Designing Equitable Algorithms

Alex Chohlas-Wood
Stanford University
alexcw@stanford.edu

Madison Coots
Harvard University
mcoots@g.harvard.edu

Sharad Goel
Harvard University
sgoel@hks.harvard.edu

Julian Nyarko
Stanford University
jnyarko@law.stanford.edu

Abstract

Predictive algorithms are now used to help distribute a large share of our society's resources and sanctions, such as healthcare, loans, criminal detentions, and tax audits. Under the right circumstances, these algorithms can improve the efficiency and equity of decision-making. At the same time, there is a danger that the algorithms themselves could entrench and exacerbate disparities, particularly along racial, ethnic, and gender lines. To help ensure their fairness, many researchers suggest that algorithms be subject to at least one of three constraints: (1) no use of legally protected features, such as race, ethnicity, and gender; (2) equal rates of "positive" decisions across groups; and (3) equal error rates across groups. Here we show that these constraints, while intuitively appealing, often worsen outcomes for individuals in marginalized groups, and can even leave all groups worse off. The inherent trade-off we identify between formal fairness constraints and welfare improvements—particularly for the marginalized—highlights the need for a more robust discussion on what it means for an algorithm to be "fair". We illustrate these ideas with examples from healthcare and the criminal-legal system, and make several proposals to help practitioners design more equitable algorithms.

1 Introduction

With the advancement of statistical methods, computational resources, and data availability, the last decade has seen a dramatic increase in the use of algorithmic decision-making across all facets of life. Banks make algorithmic predictions to assess who is at risk of default and should thus not be offered a loan (68), and to identify possible instances of money laundering (103). In healthcare, algorithms are used to decide who gets screened for diseases like diabetes (1), and to allocate resource-limited benefits such as kidney transplants (41) and HIV-prevention counseling (94). Criminal justice agencies use algorithms to inform the allocation of police resources (35; 78), to assist

investigators (21; 84), to inform incarceration and sentencing decisions (32; 45; 91), and to limit the impact of perceived race on prosecutorial charging decisions (22). Technology companies use algorithms to decide who sees ads for housing (92) and employment opportunities (67), among others. Child services agencies use algorithms to estimate the risk of an adverse event like child abuse (12; 26; 31; 89). City agencies use algorithms to prioritize building inspections (73). School districts use algorithms to assign students to their preferred school (3) and to identify students who are at risk of falling behind on learning material (18).

While it appears that the use of algorithms for critical decision-making will only increase in the near future, some have pointed to a heightened danger that these same algorithms could influence decision-making in a way that is unfair to marginalized groups, such as racial or ethnic minorities (38; 83). Legal scholars have argued that certain forms of algorithmic decision-making may even be in conflict with important constitutional or regulatory protections granted to groups defined by race and ethnicity, rendering them impermissible (7; 51; 55; 74; 97). In response to these concerns, researchers have developed several fairness criteria with the goal of ensuring that algorithms achieve equitable decision-making (24; 28; 77). These criteria range from excluding certain legally protected characteristics—such as race, ethnicity, gender, and their close correlates—from an algorithm's inputs, to requiring certain key metrics, like error rates, be equal across demographic groups.

Today, adherence to these fairness constraints has become common practice in the design of algorithms across many contexts. However, a dogmatic implementation of these constraints often comes at the cost of inflicting additional burdens on individuals in all groups, including those in marginalized communities. For instance, in medicine, common diabetes risk calculators that ignore a patient's race and ethnicity systematically underestimate diabetes risk for Asian, Hispanic, and Black patients and overestimate diabetes risk for White patients (1). There may well be good reasons to exclude the use of race in medical diagnoses—e.g., to guard against pernicious attitudes of biological determinism—but this constraint comes at the cost of poorer treatment for patients in every group. In Section 2 we lay out this tension between fairness constraints and welfare in more detail. Then, in Section 3, we make several recommendations to address this tension between fair processes and fair outcomes, as well as other problems commonly encountered when building algorithms. We hope our discussion helps researchers, policymakers, and practitioners understand the subtleties of popular fairness constraints, and leads to the design of more equitable algorithms.

