one would thus achieve error-rate parity even if groups have different base rates. In general, though, when there are differences in base rates, there are also differences in false positive rates.

[15] Like the other error-rate metrics we have discussed, AUC is closely tied to the shape of a group's risk distribution, and we would thus expect AUC to in general differ across groups. Nevertheless, because AUC naturally adjusts for differences in base rates, it is arguably a reasonable—if imperfect—measure of aggregate predictive performance.

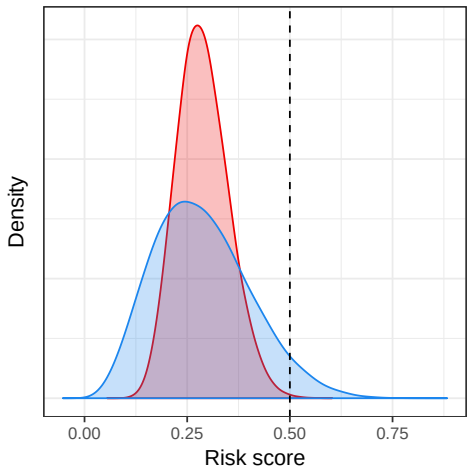

Figure 6: *Calibration is not sufficient to prevent discrimination. Starting from the true risk distribution (in blue), one can pair low-risk and high-risk individuals to produce new, calibrated scores (in red) that are concentrated around the group mean. In this manner, a decision maker can alter the proportion of individuals that lie above the decision threshold.*

3.3 Limitations of calibration

When criminologists and practitioners develop or audit a risk assessment tool, they typically check that risk scores are calibrated (Danner et al., 2016; DeMichele et al., 2018; Flores et al., 2016; Skeem and Lowenkamp, 2016). Calibration ensures that risk scores $s(x)$ mean the same thing for all protected groups—for example, that white and black defendants given a risk score of 7 indeed recidivate at comparable rates. Without this property, it is unclear how $s(x)$ can be said to quantify risk at all. However, while important, calibration is insufficient to ensure that risk scores are accurate or that decisions are equitable.

To see this, imagine a bank that wants to discriminate against black applicants. Further suppose that: (1) within zip code, white and black applicants have similar default rates; and (2) black applicants live in zip codes with relatively high default rates. Then the bank can surreptitiously discriminate against blacks by basing risk estimates only on an applicant's zip code, ignoring all other relevant information. Such scores would be calibrated (white and black applicants with the same score would default equally often), and the bank could use these scores to justify denying loans to nearly all black applicants. The bank, however, would be sacrificing profit by refusing loans to creditworthy black applicants,[16] and is thus engaged in taste-based discrimination. This discriminatory lending strategy is indeed closely related to the historical (and illegal) practice of redlining, and illustrates the limitations of calibration as a measure of fairness.

This redlining example can be generalized by directly altering a group's risk distribution. As in Section 3.2 above, we can aggregate low-risk and high-risk individuals and re-assign them risk scores equal to the group average. This procedure results in calibrated scores having a distribution that is concentrated around the group mean, as shown in Figure 6. Depending on where the mean lies relative to the decision threshold, this process can be used to change the number of individuals receiving positive or negative classifications. For example, assuming that the average minority bank applicant would not qualify for a loan, one could discriminate against minorities by altering the minority risk distribution to be concentrated around its mean, as in the redlining example above. Alternatively, assuming that the average white defendant is relatively low risk, one could discriminate against minorities by concentrating the white distribution around its mean, ensuring that no white defendants are detained. These examples illustrate the importance of considering all available data when constructing statistical risk estimates; assessments that either intentionally or inadvertently ignore predictive information may facilitate discriminatory decisions while satisfying calibration.

[16] These applicants are creditworthy in the sense that they would have been issued a loan had the bank used all the information it had available to determine their risk.

