ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Algorithmic Fairness: Choices, Assumptions, and Definitions

Shira Mitchell,[1] Eric Potash,[2] Solon Barocas,[3,4] Alexander D'Amour,[5] and Kristian Lum[6]

[1]Port Jefferson, New York 11777, USA

[2]Harris School of Public Policy, University of Chicago, Chicago, Illinois 60637, USA; email: epotash@uchicago.edu

[3]Microsoft Research, New York, NY 10012, USA

[4]Department of Information Science, Cornell University, Ithaca, New York 14853, USA; email: sbarocas@cornell.edu

[5]Google Research, Cambridge, Massachusetts 02124, USA; email: alexdamour@google.com

[6]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA; email: kl1@seas.upenn.edu

Annu. Rev. Stat. Appl. 2021. 8:141–63

First published as a Review in Advance on November 9, 2020

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-042720-125902

Keywords

algorithmic fairness, predictive modeling, statistical learning, machine learning, decision theory

Abstract

A recent wave of research has attempted to define fairness quantitatively. In particular, this work has explored what fairness might mean in the context of decisions based on the predictions of statistical and machine learning models. The rapid growth of this new field has led to wildly inconsistent motivations, terminology, and notation, presenting a serious challenge for cataloging and comparing definitions. This article attempts to bring much-needed order. First, we explicate the various choices and assumptions made—often implicitly—to justify the use of prediction-based decision-making. Next, we show how such choices and assumptions can raise fairness concerns and we present a notationally consistent catalog of fairness definitions from the literature. In doing so, we offer a concise reference for thinking through the choices, assumptions, and fairness considerations of prediction-based decision-making.

1. INTRODUCTION

Prediction-based decision-making has swept through industry and is quickly making its way into government. These techniques are already common in lending (Hardt et al. 2016, Liu et al. 2018, Fuster et al. 2020), hiring (Miller 2015a,b; Hu & Chen 2018a), and online advertising (Sweeney 2013), and they increasingly figure into decisions regarding pretrial detention (Angwin et al. 2016, Dieterich et al. 2016, Larson et al. 2016), immigration detention (Koulish 2016), child maltreatment screening (Vaithianathan et al. 2013, Chouldechova et al. 2018, Eubanks 2018), public health (Mayor's Off. Data Anal. 2018, Potash et al. 2015), and welfare eligibility (Eubanks 2018). Across these domains, decisions are based on predictions of an outcome deemed relevant to the decision. In recent years, attention has focused on how consequential predictive models may be biased—a now overloaded word that, in popular media, has come to mean that the model's performance (however defined) unjustifiably differs along social axes such as race, gender, and class. Uncovering and rectifying such biases in statistical and machine learning models has motivated a field of research we call algorithmic fairness.

Algorithmic fairness has been explored in popular books (O'Neil 2016, Eubanks 2018) and an influential White House report (Exec. Off. Pres. 2016), has been surveyed in technical review articles (Berk et al. 2017, Chouldechova & Roth 2018, Friedler et al. 2018, Verma & Rubin 2018) and an in-progress textbook (Barocas et al. 2018), and has inspired a number of software packages (Galhotra et al. 2017, Zehlike et al. 2017, Angell et al. 2018, Wexler 2018).

Though the algorithmic fairness conversation is somewhat new, it resembles older work. For example, since the 1960s, psychometricians have studied the fairness of educational tests based on their ability to predict performance (at school or work) (Cleary 1966, Darlington 1971, Einhorn & Bass 1971, Thorndike 1971, Hunter & Schmidt 1976, Petersen & Novick 1976, Lewis 1978, Hutchinson & Mitchell 2019). More recently, Dorans & Cook (2016) reviewed broader notions of fairness, including in test design and administration.

The goal of this article is not to advance axiomatic definitions of fairness but to summarize the definitions and results in this area that have been formalized to date. Our hope is that this article contributes a concise, cautious, and reasonably comprehensive catalog of the important fairness-relevant choices that are made in developing a predictive model; the assumptions that underlie many models where fairness is a concern; and the wide range of metrics and methods that have been proposed for evaluating the fairness of models. Alongside this summary, we point out gaps between mathematically convenient formalism and the larger social goals that many of these concepts were introduced to address.

Throughout this article, we ground our theoretical and conceptual discussion of algorithmic fairness in real-world example cases that are prevalent in the literature: pretrial risk assessment and lending models. In a typical pretrial risk assessment model, information about a person who has been arrested is used to predict whether they will commit a (violent) crime or whether they will fail to appear for court if released. These predictions are often based on an individual's demographic information, their criminal history, and sometimes their responses to more in-depth interview questions. These predictions are then used to inform a judge, who must make an extremely consequential decision: the conditions under which the person should be released from jail, if at all. Although the options available to judges are many (including simply releasing the person, setting bail, requiring participation in a supervised release program, etc.), in this literature, this decision is often reduced to a binary decision—release or detain.

In lending, the task is to predict whether an individual will repay a loan if one is granted. These predictions are based on employment, credit history, and other covariates. The loan officer then incorporates this prediction into their decision-making to decide whether the applicant should

be granted a loan and, if so, under what terms. In the algorithmic fairness literature, the decision space is often reduced to the decision to grant or deny the loan.

The article proceeds as follows. Section 2 outlines common choices and assumptions that are made that are often considered outside the scope of the model but have material consequences for the fairness of the model in practice. In Section 3, as a segue to the slightly more mathematical parts of the article, we introduce our main setup and notation. Section 4 begins the discussion of the various mathematical notions of fairness that have been introduced, including tensions among them. In Section 5, we explore causal frameworks for reasoning about algorithmic fairness. Section 6 offers some suggestions and ideas for future work in this area, and Section 7 concludes.

2. CHOICES, ASSUMPTIONS, AND CONSIDERATIONS

Several recent papers have demonstrated how policy goals are, sometimes clumsily, abstracted and formulated to fit into a prediction task (Dobbe et al. 2018, Eckhouse et al. 2018, Green 2018, Green & Hu 2018, Ochigame et al. 2018, Silva & Kenney 2018, Passi & Barocas 2019, Selbst et al. 2019). In this section, we link these concerns to choices and assumptions made in the policy design process. Broadly, these assumptions take the problem of evaluating the desirability of a policy and reduce it to the simpler problem of evaluating the characteristics of a model that predicts a single outcome.

2.1. The Policy Question

Much of the technical discussion in algorithmic fairness takes as given the social objective of deploying a model, the set of individuals subject to the decision, and the decision space available to decision-makers who will interact with the model's predictions. Each of these is a choice that—although sometimes prescribed by policies or people external to the immediate model building process—is fundamental to whether the model will ultimately advance fairness in society, however defined.

2.1.1. The overarching goal. Models for which fairness is a concern are typically deployed in the service of achieving some larger goal. For a benevolent social planner, this may be some notion of justice or social welfare (Hu & Chen 2018b). For a criminal justice actor, this goal may be reducing the number of people who are detained pending trial while simultaneously protecting public safety. For a bank making lending decisions, the goal may be maximizing profits. Often, different stakeholders have genuinely different and conflicting goals, which cannot be resolved by more data or different modeling choices (Eubanks 2018, O'Neil & Gunn 2020). If one disagrees with the overarching goal, a model that successfully advances this goal—regardless of whether it attains any of the mathematical notions of fairness discussed here—will not be acceptable.

Prediction-based decision systems often implicitly assume the pursuit of the overarching social goal will be served by better predicting some small number of outcomes. For example, in pretrial decisions, outcomes of interest are typically crime (measured as arrest or arrest for a violent crime) and nonappearance in court. In contrast, human decision-makers may consider several outcomes, including impacts on a defendant's well-being or the well-being of a defendant's dependents (Brownsberger 2017). In making decisions about college admissions, it may be tempting to narrow the larger goals of higher education to simply maximizing the future grade point averages of admitted students (Kleinberg et al. 2018). Narrowing focus to a chosen, measured outcome can fall short of larger goals [a problem sometimes called omitted payoff bias (Kleinberg et al. 2017)].

Furthermore, prediction-based decision systems usually only focus on outcomes under one decision (e.g., crime if released) and assume the outcome under an alternative decision is known (e.g., crime if detained). More recent work has formalized this oversight using the language of counterfactuals and potential outcomes (Coston et al. 2020). Finally, prediction-based decisions are formulated by assuming that progress toward the overarching goal can be expressed as a scalar utility function that depends only on decisions and outcomes (see Section 3).

2.1.2. The population. A model's predictions are not applied to all people, but typically to a specific subpopulation. In some cases, individuals choose to enter this population; in other cases, they do not. In pretrial decisions, the population is people who have been arrested. In lending decisions, the population is loan applicants. These populations are sampled from a larger population by some mechanism—for example, people are arrested because a police officer determines that the individual's observed or reported behavior is sufficiently unlawful to warrant an arrest, or creditors target potential applicants with offers or applicants independently decide to apply to a particular lending company for a loan. The mechanism of entry into this population may reflect objectionable social structures, e.g., policing that targets racial minorities for arrest (Alexander 2012) and discrimination in loan preapplication screening (Courchane et al. 2000). A model that satisfies fairness criteria when evaluated only on the population to which the model is applied may overlook unfairness in the process by which individuals came to be subject to the model in the first place.