2 Popular fairness constraints and their consequences

Over the last several years, researchers across numerous fields have considered the equitable design of algorithms, including in computer science and statistics (13; 16; 20; 23–25; 27; 29; 36; 43; 50; 59; 61; 66; 71; 76; 79; 80; 93; 95; 96; 98; 99; 101; 102), law (8; 19; 52; 55; 60; 74; 97), medicine (46; 75;

82; 85; 86), the social sciences (9; 30; 48; 56; 57; 62; 70; 81), and philosophy (15; 54; 58). Many of these studies have proposed formal statistical principles for designing "fair" algorithms. Here we group these myriad fairness principles into three conceptual categories:

1. *Blinding*, in which one limits the effects of demographic attributes—like race—on decisions;
2. *Equalizing decision rates* across demographic groups; and
3. *Equalizing error rates* across demographic groups.

To many, these principles represent intuitively appealing understandings of fairness, and they have been applied to a variety of contexts in which algorithms guide decisions. They are often implemented with the explicit goal of protecting members of disadvantaged communities, but, as we discuss next, strict adherence to these principles often leads to worse outcomes for those in marginalized groups—and society as a whole (28).

To illustrate with a practical example, consider the case of diabetes risk estimation. Approximately one in ten Americans suffer from Type 2 diabetes, which can lead to other serious health problems, including heart disease, kidney disease, and vision loss. Upon learning of their diagnosis, patients can better manage their condition—for example, through changes in diet and physical activity—making early detection critical to improving health outcomes. In theory, every patient could be screened at regular intervals in an effort to detect diabetes early. But screening itself comes with monetary and non-monetary costs (e.g., patients may need to take time off from work, resulting in lost income). The medical community accordingly recommends that only those with at least a moderate risk of developing diabetes undergo screening. For example, results by Aggarwal et al. (1) suggest that patients will typically benefit from screening if their risk of diabetes is above 1.5%. To follow this recommendation, statistical risk algorithms can be used to estimate the diabetes risk for every patient, offering screening to those with an estimated risk above 1.5%.

We empirically ground our discussion by training statistical models that estimate diabetes risk using data from the National Health and Nutrition Examination Survey (NHANES). NHANES combines interview responses with laboratory data to provide insight into the health and nutritional status of adults and children in the U.S. The survey is conducted every two years by the National Center for Health Statistics and is frequently used by researchers to assess the prevalence of major diseases and their risk factors across the U.S. population. In our analysis, we use the four NHANES cycles from 2011–2018. Following Aggarwal et al. (1), we restricted our sample to 18,000 patients who were not pregnant, were 18–70 years old, and had a BMI between 18.5 kg/m^2 and 50.0 kg/m^2.

We now discuss the three fairness constraints above, in turn showing how statistical risk algorithms that adhere to each constraint may lead to worse outcomes for minority and majority groups alike.

The consequences of blinding

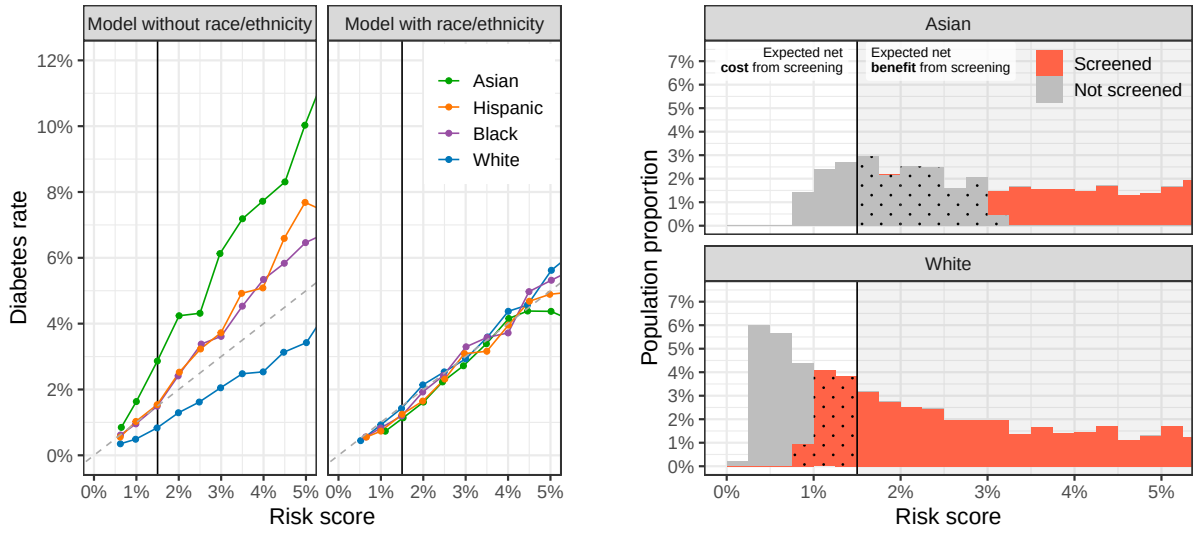
In an attempt to limit the effects of demographic attributes on risk assessments, the principle of blinding mandates that algorithms not have access to certain demographic characteristics, such as race or ethnicity, when estimating patient risk.[1] For example, diabetes risk may be estimated by a statistical risk algorithm that considers one's age and body mass index (BMI), but not their race or ethnicity. This principle is also sometimes called "fairness through unawareness."