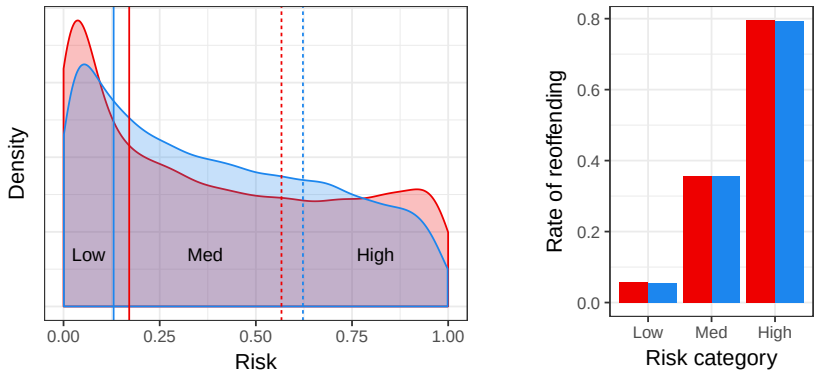

Figure 7: *Discriminating using calibrated, coarsened scores. The left panel shows the risk distributions for two groups, and the group-specific thresholds used to define the three risk categories. The right panel shows that these categories are calibrated. However, by detaining only defendants in the "high risk" category (those above the dotted thresholds in the left panel), the decision maker has successfully discriminated against the red group, who face a lower threshold than defendants from the blue group; in this case, there are red defendants who are detained while equally-risky blue defendants are released.*

In practice, it is common for risk assessments to be reported on a discrete scale (e.g., "low", "medium", or "high") rather than as real-valued probability estimates. Although this strategy may aid interpretation by human decision makers, it further complicates the role of calibration in assessing fairness. One sensible way to create discrete risk categories is to bin the underlying continuous probability estimates. Applying a threshold at any point on the discrete scale is then equivalent to thresholding the true risk, ensuring that no taste-based discrimination is present. However, when risk distributions differ and the number of categories is small, this strategy will typically not produce scores that are calibrated on the discrete scale—a phenomenon akin to the problem of infra-marginality.[17] In the extreme case—where risk scores are coarsened into just two categories, "low" and "high"—calibration on the discrete scale is equivalent to requiring a form of classification parity (specifically, both parity of positive predictive values and parity of negative predictive values), and is problematic for the same reasons. For example, Figure 7 shows how calibrated, discrete scores can mask taste-based discrimination, where different thresholds are used to bin individuals into risk categories. One must accordingly be careful when assessing the calibration of discretized risk scores.

4 Open challenges for designing equitable algorithms

We have thus far focused on the shortcomings of mathematical definitions of fairness, but it is equally important to identify a path forward, both for researchers and for policymakers. Unfortunately, there is no simple procedure or metric to ensure algorithmic decisions are fair. We can, however, enumerate some of the key principles and challenges for designing equitable algorithms. We specifically focus here on four broad issues: (1) measurement error; (2) sample bias; (3) model form, including model interpretability; and (4) externalities and equilibrium effects.

Measurement error. In our above discussion, we have assumed that algorithmic decisions are based on an individual's *true* risk $r(x) = \Pr(Y = 1 \mid X = x)$, a condition that implicitly requires that we have accurate measures of both y and x. We call measurement errors in these quantities *label bias* and *feature*

[17]Note, however, that the miscalibration produced by this strategy will never exceed one unit on the discrete scale; members of every group in a given risk category will always be riskier on average than members of any group in a lower risk category. Thus, COMPAS's 2-point gender miscalibration shown in Figure 1 is still problematic even though we should not in general expect perfect calibration when risk is converted into deciles.

bias, respectively, and address them both in turn below. We argue that label bias often poses significant challenges to constructing equitable risk scores, and indeed label bias is perhaps the most serious obstacle facing fair machine learning. Feature bias, however, can often be dealt with more easily in practice, though complications still remain.

In our running pretrial example, we take y to indicate whether a given defendant would commit a violent crime if released. But there are two key difficulties with this assumption. First, though we might want to measure violent crime conducted by defendants awaiting trial, we typically only observe crime that results in a conviction or an arrest. These observable outcomes, however, are imperfect proxies for the underlying criminal act. Further, heavier policing in minority neighborhoods might lead to black and Hispanic defendants being arrested, and later convicted, more often than whites who commit the same offense (Lum and Isaac, 2016). Poor outcome data might thus cause one to systematically underestimate the risk posed by white defendants. The second, related, issue is that y is a *counterfactual* outcome; it denotes what would have happened had a defendant been released. In reality, we only observe what happened conditional on the judge's actual detention decision.