2.1.3. The decision space. The decision space is the set of actions available to a decision maker. For example, in a simplified pretrial context, the decision space might consist of three options: release the arrested person on recognizance, set bail that must be paid to secure the individual's release, or detain the individual. In the lending example, the decision space might only consist of the options to grant or deny the loan application. Both of these decision spaces leave out many other possible interventions or options. In lending, a broader decision space could include offering different interest rates and loan terms. While mathematical definitions in the algorithmic fairness literature discussed below may be able to certify the fair allocation of decisions across a population, they have nothing to say about whether any of the available actions are acceptable in the first place.

2.2. The Statistical Learning Problem

Mathematical definitions of fairness generally treat the statistical learning problem that is used to formulate a predictive model as external to the fairness evaluation. Here, too, there are a number of choices that can have larger social implications but go unmeasured by the fairness evaluation. We focus specifically on the choices of training data, model, and predictive evaluation.

2.2.1. The data. The foundation of a predictive model is the data on which it is trained. The data available for any consequential prediction task, especially data measuring and categorizing people, can induce undesirable properties when used as a basis for decision-making. In the algorithmic fairness literature, data with these undesirable properties are often labeled informally as biased (e.g., Kamiran & Calders 2009, Barocas & Selbst 2016, Chouldechova 2017, Lipton et al. 2018). Here, we decompose this notion of bias into more precise notions of statistical bias (i.e., concerns about nonrepresentative sampling and measurement error) and societal bias (i.e., concerns about objectionable social structures and past injustice that are represented in the data) (**Figure 1**). We treat each of these notions in turn.

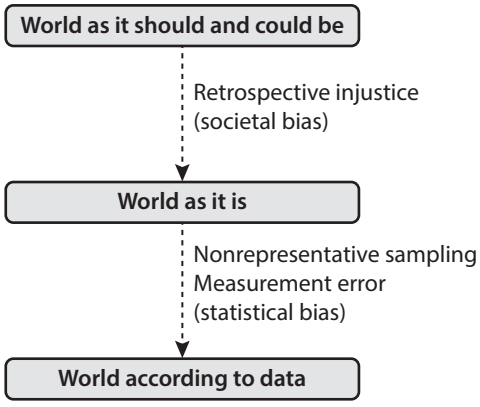

Figure 1

A schematic showing two components of biased data: societal bias and statistical bias.

2.2.1.1. Statistical bias. Here we consider statistical bias to be a systematic mismatch between the sample used to train a predictive model and the world as it currently is. Specifically, we consider how sampling bias and measurement error can induce fairness implications that are usually unmeasured by mathematical fairness definitions.

Sampling bias occurs when a data set is not representative of the full population to which the resulting model will be applied. For example, in pretrial decisions, the sample used to train the model is typically drawn from the population of people who were released pretrial. For people who were not released before their case has concluded, it is not possible to directly measure typical prediction outcomes, such as rearrest, as they do not have the opportunity to be rearrested. These people are typically excluded from training data sets. In the lending example, models may be based only on those individuals who were granted a loan, as it is not possible to measure the counterfactual outcome, i.e., whether a person who was not granted a loan would have repaid it had it been granted. This problem is sometimes called selective labels (Lakkaraju et al. 2017). Incorrectly assuming that a sample is representative can lead to biased estimation of conditional probabilities (e.g., probability of crime given covariates), biased estimation of utility, and inadequate fairness adjustments (Kallus & Zhou 2018).

Under a missing at random assumption, modeling could hope to avoid this selection bias (Gelman et al. 2013). But there can be regions of the covariates where no data exist to fit a model; for example, there may be values of a covariate for which no defendants were ever released, and hence the outcome in that region is unobserved (Chouldechova & Roth 2018). One option is extrapolation: Fit the model to released defendants, then apply the model to all defendants even if this includes new regions of the variables. However, models can perform poorly with such a shift in covariates (Rabanser et al. 2018). Another option is to trust previous decisions: Assume that regions of the covariates where no defendants were released are regions where all defendants would have committed crimes if released (De-Arteaga et al. 2018). Though convenient from a modeling perspective, these types of assumptions are unlikely to hold.

Similarly, systematic measurement error can have profound consequences. This is particularly true when the error is greater for some groups than others, which is known as differential measurement error (VanderWeele & Hernán 2012). In lending, some measures of past success in loan repayment (measures that may be used to predict future loan repayment) only account for repayment of loans through formal banking institutions. At least historically, immigrant communities were more likely to engage in more informal lending (Day 2002). Measurements of past success in

loan repayment may then systematically understate the value of past loans repaid for people who participate in informal lending relative to those who go through formal channels. Similar issues exist in our running example of pretrial risk assessment, where commission of a crime is often measured by rearrest during the pretrial period. Individuals and groups that are more likely to be rearrested following the commission of a crime will be measured in the data as more criminal, regardless of whether those differences are the result of true underlying rates of the commission of the crime or bias in the process by which the decision to arrest is made.

Additionally, the perceived fairness of a model may hinge on measurement choices that incorporate moral or normative considerations. For example, in one author's experience, recent updates to a pretrial risk assessment tool have incorporated changes such that past failures to appear for court appointments have a sunset window for inclusion in a model. In the name of fairness, after the sunset window, they can no longer be counted against the defendant. In lending, it is already the case that bankruptcy is erased from one's credit report after 7 years (chapter 13 bankruptcy) or 10 years (chapter 7 bankruptcy). These measurement considerations have less to do with the accuracy or bias of the measurement and more to do with normative decisions about how a person ought to be evaluated.

2.2.1.2. Societal bias. Even if training data are representative and accurate, they may still record objectionable aspects of the world that, if encoded in a policy, run counter to the decision-maker's goals. This corresponds to a nonstatistical notion of societal bias (Suresh & Guttag 2019). For example, in the lending setting, one could compile representative and accurate information on loan applicants' income. This, by definition, would not suffer from statistical bias. However, it could suffer from societal bias if some people (e.g., women) systematically received lower wages. Though the data may accurately reflect the reality of compensation, including data that encode such differences may inadvertently perpetuate them.

In addition, there can be a connection between the two concepts. For example, societal bias that leads to the overpolicing of people of color can cause statistical bias in arrest data if they are used as a measure of crime (Alexander 2012, Lum & Isaac 2016). But suppose we could perfectly measure crime: Does this make the data free from bias? In a statistical sense, the data are free from bias, but in a normative sense, they are not—crime rates reflect unequal social structures, and societal bias arises even in how crime is defined (Raphling 2018).[1]

In general, addressing issues of societal bias may require adjusting data collection processes or manually incorporating an understanding of this bias into the model building process, or these problems may not have technical solutions at all.

2.2.2. The predictive model. Statistical and machine learning models are designed to identify patterns in the data used to train them. As such, they will reproduce, to the extent that they are predictive, unfair patterns encoded in the data or unfairness in the problem formulation as described above.

Furthermore, there are many choices for building a model. Indeed, this is the subject of much of statistics and machine learning research. Choices include the class of model, the functional form, and model parameters, among others. These choices, too, have ramifications for fairness. For example, it has been shown that different functional forms can differ in their estimates for individuals and disparities between groups of individuals (Chouldechova & G'Sell 2017).

[1] In statistics, bias generally refers to properties of an estimator, not data. Here, we mean bias in the estimation of conditional probabilities or fairness metrics that could result from nonrepresentative data, measurement error, or model misspecification.

Models also differ in their interpretability. For example, some use point systems, where the presence and absence of characteristics are assigned integer point values. The individual's final score or prediction is then the sum of those integer-valued points. These and other simple models can help humans understand the basis upon which a model's predictions are being made and facilitate normative debate about the legitimacy of that basis, e.g., revealing how a risk assessment model makes use of potentially contentious covariates like a defendant's financial, employment, or family situation. These techniques can also clarify how perturbations to the inputs might impact the model's predictions, e.g., providing stakeholders in the criminal justice system with insights into the differences that make a difference to the pretrial detention decision. The disclosure of such details is often essential to ensuring the perceived fairness of the decision-making process and is a crucial requirement of due process. This can also have a material effect on the quality of decisions by helping stakeholders to foresee and account for cases in which a model might produce counterintuitive or nonsensical results—and thus help stakeholders avoid or catch errors (Rudin 2019).

Finally, there is the choice of which covariates to include in the model. Outcome predictions can change depending on what we condition on. Person A can have a higher predicted value than person B with a choice of covariates V but a lower predicted value with a choice of covariates V'.

2.2.3. Evaluation. Fairness concerns aside, predictive models are typically built, evaluated, and selected using various measures of their predictive performance such as (for continuous predictions) mean squared error or (for binary predictions) positive-predictive value, or sensitivity, specificity, and so on.