In Figure 1(a), we compare diabetes risk estimated by models that either exclude (left panel) or include (right panel) information on a patient's race and ethnicity against empirical rates of diabetes prevalence. The model that is blind to a patient's race and ethnicity systematically underestimates diabetes risk for Asian, Black, and Hispanic patients, while it systematically overestimates diabetes risk for White patients—a problem that does not occur in the model that considers race and ethnicity.

Under the blind model, the "miscalibrated" risk scores could in turn result in erroneous screening decisions for some patients. Imagine, for concreteness, a hypothetical 30-year-old Asian patient with a BMI of 21.5 kg/m^2. Under the blind model, our hypothetical patient would have an estimated diabetes risk of approximately 1.15%, and so would not be screened based on the 1.5% screening threshold. However, as shown in the left panel of Figure 1(a), Asian patients with a nominal, race-blind diabetes risk of 1.15% have an empirical rate of diabetes close to 2%. We accordingly expect our hypothetical patient to benefit from screening, even though the race-blind model would recommend against it. The race-aware model, in contrast, correctly estimates that patients like this have an elevated risk of diabetes and thus recommends they be screened.

Analogously, consider a hypothetical 40-year-old White patient with a BMI of 20.5 kg/m^2. The race-blind model estimates our hypothetical patient has a 1.9% risk of diabetes, but, in reality, only 1.3% of patients like this have diabetes. The race-blind model would recommend our patient be screened, even though we expect screening to impose a net cost in this case. As before, the race-aware model correctly estimates that patients like this have a relatively low risk of diabetes and thus advises against screening.

[1] A related family of causal fairness criteria seeks to reduce both the direct and indirect effects of race on decisions (16), since even if algorithms are formally "blind" to race, race may still impact decisions indirectly through the algorithm's other inputs. This line of work, however, suffers from at least two serious limitations. First, it requires formalizing a causal effect of race, a long-standing statistical and conceptual problem replete with challenges that are succinctly captured by the mantra, "no causation without manipulation" (53). In particular, one must make sense of counterfactuals in which a person's race is altered (42; 47; 54; 88), a notion that is difficult, and perhaps impossible, to make precise. Second, recent mathematical results have shown that these causal fairness definitions constrain algorithms so severely that they often produce unintended results (28; 80). For example, under one prominent causal fairness definition, the only permissible algorithm in many situations is one that makes the same decision for every individual, irrespective of their risk factors.