There are no perfect solutions to the general problem of label bias. However, at least in certain applications, one can mitigate these tricky statistical issues. For example, criminologists have found that arrests for violent crime—as opposed to drug crime—may not in fact suffer from substantial racial bias.[18] In particular, Skeem and Lowenkamp (2016) note that the racial distribution of individuals arrested for violent offenses is in line with the racial distribution of offenders inferred from victim reports, and is also in line with self-reported offending data. In other cases, like lending, where one may seek to estimate default rates, the measured outcome (e.g., failure to pay) corresponds exactly to the event of interest. Even the problem of estimating counterfactuals can be partially addressed in many applications. In the pretrial setting, Angwin et al. (2016) measure recidivism rates in the first two-year period during which a defendant is not incarcerated; this is not identical to the desired counterfactual outcome—since the initial detention may be criminogenic, for example—but it seems like a reasonable estimation strategy. Further, unaided human decisions often exhibit considerable randomness, a fact that can be exploited to facilitate statistical estimation of counterfactual outcomes (Jung et al., 2017; Kleinberg et al., 2017a). More generally, a spate of recent work at the intersection of machine learning and causal inference (Hill, 2011; Jung et al., 2018; Mullainathan and Spiess, 2017) offers hope for more gains in counterfactual estimation.

We next turn to measurement errors in the predictors x. As alluded to above, minorities who commit drug crimes are more likely to be arrested than whites who commit the same offenses (Ramchand et al., 2006). Consequently, when past criminal behavior is inappropriately used to predict future activity, such feature bias can skew risk estimates. Suppose, for example, that true risk increases monotonically in the number of past drug sales one has *committed*, and further suppose that one uses arrests for past drug sales as a proxy for those committed transactions. Now consider a white and black defendant who have carried out the same number of drug sales, and who accordingly have similar true risk. Because the black defendant is likely to have had more drug arrests, he would (incorrectly) be rated higher risk than the white defendant, illustrating the problem of feature bias. Fortunately, in the absence of label bias (i.e., measurement error in y), such feature bias is statistically straightforward to address. Specifically, one can include group membership (e.g., race and gender) in the predictive model itself, or alternatively, one can fit separate risk models for each group. In this manner—and under the assumption that y is accurately measured—the statistical models would automatically learn to appropriately weight predictors according to group membership. For example, when predicting violent recidivism, a model might learn to down-weight past drug arrests for black defendants.

Such measurement error in x is closely related to what Ayres (2002) calls *subgroup validity*, the property that features are equally predictive across groups. However, subgroup validity is a more general phenomenon, as the relationship between x and y may plausibly differ across group even when predictors and labels are

[18] D'Alessio and Stolzenberg (2003) find evidence that white offenders are even somewhat more likely than black offenders to be arrested for certain categories of crime, including robbery, simple assault, and aggravated assault. Measurements of minor criminal activity, like drug offenses, are more problematic. For example, there is evidence that drug arrests in the United States are biased against black and Hispanic individuals, with minorities who commit drug crimes substantially more likely to be arrested than whites who commit the same offenses (Ramchand et al., 2006). Although this pattern is well known, many existing risk assessment tools still consider arrests or convictions for *any* new criminal activity—including drug crimes—which may lead to biased estimates.

accurately measured. For example, among defendants with (truly) similar past criminal behavior, men and women may in fact recidivate at different rates. As with feature bias, a simple statistical solution to this problem is to explicitly account for group membership when estimating risk. As discussed in Section 3.1, some jurisdictions indeed apply gender-specific risk models to address this issue. It is also common for algorithm designers to exclude features with differential predictive power (Danner et al., 2015). However, while perhaps a reasonable strategy in practice, we note that discarding information may inadvertently lead to the redlining effects mentioned in Section 3.3.