We note that such evaluations generally make three important assumptions. First, they assume that decisions can be evaluated as an aggregation of separately evaluated individual decisions. This includes assuming that outcomes are not affected by the decisions for others, an assumption known as no interference (Imbens & Rubin 2015). In the loan setting, denying one family member a loan may directly impact another family member's ability to repay their own loan. If both are evaluated by the same model, this dependence is in conflict with the assumption that they can be accurately evaluated separately.

This assumption resembles beliefs from utilitarianism, which represents social welfare as a sum of individual utilities (Rawls 1971). With binary predictions and decisions, each individual decision's utility can in turn be expressed in terms of four utilities for each possibility: true positive, false positive, false negative, and true negative.

Second, these evaluations assume that all individuals can be considered symmetrically, i.e., identically. This assumes, for example, that the harm of denying a loan to someone who could repay is equal across people. Denying someone a loan for education will likely have a very different impact on their life than denying a different person a similarly sized loan for a vacation home.

Third, these evaluations assume that decisions are evaluated simultaneously. That is, they are evaluated in a batch as opposed to serially and so do not consider potentially important temporal dynamics (Chouldechova & Roth 2018). For example, in the context of predictive policing, Harcourt (2008) shows that if the population changes their behavior in response to changing decision probabilities (i.e., the likelihood of being subject to policing), prediction-based decisions can backfire and diminish social welfare.

A fundamental question of algorithmic fairness is to what extent prediction measures making these assumptions are relevant to fairness.

2.3. Axes of Fairness and Protected Groups

A final choice in mathematical formulations of fairness is the axis (or axes) along which fairness is measured. For example, much of the algorithmic fairness literature considers the simple case of

two groups (advantaged and disadvantaged). In this setting, typically only one variable is selected as a sensitive attribute, which defines the mapping to the advantaged and disadvantaged groups. Deciding how attributes map individuals to these groups is important and highly context specific. Does race determine advantaged group membership? Does gender? In regulated domains such as employment, credit, and housing, these so-called protected characteristics are specified in the relevant US discrimination laws. Even in the absence of formal regulation, though, certain attributes might be viewed as sensitive, given specific histories of oppression, the task at hand, and the context of a model's use. Cultural context will also often dictate how any given sensitive attribute is thought to carve up the population, e.g., the relevant ethnic categories in society.

Differential treatment or outcome by race is often concerning, even when it does not violate a specific law, but which racial groups are salient will often vary by cultural context. Though most algorithmic fairness work considers only one sensitive attribute at a time, discrimination might affect members at the intersection of two groups, even if neither group experiences discrimination in isolation. This fact was most famously highlighted by the influential work of Crenshaw (1989), who analyzed failed employment discrimination lawsuits involving black women. She revealed that black women were unable to establish discrimination simply as sex discrimination (because the discrimination they experienced did not apply to white women) or as race discrimination (because it did not apply to black men). More recently, the importance of considering combinations of attributes to define advantage and disadvantage in algorithmic fairness applications is shown by Buolamwini & Gebru (2018), who evaluated commercial gender classification systems and found that darker-skinned females are the most misclassified group.

3. SETUP AND NOTATION

We now turn to mathematical formulations of fairness, having presented a number of key choices, assumptions, and consideration that make this abstraction possible. Here, we introduce notation for some canonical problem formulations considered in the algorithmic fairness literature. We follow much of the recent discourse in this literature and focus on fairness in the context of binary decisions that are made on the basis of predictions of binary outcomes.

We consider a population about whom we want to make decisions. We index people by $i = 1, \ldots, n$. We assume this finite population (of size n) is large enough to approximate the superpopulation distribution from which they were drawn and refer to both as the population. Each person has covariates (i.e., features, variables, or inputs) $v_i \in \mathcal{V}$ that are known at decision time. In some cases, we can separate these into sensitive variable(s) a_i (e.g., race, gender, or class) and other variables x_i, writing $v_i = (a_i, x_i)$. We denote the person's outcome by y_i.

A binary decision d_i is made for each person. We restrict decisions to be functions of variables known at decision time, $\delta : \mathcal{V} \to \{0, 1\}$, where $d_i = \delta(v_i)$. We define random variables V, Y, $D = \delta(V)$ as the values of a person randomly drawn from the population.

In prediction-based decisions, decisions are made based on a prediction of an outcome, y_i, that is unknown at decision time. Specifically, decisions are made by first estimating the conditional probability

$$\mathrm{P}[Y = 1 | V = v_i].$$

The decision system does not know the true conditional probability; instead it uses $\psi : \mathcal{V} \to [0, 1]$ where $s_i = \psi(v_i)$ is a score intended to estimate $\mathrm{P}[Y = 1 | V = v_i]$. Let $S = \psi(V)$ be a random score from the population. A prediction-based decision system then has a decision function δ that is a function of the score alone, i.e., $\delta(v) = f(\psi(v))$ for some function f. A common choice for f is an indicator function of whether $\psi(v)$ is greater than some threshold. Both ψ and δ are functions of a sample of $\{(v_i, y_i)\}$ that we hope resembles the population.

For example, in pretrial risk assessment, we predict whether an individual i will be rearrested ($y_i = 1$) or not ($y_i = 0$) using v_i, summaries of criminal history and basic demographic information. A decision is then made to detain ($d_i = 1$) or release ($d_i = 0$) the individual on the basis of the prediction. In the lending setting, the goal is to predict whether an individual will default on ($y_i = 0$) or repay ($y_i = 1$) the loan as a function of v_i, their credit history. The decision space consists only of the decision to deny ($d_i = 0$) or grant ($d_i = 1$) the loan. In both examples, we choose notation such that when $y_i = d_i$, the "correct" decision has been made.

4. FLAVORS OF FAIRNESS DEFINITIONS FROM DATA ALONE

We begin our exposition of formal fairness definitions with so-called oblivious definitions of fairness that depend on the observed data alone (Hardt et al. 2016). These definitions equate fairness with certain parities that can be derived from the distributions of the observed features V, outcomes Y, scores S, and decisions D, without reference to additional structure or context. These stand in contrast to nonoblivious fairness definitions, presented in Section 5.

4.1. Unconstrained Utility Maximization and Single-Threshold Fairness

As a default, we consider a definition of fairness, which we call single-threshold fairness, that is fully compatible with simply maximizing a specific kind of utility function without treating fairness as a separate consideration. Here, a decision is considered to be fair if individuals with the same score $s_i = \psi(v_i)$ are treated equally, regardless of group membership (Corbett-Davies & Goel 2018).

This notion of fairness is connected to utility maximization by a set of results showing that, for a certain set of utility functions and scores $\psi(v)$, the utility-maximizing decision rule δ is necessarily a single-threshold rule (Karlin & Rubin 1956, Berger 1985, Corbett-Davies et al. 2017, Lipton et al. 2018). Rules of this form select a threshold c and apply one decision to individuals with scores below c and another to individuals with scores above c. Formally, we have

$$\delta(v) = I(\psi(v) \geq c).$$

For example, if the outcome is loan repayment, individuals with scores above the threshold would be granted a loan while those with scores below it would be denied. A key condition for the optimality of single-threshold rules is that the score $\psi(v)$ is a good approximation of the true conditional probability $\mathrm{P}[Y = 1|V = v]$.

Moreover, applying the same threshold to all individuals, regardless of subgroup membership, maximizes the total utility for each subgroup. To the extent that the benefits and harms of decisions for a member of a subgroup are contained entirely within that subgroup, the single-threshold rule is a viable group-sensitive definition of fairness.

That said, the desirability of single-threshold rules is sensitive to a number of the choices outlined in Section 2. First, these rules are a direct function of the score $\psi(v)$ and the utility function used to evaluate the decision $\delta(v)$. These functions are specified by the decision-maker and are sensitive to data collection, measurement, and modeling choices (Section 2.2). In addition, the optimality results here make strong assumptions about the form of the utility function used to evaluate the decision $\delta(v)$. In particular, they only hold for utility functions that satisfy the separate, symmetric, and simultaneous assumptions of Section 2.2.3.

These sensitivities motivate notions of fairness that are external to the utility maximization problem, and which can be evaluated without taking scoring models or utility functions for granted.

	$Y=1$	$Y=0$	$P(Y=1\|D)$	$P(Y=0\|D)$
$D=1$	True positive	False positive	$P(Y=1\|D=1)$: Positive predictive value	$P(Y=0\|D=0)$: False discovery rate
$D=0$	False negative	True negative	$P(Y=1\|D=0)$: False omission rate	$P(Y=0\|D=0)$: Negative predictive value
$P(D=1\|Y)$	$P(D=1\|Y=1)$: True positive rate	$P(D=1\|Y=0)$: False positive rate		
$P(D=0\|Y)$	$P(D=0\|Y=1)$: False negative rate	$P(D=0\|Y=0)$: True negative rate		$P(D=Y)$: Accuracy

Figure 2

A confusion matrix defining terminology used in this article for relationships between Y, the outcome, and D, the decision. In our prediction-based decision setup, when $Y = D$, the correct decision has been made. Equality among groups of each of these measures defines several mathematical notions of fairness.