(a) *Estimated vs. actual diabetes risk under models that exclude (left) or use (right) race and ethnicity.*

(b) *The relative cost of excluding race and ethnicity from diabetes risk estimates.*

Figure 1: *The plots in (a) compare the estimated risk from race-blind and race-aware models against the observed rates of diabetes across demographic groups. The plots in (b) illustrate the cost of using a race-blind model for Asian and White patients when compared against (more accurate) race-aware risk scores. The plot shows the distribution of race-aware risk scores for Asian and White patients, and the shaded areas show which patients would receive screening under the race-blind model. The dotted area, in particular, covers patients for whom the race-blind model makes a screening error, either because it fails to recommend Asian patients for screening (even though they would expect to benefit from a test) or because it recommends White patients for screening (when they would not expect to benefit from a test).*

Generally, if diabetes risk is estimated without the use of race and ethnicity, some non-White patients expected to benefit from screening would be counseled against screening, while some White patients expected to incur a net cost from screening would be screened anyway. On the other hand, models that use a patient's race and ethnicity are able to account for differences in diabetes risk across groups, and so do not make these systematic errors.

In Figure 1(b), we show the overall consequence of banning race and ethnicity from the algorithmic inputs on the two groups with the largest disparity, Asian and White patients. When avoiding the use of race and ethnicity, nearly 14% of Asian patients would not receive a screening even though they would be expected to benefit from it. Similarly, about 9% of White patients would be screened for diabetes even though they would be expected to incur a net cost from screening.

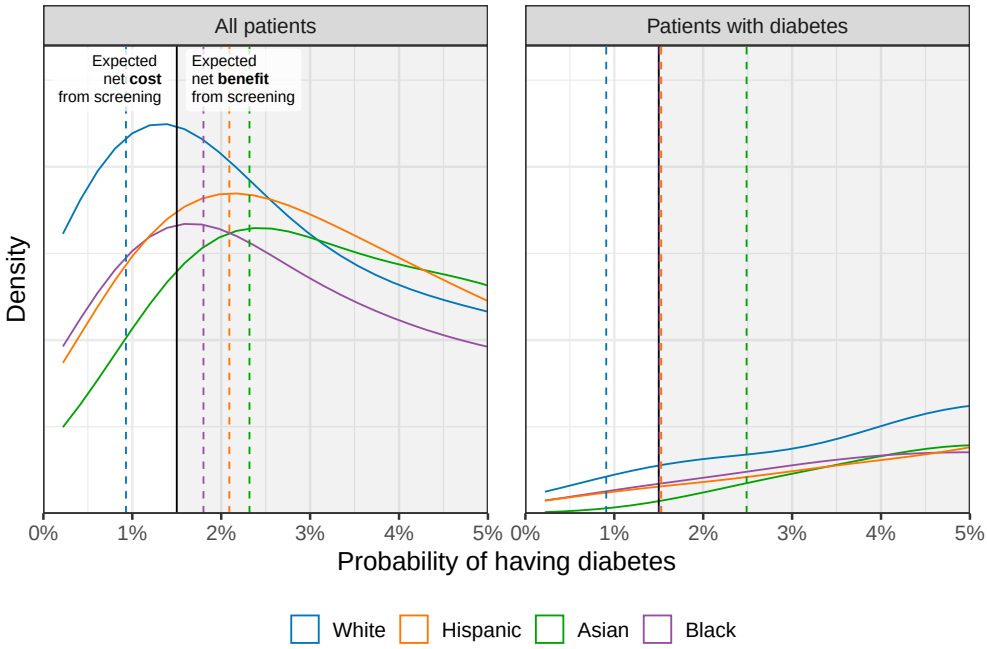


Figure 2: *The distribution of diabetes risk for all patients (left) and patients with diabetes (right). Estimates of risk were generated using patients' age, BMI, and race or ethnicity. The dashed vertical lines correspond to screening thresholds that equalize decision rates across groups (left) and false negative rates across groups (right).*

The consequences of equalizing decision rates

A second common fairness constraint requires that algorithmic decisions be made at equal rates across demographic groups—defined, for example, by race and ethnicity. For instance, a policy following this constraint might enforce that the proportion of White patients who are recommended for diabetes screening is approximately equal to the proportion of Asian patients recommended for screening. As with blinding, equalizing decision rates may feel intuitively appealing. But—also as with blinding—equalizing decision rates can likewise impose considerable costs on members of every group, due to the "problem of infra-marginality" (4; 6; 17; 37; 44; 63; 87; 90).