We note one final complication with measuring labels and features. In many settings, and one may be able to gather better data with greater investment of time and money. For example, recidivism predictions might be improved by collecting more comprehensive information on a defendant's criminal history, a process that is often costly and requires coordinating with multiple jurisdictions. In theory, this additional information may lead to welfare gains, and policymakers must accordingly evaluate the relative costs and benefits to all groups of exerting this extra effort when designing algorithms. In practice, there is often diminishing returns to information, with a relatively short list of key features providing most of the predictive power (Jung et al., 2017), at least partially mitigating this concern.

Sample bias. In addition to addressing measurement error, it is important to minimize sample bias when constructing risk scores. Ideally, one should train algorithms on datasets that are representative of the populations on which they are ultimately applied. Though this is often challenging in practice, failure to do so can lead to unintended, and potentially discriminatory, consequences. For example, Buolamwini and Gebru (2018) found that commercial facial analysis tools struggle to correctly classify the gender of dark-skinned individuals—and of dark-skinned women in particular—a disparity that is potentially attributable to the relative dearth of dark-skinned faces in two popular facial analysis datasets. Sample bias can similarly plague pretrial risk assessments, as it is often logistically challenging to develop tools that are tailored to local criminal justice populations. For example, the Ohio Risk Assessment System (ORAS) was developed on a sample of 452 Ohio defendants, but is now used in jurisdictions nationwide (Latessa et al., 2010). To the extent that Ohio defendants are not representative of those in other jurisdictions, ORAS may provide poor risk estimates. More recent criminal justice assessments address this shortcoming by training models on larger and more diverse samples. For example, the Public Safety Assessment (PSA) is based on 1.5 million cases from approximately 300 jurisdictions across the United States (Milgram et al., 2014). A further challenge is that risk assessments are often adopted as part of a broader bail reform effort that changes the distribution of defendants for whom pretrial decisions must be made. Therefore, even data from the same jurisdiction could suffer from sample bias if used to train a model to make predictions in the new regime.

As with measurement error, there is no complete solution to the problem of sample bias. In many settings, it may be prohibitively difficult to obtain representative data. For example, smaller jurisdictions may simply not have enough historical cases to train statistical models on local populations. Nevertheless, as in all situations, one must carefully weigh the potential costs and benefits of adopting a necessarily imperfect risk assessment tool relative to the other feasible options. In particular, even an imperfect algorithm may in some circumstances be better than leaving decisions to similarly imperfect humans who have their own biases. As we discuss briefly below, simple, transparent models can also mitigate some of these concerns with the training data.

Model form and interpretability. To create risk scores from a set of features x and labels y, one must faithfully map the relationship between x and y. When the feature space is low-dimensional and the training data are abundant, the precise statistical strategy chosen has little effect on the resulting estimates. Though this is an admittedly best-case scenario, it is not far from reality in some common risk assessment applications. For example, statistical models for estimating recidivism or default risk are often based on hundreds of thousands of examples and a small number of predictive features (e.g., the COMPAS risk model uses only six variables).[19] When the feature space is high-dimensional or the training data are less plentiful,

[19]Dressel and Farid (2018) write that COMPAS's risk assessment is based on 137 inputs. However, Equivant, the creator of COMPAS, states that their pretrial risk assessment algorithm uses only six features, writing that the additional information relates to "needs factors and are NOT used as predictors in the COMPAS risk assessment" (cf. `http://www.equivant.com/blog/official-response-to-science-advances`, emphasis in original). Other common

it becomes important to carefully consider the precise functional form of the statistical estimator, an ongoing challenge in supervised machine learning more broadly. One promising approach was recently proposed by Hashimoto et al. (2018), who describe a strategy for controlling the worst-case estimation error for arbitrary groups.