4.2. Equal Prediction Measures

We can get group-specific utilities if, as noted above, the impacts of decisions are contained within groups. In this case, a notion of fairness might ask these to be equal.

When false positives and false negatives have equal cost, this corresponds to the fairness definition of equal accuracy: $\mathrm{P}[D = Y|A = a] = \mathrm{P}[D = Y|A = a']$. This definition is based on the understanding that fairness is embodied by the predictions being correct at the same rate among groups. For example, we might want a medical diagnostic tool to be equally accurate for people of all races or genders.

Instead of comparing overall accuracies, we could restrict the comparison to subsets of our advantaged and disadvantaged groups defined by their predictions or their outcomes. All such possible subsets are summarized by a confusion matrix, which illustrates match and mismatch between Y and D, with margins expressing conditioning on subsets; this is shown in **Figure 2**, which defines common terminology for quantities that will be discussed in this article.

4.2.1. Definitions from the confusion matrix. For any box in the confusion matrix involving the decision D, we can require equality across groups. For example, we could define fairness by equality of false positive rates by requiring that the model satisfy $\mathrm{P}[D = 1|Y = 0, A = a] = \mathrm{P}[D = 1|Y = 0, A = a']$. All other cells of the confusion matrix can similarly define fairness analogously by adding the conditioning on sensitive group attribute, A. We list common definitions of fairness that have been proposed in the literature from the margins of the confusion matrix, grouped by pairs that sum to one. Equality of one member of the pair immediately implies equality of the other.

4.2.1.1. Conditional on outcome. First consider conditioning on the outcome Y. This leads to two pairs of fairness definitions. The first pair, equality of false positive rates and equality of true negative rates, conditions on $Y = 0$ and is equivalent to requiring $D \perp A \mid Y = 0$. This definition of fairness demands equality from the perspective of those individuals with the outcome defined by zero. For example, this might reflect the perspective of innocent defendants, in requiring that all

individuals who do not go on to be rearrested have the same likelihood of being released, regardless of whether they are in the advantaged or disadvantaged group (Narayanan 2018).

The second pair, equality of true positive rates and equality of false negative rates, is equivalent to $D \perp A \mid Y = 1$. This definition of fairness takes the perspective of those individuals with outcome defined by one. For example, this is equivalent to the requirement that all individuals who will go on to repay a loan have the same likelihood of receiving a loan, regardless of their sensitive group membership. This condition alone has been called equal opportunity (Hardt et al. 2016). Taken together with equality of false positive rates/true negative rates, these notions of fairness are called error rate balance (Chouldechova 2017), separation (Barocas et al. 2018), or equalized odds (Hardt et al. 2016).

These two pairs reflect a fairness notion that people with the same outcome should be treated the same, regardless of sensitive group membership. We posit that this notion of fairness is more closely aligned with the perspective of the population evaluated by the model as it demands that people who are actually similar with respect to their outcomes be treated similarly. Hardt et al. (2016) give an algorithm for model fitting that is optimal with respect to this notion of fairness. Kearns et al. (2018) develop intersectional methods to ensure parity of fairness metrics, including false positive rates, for groups of large enough size that are defined by the sensitive variables.

4.2.1.2. Conditional on decision. Turning to the other margin of the confusion matrix, equality of negative predictive value and equality of false omission rate are defined by the statement $Y \perp A \mid D = 0$. This notion of fairness conditions on having received the decision defined by zero. For example, this definition requires that all individuals who were denied a loan be equally likely to have defaulted had the loan been granted. The other pair of definitions that appear on this margin are equality of positive predictive value and equality of false discovery rate. These are defined by the conditional independence relationship $Y \perp A \mid D = 1$. This definition of fairness is also sometimes called predictive parity (Chouldechova 2017), and it is assessed by an outcome test (Simoiu et al. 2017). This definition of fairness, for example, is met when people from both the advantaged and disadvantaged groups who are granted loans go on to repay them at the same rate. These two pairs of definitions taken together have been called sufficiency (Barocas et al. 2018). They reflect a fairness notion that people with the same decision would have had similar outcomes, regardless of group.

These two pairs of definitions of fairness reflect the viewpoint of the decision-maker or modeler, as individuals are grouped with respect to the decision or model's prediction, not with respect to their actual outcome. Dieterich et al. (2016) argue that this definition of fairness is more appropriate because at the time of the decision, the decision-maker knows only the prediction, not the eventual outcome for the individual, and so individuals should be grouped by the characteristic that is known at the time of the decision.

Zafar et al. (2017) call all four pairs of definitions (as well as equal accuracy) avoiding disparate mistreatment. Berk et al. (2017) also consider a definition based on several of the above elements of the confusion matrix. They define treatment equality in terms of the ratio of false negatives to false positives: $\frac{\mathrm{P}[Y=1,D=0|A=a]}{\mathrm{P}[Y=0,D=1|A=a]} = \frac{\mathrm{P}[Y=1,D=0|A=a']}{\mathrm{P}[Y=0,D=1|A=a']}$.

4.2.2. Analogues with scores. In the previous section, we focused on fairness definitions defined by the relationship between Y and D, a binary decision. As discussed in Section 4.1, the ultimate decision is typically arrived at by thresholding a score $S = \psi(V)$ that is intended to estimate $\mathrm{P}[Y = 1|V = v]$. In this section, we consider definitions based on the relationship between S and Y and draw connections to the confusion matrix–based definitions.

Area under curve (AUC) parity is the requirement that the area under the receiver operating characteristic curve is the same across groups. This is analogous to equality of accuracy in the previous section, as the AUC is a measure of model accuracy.

Balance for the negative class is defined by $\mathrm{E}(S|Y = 0, A = a) = \mathrm{E}(S|Y = 0, A = a')$. This is similar to equality of false positive rates in that if the score function is the same as the binary decision, i.e., $S = D$, then achieving equality of false positive rates implies balance for the negative class. This is easily seen because the conditional expectation of a binary variable is, by definition, the same as the conditional probability that the variable takes value one. Balance for the positive class is similarly defined as $\mathrm{E}(S|Y = 1, A = a) = \mathrm{E}(S|Y = 1, A = a')$. By an analogous argument to that above, this definition of fairness is closely related to equality of true positive rates. Separation denotes conditional independence of the protected group variable with the score or decision given Y and covers all definitions discussed in this paragraph.

Calibration within groups is satisfied when $\mathrm{P}[Y = 1|S, A = a] = S$—that is, when the score function accurately reflects the conditional probability of $Y|V$. Barocas et al. (2018) point out that calibration within groups is satisfied without a fairness-specific effort. With enough (representative, well-measured) data and model flexibility, a score S can be very close to $\mathrm{E}(Y|A, X)$. With many X variables, A may be well-predicted by them, i.e., there is a function $a(X)$ that is approximately A. Then we can get calibration within groups even without using A because $\mathrm{E}(Y|A, X) = \mathrm{E}(Y|X)$.

Calibration within groups is the multi-valued analogue of equality of positive predictive value and equality of negative predictive value. Sufficiency, $Y \perp A \,|\, D$ or $Y \perp A \,|\, S$, is closely related to both of these cases. In terms of S, calibration within groups implies sufficiency. Conversely, if S satisfies sufficiency then there exists a function l such that $l(S)$ satisfies calibration within groups (Barocas et al. 2018).

4.3. Equal Decision Measures

We now turn to fairness notions that focus on decisions D without consideration of Y. These can be motivated in a few ways. Suppose that, from the perspective of the individuals about whom we make decisions, one decision is always preferable to another, regardless of Y (e.g., nondetention).[2] In other words, allocation of benefits and harms across groups can be examined by looking at the decision (D) alone. Furthermore, while the decisions (e.g., detentions) are observed, the outcomes being predicted (e.g., crime if released) may be unobserved or poorly measured, making error rates unknown. Therefore, disparity in decisions (e.g., racial disparity in detention rates) may be more publicly visible than disparity in error rates (e.g., racial disparity in detention rates among those who would not have committed a crime).

Yet another motivation to consider fairness constraints without the outcome Y is measurement error (see Section 2.2.1.1). For example, if Y suffers from differential measurement error, fairness constraints based on Y may be unappealing (Johndrow & Lum 2019). One might believe that all group differences in Y are a result of measurement error and that the true outcomes on which we want to base decisions are actually similar across groups (Friedler et al. 2016).

Even more broadly, we might consider the relationship between A and Y to be unfair, even if the observed relationship in the data is accurately capturing a real-world phenomenon. These considerations can all motivate requiring demographic parity: equal decision rates across groups regardless of outcome Y. This is also sometimes called statistical parity or group fairness. This

[2] In contrast, lending to someone unable to repay could hurt their credit score (Liu et al. 2018). Of course, the ability to repay may strongly depend on the terms of the loan.

fairness definition can be thought of in terms of (unconditional) independence: $S \perp A$ or $D \perp A$. Calders & Verwer (2010), Feldman et al. (2015), and Johndrow & Lum (2019) give algorithms to build models achieving this notion of fairness

A related definition considers parity within strata: Conditional demographic parity is defined by the condition that $D \perp A \mid \text{Data}$. When $\text{Data} = Y$, conditional demographic parity is equivalent to separation. When $\text{Data} = X$ (the insensitive variables), it is equivalent to fairness through unawareness (Kusner et al. 2017), anticlassification (Corbett-Davies & Goel 2018), or treatment parity (Lipton et al. 2018). This is easily achieved by not allowing a model to directly access information about A. Unawareness implies that people with the same x are treated the same, i.e., $\delta(v_i) = \delta(v_{i'})$ if $x_i = x_{i'}$. Note this does not imply that all people who are treated the same have the same covariates, since it is possible that many people receive the same treatment while no two people have the same x. A related idea requires people who are similar in x to be treated similarly. More generally, we could define a similarity metric between people that is aware of the sensitive variables, motivating the next flavor of fairness definitions (Dwork et al. 2012).

4.4. Impossibilities

Although each flavor of fairness definition presented in this section formalizes an intuitive notion of fairness, these definitions are not mathematically or morally compatible in general. In this section, we review several impossibility results about definitions of fairness, providing context for how this discussion unfolded in the algorithmic fairness literature.

4.4.1. The COMPAS debate. In the algorithmic fairness literature, incompatibilities between fairness definitions were brought to the fore in a public debate over a tool called COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) deployed in criminal justice settings.

In 2016, ProPublica published a highly influential analysis based on data obtained through public records requests (Angwin et al. 2016). Their most discussed finding was that COMPAS does not satisfy equal false positive rates by race: Among defendants who did not get rearrested, black defendants were twice as likely to be misclassified as high risk. Based largely on this and other similar findings, they described the tool as biased against blacks.

Northpointe (now Equivant), the developers of COMPAS, critiqued ProPublica's work and pointed out that COMPAS satisfies equal positive predictive values: among those called higher risk, the proportion of defendants who got rearrested is approximately the same regardless of race (Dieterich et al. 2016). COMPAS also satisfies calibration within groups (Flores et al. 2016). Much of the subsequent conversation consisted of either trying to harmonize these definitions of fairness or asserting that one or the other is correct. As it turns out, there can be no harmony among definitions in a world where inequality and imperfect prediction are the reality.

4.4.2. Separation and sufficiency. Tension between margins of the confusion matrix is expressed in three very similar results. Barocas et al. (2018) and Wasserman (2010) show that under the assumption of separation ($S \perp A|Y$) and sufficiency ($Y \perp A|S$), either $(Y, S) \perp A$ or an event in the joint distribution has probability zero.

Putting this in context and recalling that sufficiency implies equality of positive predictive values and that separation implies equality of false positive rates, this result shows that both supported definitions of fairness in the COMPAS debate can only be met when (*a*) the rate of recidivism and the distribution of scores are the same for all racial groups or (*b*) there are some groups that never experience some of the outcomes (e.g., white people are never rearrested).

A related result was given by Kleinberg et al. (2016), who showed that if a model satisfies balance for the negative class, balance for the positive class, and calibration within groups, then either there are equal base rates ($Y \perp A$) or there was perfect prediction ($\mathrm{P}[Y = 1|V = v] = 0$ or 1 for all $v \in \mathcal{V}$). A very similar result was shown by Chouldechova (2017). In the context of the COMPAS debate, this requires either that reality is equal and fair (i.e., there is no racial disparity in recidivism rates) or that the model is perfectly able to predict recidivism (a reality that is, as of now, unattainable). Equal base rates and perfect prediction can be called trivial, degenerate, or even utopian (representing two very different utopias). Regardless of description, these conditions were not met in ProPublica's data on COMPAS, and so the definitions of fairness championed by the different sides of the debate cannot be achieved simultaneously.

4.4.3. Incompatibilities with demographic parity. Here we describe several impossibility results involving demographic parity, though they have not played so prominent a role in the public debate about fairness. Barocas et al. (2018) showed that when Y is binary and the score exhibits separation ($S \perp A|Y$) and demographic parity ($S \perp A$), then there must be at least one of equal base rates ($Y \perp A$) or the score is useless for predicting the outcome ($Y \perp S$).

Similarly, Barocas et al. (2018) also showed that if a model satisfies sufficiency ($Y \perp A|S$) and demographic parity ($S \perp A$), then there must also be equal base rates: $Y \perp A$. Taken together, in the context of the COMPAS debate, even if we could decide that equality of positive predictive values or equality of negative predictive values were the relevant notions of fairness, if we also want to constrain the model to avoid disparate impact by requiring demographic parity, this would only be possible if we lived in a world in which there are no racial disparities in rearrest and/or we have a completely useless predictive model.

Finally, Corbett-Davies et al. (2017) and Lipton et al. (2018) both note that a decision rule δ that maximizes utility under a demographic parity constraint (in general) uses the sensitive variables a both in estimating the conditional probabilities and for determining their thresholds. Therefore, solutions such as disparate learning processes, which allow the use of sensitive variables during model building but not prediction, are either sub- or equi-optimal (Pedreshi et al. 2008, Kamiran & Calders 2009).

5. FLAVORS OF FAIRNESS DEFINITIONS INCORPORATING ADDITIONAL CONTEXT

So far, we have discussed oblivious fairness based on summaries of the joint and marginal distributions of Y, V, S, and D. In this section, we consider fairness definitions that incorporate additional context, in the form of metrics and causal models, to inform fairness considerations. This external context provides additional degrees of freedom for mapping social goals onto mathematical formalism.

5.1. Metric Fairness

Assume there is a metric that defines similarity based on all variables, $m : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$. Then, metric fairness is defined such that for every $v, v' \in \mathcal{V}$, their closeness implies closeness in decisions $|\delta(v) - \delta(v')| \leq m(v, v')$. This is also known as individual fairness, the m-Lipschitz property (Dwork et al. 2012), and perfect metric fairness (Rothblum & Yona 2018). In cases where the metric only considers insensitive variables, $m(x, x')$, metric fairness implies unawareness.

Definitions of the metric differ. In the original work, Dwork et al. (2012) consider a similarity metric over individuals. But in subsequent research, the metric is often defined over the variables input to the classifier (Kim et al. 2018, Rothblum & Yona 2018).

Either way, the metric is meant to capture ground truth. This inspired Friedler et al. (2016) to define the construct space, the variables on which we want to base decisions. For example, suppose we want to base decisions on the probability of the outcome for an individual i. Let I be a random individual drawn from the population. We can express the construct space as $\mathcal{CS} = \{t_i\}$, where $t_i = \mathrm{P}[Y = 1|I = i]$, but we cannot estimate $\mathrm{P}[Y = 1|I = i]$ because we only have one individual i and we do not observe their outcome in time to make the decision. Instead, we calculate scores $s_i = \psi(v_i)$ intended to estimate $\mathrm{P}[Y = 1|V = v_i]$. As noted above, the conditional probabilities $\mathrm{P}[Y = 1|V = v_i]$ are sometimes called an individual's true risk. However, despite the tendency in the computer science literature to conflate an individual with their probabilities, the probabilities do not condition on the individual, only some measured variables.

Friedler et al. (2016) introduce an assumption they call what you see is what you get (WYSIWYG), i.e., that we can define a metric in the observed space that approximates a metric in the construct space: $m(v_i, v_{i'}) \approx m_{\mathcal{CS}}(t_i, t_{i'})$. To satisfy WYSIWYG, m may need to be aware of the sensitive variables (Dwork et al. 2012). One reason is that the insensitive variables X may predict Y differently for different groups. For example, suppose we want to predict who likes math so we can recruit them to the school's math team. Let $Y = 1$ be liking math and X be choice of major. Suppose students who like math in one group are steered toward economics, and in the other group they are steered toward engineering. To predict liking math, we should use group membership in addition to X. Using this terminology, the metric of Dwork et al. (2012) can be defined as aligning differences in the construct space with differences in the observed space.

Friedler et al. (2016) also introduce an alternate assumption called we're all equal (WAE), i.e., that the groups on average have small distance in the construct space. On this basis, we could adjust a metric in the observed space so that the groups have small distance. Methods for adjusting the insensitive variables X so that they are independent of group are consistent with adjusting the observed space so that it is consistent with a WAE understanding of the construct space (Calders & Verwer 2010, Feldman et al. 2015, Johndrow & Lum 2019).

Though metric fairness is conceptually appealing, one major difficulty of implementing it is defining the metric itself, especially in high dimensions. Jung et al. (2019) bypass explicit elicitation of m and instead query individuals only on whether $|\delta(v) - \delta(v')|$ is small for many pairs of individuals i and i'. For example, the method requires data on whether individuals i and i' ought to be treated similarly without explicitly requiring the decision-maker to give an analytical expression for m. The objective function for model fitting then incorporates this notion of fairness by enforcing the elicited pairwise constraints.