In the risk assessment community, there is a growing push to design statistical models that are simple, transparent, and explainable to domain experts. In traditional machine learning applications, researchers have often willingly accepted complexity in exchange for accuracy. This drive for predictive performance has resulted in algorithms capable of extraordinary feats, but ones that are also increasingly hard to understand. However, a risk assessment tool is only useful if it is adopted by practitioners; a complicated or opaque algorithm may engender mistrust from policymakers and other stakeholders, hindering implementation. Transparency can even become a legal requirement when society demands to know how algorithmic decisions are made (Goodman and Flaxman, 2016). Finally, simpler models may better transfer from one population to another by capturing general relationships rather than idiosyncratic patterns, partially alleviating concerns about sample bias in the training data. The newly active field of *interpretable machine learning* has already made significant strides, developing predictive algorithms that are both accurate and explainable (Doshi-Velez and Kim, 2017; Jung et al., 2017; Zeng et al., 2016), but there is still more work to do.

Externalities and equilibrium effects. Finally, we highlight the challenge of understanding fairness in more complex environments. Most of our discussion has ignored potential externalities and equilibrium effects, but there are important settings where these considerations come to the fore. For example, some decisions are better thought of as group rather than individual choices. In university admissions, a diverse student body may benefit the entire institution (Page, 2008), creating interdependencies between applicants. Predictive algorithms can also create feedback loops, leading to unintended consequences. As Lum and Isaac (2016) note, if police officers are deployed based on statistical predictions of criminal activity, that may entrench historical patrolling patterns, since crime is more likely to be recorded in locations were officers were previously stationed. Such a scenario can be viewed as a particularly pernicious form of label bias. In an employment scenario, Hu and Chen (2018) describe a dynamic, theoretical model where applying group-based thresholds incentivizes under-represented groups to invest in education and training, leading to a better long-term equilibrium in which group-based thresholds are no longer needed. Deploying an algorithm can also change the behavior of more distant actors in a complex system. For example, a pretrial risk assessment tool could change the upstream actions of officers and prosecutors who compensate for the expected outcomes of their decisions. Such potential equilibrium effects have implications for the overall utility of an algorithmic intervention; moreover, if these tools alter the populations to which they are applied, risk assessments may also become less accurate over time if they are not continually updated. While it is easy to enumerate such potential complications, it is admittedly difficult to quantify their effects or to translate such observations to actionable insights. Nevertheless, it is useful for researchers and policymakers to at least be aware of these broader issues when designing and deploying algorithmic systems.

5 Conclusion

From criminal justice to medicine, practitioners are increasingly turning to statistical risk assessments to help guide and improve human decisions. Algorithms can avoid many of the implicit and explicit biases of human decisions makers, but they can also exacerbate historical inequities if not developed with care. Policymakers, in response, have rightly demanded that these high-stakes decision systems be designed and audited to ensure outcomes are equitable. The research community has responded to the challenge, coalescing around three mathematical definitions of fairness: anti-classification, classification parity, and calibration. Over the last several years, dozens of papers have applied one or more of these measures either to audit existing systems or as constraints when developing new algorithms. However, as we have aimed to articulate, these popular measures of fairness suffer from significant statistical limitations. Indeed, enforcing anti-classification or classification parity can often harm the very groups that these measures were designed to protect.

pretrial risk assessments, like the Public Safety Assessment (PSA), similarly base their estimates on a small number of factors (Milgram et al., 2014).

How, then, should one design equitable algorithms? At a high-level, we first stress the importance of grounding technical and policy discussions of fairness in terms of real-world quantities. For example, in the pretrial domain, one might consider a risk assessment's short and long-term impacts on public safety and the size of the incarcerated population, as well as a tool's alignment with principles of due process. In lending, one could similarly consider a risk assessment's immediate and equilibrium effects on community development and the sustainability of a loan program. Formal mathematical measures of fairness only indirectly address such issues, and can inadvertently lead discussions astray. Of course, it is not always clear how best to quantify or to balance the relevant costs and benefits of proposed algorithmic interventions. In some cases, it may be possible to conduct randomized controlled trials; in other cases, the best one can do is hypothesize about an algorithm's potential effects. Regardless, we believe a more explicit focus on consequences is necessary to make progress.