5.2. Causal Definitions

In this section, we discuss an alternative framework for conceiving of model fairness: causality. Framing fairness issues with causal language can make value judgments more explicit. In particular, this framing allows practitioners to designate which causal pathways from sensitive attributes to decisions constitute acceptable or unacceptable sources of dependence between sensitive attributes and decisions. A number of the key questions involved in mapping social goals to mathematical formalism can thus be addressed by examining a causal graph and discussing these value judgments.

We have already touched on causal notions, considering the potential or counterfactual values under different decisions in Section 2.1.1. Causal fairness definitions consider instead counterfactuals under different settings of a sensitive variable. Let $v_i(a) = (a, x_i(a))$ be the covariates if the

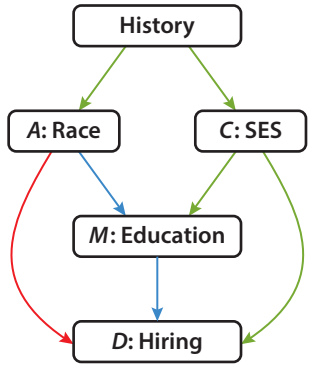

Figure 3

An example causal graph where a complex historical process creates an individual's race (sensitive variable A) and socioeconomic status (SES) at birth (confounder C). Race and SES are correlated due to this upstream history. They both affect hiring (decision D), including through education (mediator M). We are concerned with paths between race and hiring, which include direct (*red*), indirect (*blue*), and back-door (*green*) paths.

individual had their sensitive variable set to a. We write $d_i(a) = \delta(v_i(a)) = \delta(a, x_i(a))$ for the corresponding decision, e.g., what would the hiring decision be if a black job candidate had been white? We define random variables $V(a)$, $D(a)$ as values randomly drawn from the population (Kohler-Hausmann 2018).

There is debate over whether these counterfactuals are well-defined. Pearl (2009) allows counterfactuals under conditions without specifying how those conditions are established, e.g., if the candidate had been white. In contrast, Hernán & Robins (2018) introduce counterfactuals only under well-defined interventions, e.g., the intervention studied by Greiner & Rubin (2011): if the name on their resume were set to be atypical among black people.

Putting these issues to the side, we can proceed to define fairness in terms of counterfactual decisions under different settings of a sensitive variable. Ordered from strongest to weakest, we have individual counterfactual fairness ($d_i(a) = d_i(a')$ for all i), conditional counterfactual fairness (Kusner et al. 2017) ($\mathrm{E}[D(a) \mid \mathrm{Data}] = \mathrm{E}[D(a') \mid \mathrm{Data}]$), and counterfactual parity ($\mathrm{E}[D(a)] = \mathrm{E}[D(a')]$). These first three causal definitions consider the total effect of A (e.g., race) on D (e.g., hiring). However, it is possible to consider some causal pathways from the sensitive variable to be fair. For example, suppose race affects education obtained. If hiring decisions are based on the applicant's education, then race affects hiring decisions. Perhaps one considers this path from race to hiring through education to be fair. It often helps to visualize causal relationships graphically (see **Figure 3**).[3]

In this cartoon version of the world, a complex historical process creates an individual's race and socioeconomic status at birth (VanderWeele & Robinson 2014, Jackson & VanderWeele 2018). These both affect the hiring decision, including through education. Let $x_i = (c_i, m_i)$, where m_i are variables possibly affected by race, so $x_i(a) = (c_i, m_i(a))$. We can define effects along paths by defining more nuanced counterfactuals. Let $d_i(a', m_i(a))$ be the decision if applicant i had their race set to a', while their education was set to whatever value it would have attained if they had their race set to a (Nabi & Shpitser 2018). To disallow the red path in **Figure 3**, we can define no direct effect fairness: $d_i(a) = d_i(a', m_i(a))$ for all i.

[3]See Pearl (2009), section 1.3.1, for a definition of causal graphs, which encode conditional independence statements for counterfactuals.

However, $m_i(a)$ is only observed when $a_i = a$, so it is not possible to confirm no direct effect fairness without direct access to the model's inner workings. If we are willing to make stronger assumptions and assume ignorability, $M(a) \perp A \mid C$, we can, however, check no average direct effect fairness: $\mathrm{E}[D(a)] = \mathrm{E}[D(a', M(a))]$.

Beyond direct effects, one could consider other directed paths from race to be fair or unfair. For example, no unfair path-specific effects specifies no average effects along unfair (user-specified) directed paths from A to D (Nabi & Shpitser 2018). Relatedly, no unresolved discrimination states that there should exist no directed path from A to D unless through a resolving variable—a variable that is influenced by A in a manner that we accept as nondiscriminatory (Kilbertus et al. 2017).

All of the above causal definitions consider only directed paths from race. In **Figure 3**, these include the red and blue paths. But what about the green paths? Known as back-door paths, these do not represent causal effects of race and therefore are permitted under causal definitions of fairness (Pearl 2009). However, back-door paths can contribute to the association between race and hiring. Indeed, they are why we say "correlation is not causation."[4] Zhang & Bareinboim (2018) decompose the total disparity into disparities from each type of path (direct, indirect, back-door). In contrast to the causal fairness definitions, health disparities are defined to include contributions from back-door paths (e.g., through socioeconomics at birth) (IOM 2003, Jackson & VanderWeele 2018).

Causal definitions of fairness focus our attention on how to compensate for causal influences at decision time. Causal reasoning can be used instead to design interventions (to reduce disparities and improve overall outcomes) rather than to define fairness. In particular, causal graphs can be used to develop interventions at earlier points, prior to decision-making (Jackson & VanderWeele 2018, Barabas et al. 2018).

6. WAYS FORWARD

In this article, we have been careful to identify assumptions, choices, and considerations in prediction-based decision-making that are and are not challenged by various mathematically defined notions of fairness. This article does not, however, discuss what to do with a flavor of fairness. The dominant focus of the algorithmic fairness literature has been to constrain decision functions to satisfy particular fairness flavors and to treat fair decision-making as a constrained optimization problem (see, e.g., Hardt et al. 2016). Another way forward is to address the choices and assumptions outlined in Section 2 directly. Here, we sketch that approach for some of the choices and assumptions for our running pretrial risk assessment example.

Starting with clearly articulated goals can improve both fairness and accountability. One example of a clearly stated goal for pretrial decisions has been articulated in a statement of concern by the Leadership Conference on Civil and Human Rights (Leadersh. Conf. Civ. Hum. Rights 2020, p. 2). It reads "If in use, a pretrial risk assessment instrument must be designed and implemented in ways that reduce and ultimately eliminate unwarranted racial disparities across the criminal justice system." Another commonly stated goal in pretrial risk assessment is reducing the number of people detained pretrial.

Similarly, careful consideration of how individuals enter the population that is subject to the predictive model can reveal ways to limit or expand that population to help meet the stated goals.

[4]If C satisfies the back-door criterion (C includes no descendants of A and blocks all back-door paths between A and D) in a causal graph, then unconfoundedness ($D(a) \perp A | C \; \forall a$) holds (Pearl 2009, Perković et al. 2015). The converse is not true in general.

One possibility to reduce the number of people subject to a pretrial risk assessment model is to enact policies that result in fewer arrests in the first place. One such policy would be decriminalization, e.g., decriminalizing marijuana use, which has broad support (Daniller 2019).

Expanding the decision space to include less harmful and more supportive interventions can benefit all groups and mitigate fairness concerns. One example of this in the pretrial context can be seen in the New Jersey Criminal Justice Reform Act, which essentially eliminated money bail as an option as a condition of pretrial release. Other, supportive interventions that have been proposed include funding transportation to court, paying for child care, and text message reminders of court dates.

If a predictive tool is built, it is important to choose outcomes carefully, considering data limitations. Although there remains debate about whether this adequately addresses problems of unfairness, one proposed solution is to avoid predicting rearrest for any crime and instead predict rearrest for violent crime, as it is argued that the latter outcome variable suffers from less measurement error (Skeem & Lowenkamp 2016).

Documenting all of these choices, from data collection, sampling, and measurement to data processing, enables modeling that appropriately accounts for the specific conditions under which data have been collected (Gebru et al. 2018, Holland et al. 2018). Similarly, documenting model performance, both overall and within subgroups, is crucial to effective evaluation and can also help check decision systems against some of the fairness definitions from Section 4 (Mitchell et al. 2018, Stoyanovich 2018, Yang et al. 2018). For our pretrial example, laws could require publicly available documentation of risk assessment tools, including details of data collection, development, and model performance.

Finally, in our opinion, it is crucial to involve members of the impacted community in the entire development process—from problem formulation through evaluation. This is also true in the pretrial risk assessment context. Indeed, one of the guiding principles in the previously mentioned statement of concern is that "a pretrial risk assessment instrument must be developed with community input, revalidated regularly by independent data scientists with that input in mind, and subjected to regular, meaningful oversight by the community" (Leadersh. Conf. Civ. Hum. Rights 2020, p. 8).