Second, we recommend decoupling the statistical problem of risk assessment from the policy problem of designing interventions. At their best, statistical algorithms estimate the likelihood of events under different scenarios; they cannot dictate policy. An algorithm might (correctly) infer that a defendant has a 20% chance of committing a violent crime if released, but that fact does not, in and of itself, determine a course of action. For example, detention is not the only alternative to release, as one could take any number of rehabilitative interventions (Barabas et al., 2018). Even if detention is deemed an appropriate intervention, one must still determine what threshold would appropriately balance public safety with the social and financial costs of detention. One might even decide that society's goals are best achieved by setting different thresholds for different groups. For example, a policymaker might reason that, all else equal, the social costs of detaining a single parent are higher than detaining a defendant without children, and thus decide to apply different thresholds to the two groups. When policymakers consider these options and others, we believe the primary role of a risk assessment tool is, as its name suggests, to estimate risk. We note, however, that this view is at odds with requiring that algorithms satisfy popular fairness criteria, such as anti-classification and classification parity. Such constrained algorithms typically do not provide the best available estimates of risk, and thus implicitly conflate the statistical and policy problems.

The field of fair machine learning is still in its infancy, and there are several important avenues of research that could benefit from new statistical and computational insights. From mitigating measurement error and sample bias, to understanding the effects of externalities, to building interpretable models, there is much work to be done. But the benefits are equally large. When carefully designed and evaluated, statistical risk assessments have the potential to dramatically improve both the efficacy and equity of consequential decisions. As machine-learning algorithms are increasingly deployed in all walks of life, it will become ever more important to ensure they are fair.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. In *International Conference on Machine Learning*.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Arrow, K. et al. (1973). The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33.

Ayres, I. (2002). Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4(1-2):131–142.

Balkin, J. M. and Siegel, R. B. (2003). The american civil rights tradition: Anticlassification or antisubordination. *Issues in Legal Scholarship*, 2(1).

Barabas, C., Virza, M., Dinakar, K., Ito, J., and Zittrain, J. (2018). Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Cal. L. Rev.*, 104:671.

Becker, G. S. (1957). *The economics of discrimination*. University of Chicago Press, Chicago, IL.

Berk, R. (2012). *Criminal justice forecasts of risk: a machine learning approach*. Springer Science & Business Media.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2017). Fairness in criminal justice risk assessments: The state of the art. Working paper available at `https://arxiv.org/abs/1703.09207`.

Bertrand, M., Chugh, D., and Mullainathan, S. (2005). Implicit discrimination. *The American Economic Review*, 95(2):94–98.

Bonchi, F., Hajian, S., Mishra, B., and Ramazzotti, D. (2017). Exposing the probabilistic causal structure of discrimination. *International Journal of Data Science and Analytics*, 3(1):1–21.

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

Calders, T. and Verwer, S. (2010). Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292.

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

Chouldechova, A., Benavides-Prado, D., Fialko, O., and Vaithianathan, R. (2018). A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148.

Colker, R. (1986). Anti-subordination above all: Sex, race, and equal protection. *NYUL Rev.*, 61:1003.

Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. its actually not that clear. *Washington Post*.

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

D'Alessio, S. J. and Stolzenberg, L. (2003). Race and the probability of arrest. *Social forces*, 81(4):1381–1397.

Danner, M., VanNostrand, M., and Spruance, L. (2015). Risk-Based Pretrial Release Recommendation and Supervision Guidelines.

Danner, M. J., VanNostrand, M., and Spruance, L. M. (2016). Race and gender neutral pretrial risk assessment, release recommendations, and supervision: Vprai and praxis revised.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. (2017). Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*.

DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., and Misra, S. (2018). The public safety assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in kentucky.

Dieterich, W., Mendoza, C., and Brennan, T. (2016). Demonstrating accuracy equity and predictive parity performance of the compas risk scales in broward county.

Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1):eaao5580.

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*, pages 214–226. ACM.

Edwards, H. and Storkey, A. (2015). Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM.

Fisher v. University of Texas (2016). 579 U.S.

Fiss, O. M. (1976). Groups and the equal protection clause. *Philosophy & Public Affairs*, pages 107–177.

Flores, A. W., Bechtel, K., and Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *Fed. Probation*, 80:38.

Goodman, B. and Flaxman, S. (2016). European union regulations on algorithmic decision-making and a" right to explanation". *arXiv preprint arXiv:1606.08813*.