7. CONCLUSION

The fact that much of the work on fairness in prediction-based decision-making has emerged from computer science and statistics has sometimes led to the mistaken impression that these concerns only arise in cases that involve predictive models and automated decision-making. Yet, many of the issues identified in the scholarship apply to human decision-making as well, to the extent that human decision-making also rests on predictions. Notably, scholars have emphasized that the trade-off between different notions of fairness would apply even if a human were the one making the prediction (Kleinberg et al. 2016). Rejecting model-driven or automated decision-making is not a way to avoid these problems.

At best, formal fairness metrics can instead illustrate when changes to prediction-based decision-making are insufficient to achieve different outcomes—and when interventions are necessary to bring about a different world. Recognizing the trade-offs involved in prediction-based decision-making does not mean that we have to accept them as a given; doing so can also spur us to think more creatively about the range of options we have to realize our policy or normative goal beyond just making predictions.

Recent criticisms of this line of work have rightly pointed out that quantitative notions of fairness can funnel our thinking into narrow silos where we aim to make adjustments to a

decision-making process, rather than to address the structural conditions that sustain inequality in society (Green & Hu 2018, Ochigame et al. 2018). While algorithmic thinking runs such risks, quantitative modeling and quantitative measures can also force us to make our assumptions more explicit and clarify what we are treating as background conditions (and thus not the target of intervention). In doing so, we have the opportunity to foster more meaningful deliberation and debate about the difficult policy issues that we might otherwise hand-wave away: What is our objective, and how do we want to go about achieving it?

Used with care and humility, the recent work on fairness can play a helpful part in revealing problems with prediction-based decision-making and provide useful tools for addressing them. While mathematical formalism cannot solve these problems on its own, it should not be dismissed as necessarily preserving the status quo. The opportunity exists to employ quantitative methods to make meaningful progress on policy goals (Mitchell et al. 2018, Fussell 2018). Moreover, the literature that we have reviewed in this article can foster critical reflection on how we choose those goals and policies for realizing them.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Jackie Shadlen provided substantial insights that contributed greatly to this article. We are grateful to all cited authors for their work and for answering our many questions. Funding for this research was provided by the Harvard-MIT Ethics and Governance of Artificial Intelligence Initiative.

LITERATURE CITED

Alexander M. 2012. *The New Jim Crow*. New York: New Press

Angell R, Johnson B, Brun Y, Meliou A. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pp. 871–75. New York: ACM

Angwin J, Larson J, Mattu S, Kirchner L. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*, May 23. **https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing**

Barabas C, Dinakar K, Ito J, Virza M, Zittrain J. 2018. Interventions over predictions: reframing the ethical debate for actuarial risk assessment. arXiv:1712.08238 [cs.LG]

Barocas S, Hardt M, Narayanan A. 2018. *Fairness and Machine Learning*. **http://www.fairmlbook.org**

Barocas S, Selbst AD. 2016. Big data's disparate impact. *Calif. Law Rev.* 104:671–732

Berger JO. 1985. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer

Berk R, Heidari H, Jabbari S, Kearns M, Roth A. 2017. Fairness in criminal justice risk assessments: the state of the art. arXiv:1703.09207 [stat.ML]

Brownsberger WN. 2017. *Bill S.770: An act providing community-based sentencing alternatives for primary caretakers of dependent children who have been convicted of non-violent crimes*. Senate Docket No. 622, Commonw. Mass. **https://malegislature.gov/Bills/190/S770**

Buolamwini J, Gebru T. 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. *Proc. Mach. Learn. Res.* 81:77–91

Calders T, Verwer S. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* 21:277–92

Chouldechova A. 2017. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5:153–63

Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. *Proc. Mach. Learn. Res.* 81:134–48

Chouldechova A, G'Sell M. 2017. Fairer and more accurate, but for whom? arXiv:1707.00046 [stat.AP]

Chouldechova A, Roth A. 2018. The frontiers of fairness in machine learning. arXiv:1810.08810 [cs.LG]

Cleary AT. 1966. *Test bias: validity of the scholastic aptitude test for negro and white students in integrated colleges*. Res. Bull. RB-66-31, Educ. Test. Serv., Princeton, NJ

Corbett-Davies S, Goel S. 2018. The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv:1808.00023 [cs.CY]

Corbett-Davies S, Pierson E, Feller A, Goel S, 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*, pp. 797–806. New York: ACM

Coston A, Mishler A, Kennedy EH, Chouldechova A. 2020. Counterfactual risk assessments, evaluation, and fairness. arXiv:1909.00066 [stat.ML]

Courchane M, Nebhut D, Nickerson D. 2000. Lessons learned: statistical techniques and fair lending. *J. Hous. Res.* 11:277–95

Crenshaw K. 1989. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *Univ. Chic. Legal Forum* 1989:8

Daniller A. 2019. Two-thirds of Americans support marijuana legalization. *FactTank Blog/Pew Research Center*, Nov. 14. **https://www.pewresearch.org/fact-tank/2019/11/14/americans-support-marijuana-legalization/**

Darlington RB. 1971. Another look at 'cultural fairness'. *J. Educ. Meas.* 8:71–82

Day JN. 2002. Credit, capital and community: informal banking in immigrant communities in the United States, 1880–1924. *Financ. Hist. Rev.* 9:65–78

De-Arteaga M, Dubrawski A, Chouldechova A. 2018. Learning under selective labels in the presence of expert consistency. arXiv:1807.00905 [cs.LG]

Dieterich W, Mendoza C, Brennan T. 2016. *COMPAS risk scales: demonstrating accuracy equity and predictive parity*. Work. Pap., Northpointe Inc., Traverse City, MI. **http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf**

Dobbe R, Dean S, Gilbert T, Kohli N. 2018. A broader view on bias in automated decision-making: reflecting on epistemology and dynamics. arXiv:1807.00553 [cs.LG]

Dorans NJ, Cook LL. 2016. *Fairness in Educational Assessment and Measurement*. New York: Taylor & Francis

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–26. New York: ACM

Eckhouse L, Lum K, Conti-Cook C, Ciccolini J. 2018. Layers of bias: a unified approach for understanding problems with risk assessment. *Crim. Justice Behav.* 46:185–209

Einhorn HJ, Bass AR. 1971. Methodological considerations relevant to discrimination in employment testing. *Psychol. Bull.* 75:261

Eubanks V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's

Exec. Off. Pres. 2016. *Big data: a report on algorithmic systems, opportunity, and civil rights*. Rep., Exec. Off. Pres., Washington, DC. **https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf**

Feldman M, Friedler SA, Moeller J, Scheidegger C, Venkatasubramanian S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–68. New York: ACM

Flores AW, Bechtel K, Lowenkamp CT. 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. Probat. J.* 80:38

Friedler SA, Scheidegger C, Venkatasubramanian S. 2016. On the (im)possibility of fairness. arXiv:1609.07236 [cs.CY]

Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D. 2018. A comparative study of fairness-enhancing interventions in machine learning. arXiv:1802.04422 [stat.ML]

Fussell S. 2018. The algorithm that could save vulnerable New Yorkers from being forced out of their homes. *Gizmodo*, Aug. 18. **https://gizmodo.com/the-algorithm-that-could-save-vulnerable-new-yorkers-fr-1826807459**

Fuster A, Goldsmith-Pinkham P, Ramadorai T, Walther A. 2020. Predictably unequal? The effects of machine learning on credit markets. *SSRN*. **https://dx.doi.org/10.2139/ssrn.3072038**

Galhotra S, Brun Y, Meliou A. 2017. Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pp. 498–510. New York: ACM

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, et al. 2018. Datasheets for datasets. arXiv:1803.09010 [cs.DB]

Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC. 3rd ed.

Green B. 2018. *"Fair" risk assessments: a precarious approach for criminal justice reform*. Paper presented at the 5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018), Stockholm, Sweden

Green B, Hu L. 2018. *The myth in the methodology: towards a recontextualization of fairness in machine learning*. Presented at Machine Learning: The Debates Workshop, 35th International Conference on Machine Learning (ICML), Stockholm, Sweden

Greiner JD, Rubin DB. 2011. Causal effects of perceived immutable characteristics. *Rev. Econ. Stat.* 93:775–85

Harcourt BE. 2008. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age*. Chicago: Univ. Chicago Press

Hardt M, Price E, Srebro N. 2016. Equality of opportunity in supervised learning. arXiv:1610.02413 [cs.LG]

Hernán MA, Robins JM. 2018. *Causal Inference*. Boca Raton, FL: Chapman & Hall/CRC

Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. 2018. The dataset nutrition label: a framework to drive higher data quality standards. arXiv:1805.03677 [cs.DB]

Hu L, Chen Y. 2018a. A short-term intervention for long-term fairness in the labor market. arXiv:1712.00064 [cs.GT]

Hu L, Chen Y. 2018b. Welfare and distributional impacts of fair classification. arXiv:1807.01134 [cs.LG]

Hunter JE, Schmidt FL. 1976. Critical analysis of the statistical and ethical implications of various definitions of test bias. *Psychol. Bull.* 83:1053

Hutchinson B, Mitchell M. 2019. 50 years of test (un)fairness: lessons for machine learning. arXiv:1811.10104 [cs.AI]

Imbens GW, Rubin DB. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge, UK: Cambridge Univ. Press

IOM (Inst. Med.). 2003. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*. Washington, DC: Natl. Acad. Press

Jackson JW, VanderWeele TJ. 2018. Decomposition analysis to identify intervention targets for reducing disparities. *Epidemiology* 29:825–35

Johndrow JE, Lum K. 2019. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *Ann. Appl. Stat.* 13:189–220

Jung C, Kearns M, Neel S, Roth A, Stapleton L, Wu ZS. 2019. Eliciting and enforcing subjective individual fairness. arXiv:1905.10660 [cs.LG]

Kallus N, Zhou A. 2018. Residual unfairness in fair machine learning from prejudiced data. arXiv:1806.02887 [stat.ML]

Kamiran F, Calders T. 2009. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pp. 1–6. Red Hook, NY: Curran

Karlin S, Rubin H. 1956. The theory of decision procedures for distributions with monotone likelihood ratio. *Ann. Math. Stat.* 27:272–99

Kearns M, Neel S, Roth A, Wu ZS. 2018. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. *Proc. Mach. Learn. Res.* 80:2564–72

Kilbertus N, Carulla MR, Parascandolo G, Hardt M, Janzing D, Schölkopf B. 2017. Avoiding discrimination through causal reasoning. arXiv:1706.02744 [stat.ML]

Kim MP, Reingold O, Rothblum GN. 2018. Fairness through computationally-bounded awareness. arXiv:1803.03239 [cs.LG]

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. 2017. Human decisions and machine predictions. *Q. J. Econ.* 133:237–93

Kleinberg J, Ludwig J, Mullainathan S, Rambachan A. 2018. Algorithmic fairness. *AEA Pap. Proc.* 108:22–27

Kleinberg J, Mullainathan S, Raghavan M. 2016. Inherent trade-offs in the fair determination of risk scores. arXiv:1609.05807 [cs.LG]

Kohler-Hausmann I. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern Univ. Law Rev.* 113:1163–228

Koulish R. 2016. Immigration detention in the risk classification assessment era. *Conn. Public Interest Law J.* 16:1

Kusner MJ, Loftus J, Russell C, Silva R. 2017. Counterfactual fairness. arXiv:1703.06856 [stat.ML]

Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J, Mullainathan S. 2017. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–84. New York: ACM

Larson J, Mattu S, Kirchner L, Angwin J. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica*, May 23. **https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm**

Leadersh. Conf. Civ. Hum. Rights. 2020. *The use of pre-trial "risk assessment" instruments: a shared statement of civil rights concerns*. Statement, Leadersh. Conf. Civil Hum. Rights, Washington, DC

Lewis MA. 1978. *A comparison of three models for determining test fairness*. Tech. Rep., Fed. Aviat. Adm., Washington, DC

Lipton Z, McAuley J, Chouldechova A. 2018. Does mitigating ML's impact disparity require treatment disparity? arXiv:1711.07076 [stat.ML]

Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M. 2018. Delayed impact of fair machine learning. arXiv:1803.04383 [cs.LG]

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

Mayor's Off. Data Anal. 2018. Legionnaires' disease response: MODA assisted in a citywide response effort after an outbreak of Legionnaires' disease. *MODA Project Library*. **https://moda-nyc.github.io/Project-Library/projects/cooling-towers/**

Miller CC. 2015a. Can an algorithm hire better than a human? *New York Times*, June 26. **https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html**

Miller CC. 2015b. When algorithms discriminate. *New York Times*, June 26. **https://www.nytimes.com/2015/06/26/upshot/can-an-algorithm-hire-better-than-a-human.html**

Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, et al. 2018. Model cards for model reporting. arXiv:1810.03993 [cs.LG]

Nabi R, Shpitser I. 2018. Fair inference on outcomes. arXiv:1705.10378 [stat.ML]

Narayanan A. 2018. *21 fairness definitions and their politics*. Presented at ACM FAT* (Fairness, Accountability and Transparency) Conference 2018, New York, NY

Ochigame R, Barabas C, Dinakar K, Virza M, Ito J. 2018. *Beyond legitimation: rethinking fairness, interpretability, and accuracy in machine learning*. Presented at Machine Learning: The Debates Workshop, 35th International Conference on Machine Learning (ICML), Stockholm, Sweden

O'Neil C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown

O'Neil C, Gunn H. 2020. Near term artificial intelligence and the ethical matrix. In *Ethics of Artificial Intelligence*, ed. SM Liao, pp. 235–69. Oxford, UK: Oxford Univ. Press

Passi S, Barocas S. 2019. Problem formulation and fairness. arXiv:1901.02547 [cs.CY]

Pearl J. 2009. *Causality*. Cambridge, UK: Cambridge Univ. Press

Pedreshi D, Ruggieri S, Turini F. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 560–68. New York: ACM

Perković E, Textor J, Kalisch M, Maathuis MH. 2015. A complete generalized adjustment criterion. arXiv:1507.01524 [math.ST]

Petersen NS, Novick MR. 1976. An evaluation of some models for culture-fair selection. *J. Educ. Meas.* 13:3–29
Potash E, Brew J, Loewi A, Majumdar S, Reece A, et al. 2015. Predictive modeling for public health: preventing childhood lead poisoning. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2039–47. New York: ACM
Rabanser S, Günnemann S, Lipton ZC. 2018. Failing loudly: an empirical study of methods for detecting dataset shift. arXiv:1810.11953 [stat.ML]
Raphling J. 2018. Criminalizing homelessness violates basic human rights. *The Nation*, July 5. **https://www.thenation.com/article/criminalizing-homelessness-violates-basic-human-rights/**
Rawls J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard Univ. Press
Rothblum GN, Yona G. 2018. Probably approximately metric-fair learning. *Proc. Mach. Learn. Res.* 80:5680–88
Rudin C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* 1(5):206–15
Selbst AD, boyd d, Friedler S, Venkatasubramanian S, Vertesi J. 2019. Fairness and abstraction in sociotechnical systems. In *FAT* '19: Proceedings of the Conference on Fairness, Accountability and Transparency*, pp. 59–68. New York: ACM
Silva S, Kenney M. 2018. Algorithms, platforms, and ethnic bias: an integrative essay. *Phylon* 55:9–37
Simoiu C, Corbett-Davies S, Goel S. 2017. The problem of infra-marginality in outcome tests for discrimination. *Ann. Appl. Stat.* 11:1193–216
Skeem JL, Lowenkamp CT. 2016. Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology* 54:680–712
Stoyanovich J. 2018. Refining the concept of a nutritional label for data and models. *Freedom to Tinker Blog*, May 3. **https://freedom-to-tinker.com/2018/05/03/refining-the-concept-of-a-nutritional-label-for-data-and-models/**
Suresh H, Guttag JV. 2019. A framework for understanding unintended consequences of machine learning. arXiv:1901.10002 [cs.LG]
Sweeney L. 2013. Discrimination in online ad delivery. *Queue* 11:10
Thorndike RL. 1971. Concepts of culture-fairness. *J. Educ. Meas.* 8:63–70
Vaithianathan R, Maloney T, Putnam-Hornstein E, Jiang N. 2013. Children in the public benefit system at risk of maltreatment: Identification via predictive modeling. *Am. J. Prev. Med.* 45:354–59
VanderWeele TJ, Hernán MA. 2012. Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. *Am. J. Epidemiol.* 175:1303–10
VanderWeele TJ, Robinson WR. 2014. On causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* 25:473–84
Verma S, Rubin J. 2018. Fairness definitions explained. In *FairWare '18: Proceedings of the International Workshop on Software Fairness*, pp. 1–7. Red Hook, NY: Curran
Wasserman L. 2010. *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer
Wexler J. 2018. The what-if tool: code-free probing of machine learning models. *Google AI Blog*, Sept. 11. **https://ai.googleblog.com/2018/09/the-what-if-tool-code-free-probing-of.html**
Yang K, Stoyanovich J, Asudeh A, Howe B, Jagadish H, Miklau G. 2018. A nutritional label for rankings. In *Proceedings of the 2018 International Conference on Management of Data*, pp. 1773–76. New York: ACM
Zafar MB, Valera I, Gomez Rodriguez M, Gummadi KP. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. arXiv:1610.08452 [stat.ML]
Zehlike M, Castillo C, Bonchi F, Baeza-Yates R, Hajian S, Megahed M. 2017. Fairness measures: a platform for data collection and benchmarking in discrimination-aware ML. *Machine Learning Software*. **https://fairnessmeasures.github.io/**
Zhang J, Bareinboim E. 2018. Fairness in decision-making–the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 2037–45. Menlo Park, CA: AAAI