Grgic-Hlaca, N., Zafar, M. B., Gummadi, K. P., and Weller, A. (2016). The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law, Barcelona, Spain*, volume 8.

Griggs v. Duke Power Co. (1971). 401 U.S. 424.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances In Neural Information Processing Systems*, pages 3315–3323.

Harrell, E. (2012). *Violent victimization committed by strangers, 1993-2010*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Hashimoto, T. B., Srivastava, M., Namkoong, H., and Liang, P. (2018). Fairness without demographics in repeated loss minimization. *arXiv preprint arXiv:1806.08010*.

Hill, J. L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

Hu, L. and Chen, Y. (2018). A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1389–1398. International World Wide Web Conferences Steering Committee.

Huq, A. (2019). Racial equity in algorithmic criminal justice. *Duke Law Journal*, 68.

Johnson, K. D., Foster, D. P., and Stine, R. A. (2016). Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528*.

Jung, J., Concannon, C., Shroff, R., Goel, S., and Goldstein, D. G. (2017). Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690*.

Jung, J., Shroff, R., Feller, A., and Goel, S. (2018). Algorithmic decision making in the presence of unmeasured confounding. *arXiv preprint arXiv:1805.01868*.

Kamiran, F., Žliobaitė, I., and Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. (2017a). Human decisions and machine predictions. *The quarterly journal of economics*, 133(1):237–293.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2017b). Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*.

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

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

Latessa, E. J., Lemke, R., Makarios, M., and Smith, P. (2010). The creation and validation of the ohio risk assessment system (oras). *Fed. Probation*, 74:16.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2018). Delayed impact of fair machine learning. In *International Conference on Machine Learning*.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, 13(5):14–19.

Milgram, A., Holsinger, A. M., Vannostrand, M., and Alsdorf, M. W. (2014). Pretrial risk assessment: Improving public safety and fairness in pretrial decision making. *Fed. Sent'g Rep.*, 27:216.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Nabi, R. and Shpitser, I. (2018). Fair inference on outcomes. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2018, page 1931. NIH Public Access.

Nurse, A. (2014). Anti-subordination in the equal protection clause: A case study. *NYUL Rev.*, 89:293.

O'Brien, R. M. (1987). The interracial nature of violent crimes: A reexamination. *American Journal of Sociology*, 92(4):817–835.

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Page, S. E. (2008). The difference: How the power of diversity creates better groups, firms, schools, and societies. *Princeton University Press*.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM.

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, 62(4):659–661.

Pierson, E., Corbett-Davies, S., and Goel, S. (2018). Fast threshold tests for detecting discrimination. In *The 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.

Primus, R. A. (2003). Equal protection and disparate impact: Round three. *Harv. L. Rev.*, 117:494.

Qureshi, B., Kamiran, F., Karim, A., and Ruggieri, S. (2016). Causal discrimination discovery through propensity score analysis. *arXiv preprint arXiv:1608.03735*.

Ramchand, R., Pacula, R. L., and Iguchi, M. Y. (2006). Racial differences in marijuana-users risk of arrest in the united states. *Drug and alcohol dependence*, 84(3):264–272.

Shroff, R. (2017). Predictive analytics for city agencies: Lessons from children's services. *Big Data*, 5(3):189–196.

Siegel, R. B. (2003). Equality talk: Antisubordination and anticlassification values in constitutional struggles over brown. *Harv. L. Rev.*, 117:1470.

Simoiu, C., Corbett-Davies, S., and Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216.

Skeem, J., Monahan, J., and Lowenkamp, C. (2016). Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40(5):580.

Skeem, J. L. and Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, 54(4):680–712.

State v. Loomis (2016). 881 N.W.2d 749 (Wis. 2016).

Washington v. Davis (1976). 426 U.S. 229.

Winkler, A. (2006). Fatal in theory and strict in fact: An empirical analysis of strict scrutiny in the federal courts. *Vand. L. Rev.*, 59:793.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2015). Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International World Wide Web Conference*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.

Zeng, J., Ustun, B., and Rudin, C. (2016). Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.