Returning to our running diabetes example, consider risk scores that are based on a patient's age, BMI, and race or ethnicity. (Similar issues result if we start with the blind risk scores.) The left panel of Figure 2 shows the distribution of estimated risk scores, disaggregated by race and ethncity. In this case, 76% of White patients have risk scores above the 1.5% screening threshold (indicated by the vertical black line), but 93% of Asian patients, 90% of Hispanic patients, and 88% of Black patients are above the threshold. As a result, if we make the optimal decision for each individual patient—screening them if their likelihood of having diabetes is above 1.5%—we would violate the principle of equalizing decision rates.

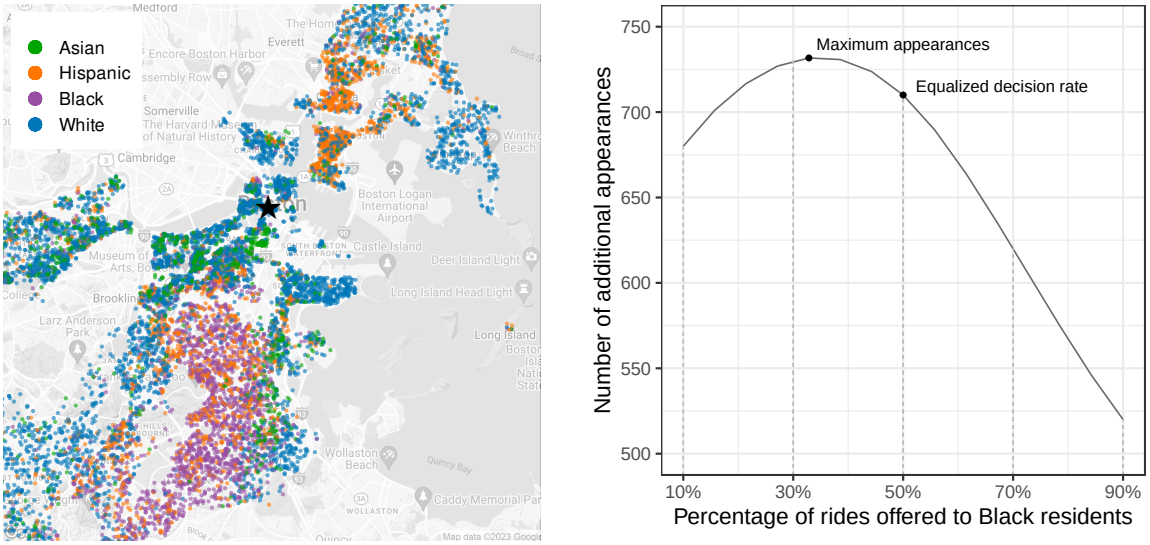
To equalize screening rates across racial and ethnic groups, we could set group-specific screening thresholds. Under a single, non-group-specific screening threshold of 1.5%, approximately 85% of individuals are screened. To equalize screening rates, we could similarly choose to screen the riskiest 85% of each group. The vertical lines in the left panel of Figure 2 show the corresponding group-specific screening thresholds for this policy. Under this approach, we would screen White patients with a risk score of approximately 1% or above, which includes many relatively low risk White patients—namely those with risk between 1% and 1.5%—for whom we expect screening to impose net costs. Conversely, we would only screen Asian patients who have relatively high risk of diabetes, above approximately 2.5%. In this case, we would fail to screen many Asian patients for whom we expect screening to have net benefits. By equalizing decisions rates, we thus harm members of all racial and ethnic groups.

The consequences of equalizing error rates

A third popular class of fairness constraints requires that error rates be equal across groups. In the context of our running example, one might, for example, demand that the false negative rate of screening decisions be the same across racial and ethnic groups. This constraint means that among those who in reality have diabetes, the proportion who are not screened is the same across groups. As with blinding and equalizing decision rates, equalizing error rates has intuitive appeal, but, as with those other constraints, it can harm members of all groups.

In right panel of Figure 2, we show the distribution of estimated risk among those patients who have diabetes, disaggregated by race and ethnicity. Under a policy of screening patients above a 1.5% threshold (indicated by the vertical black line), the false negative rate for a group corresponds to the area under that group's density curve that is to the left of the threshold. Specifically, White patients have a 1.7% false negative rate, meaning they would not be recommended for screening even though they have diabetes. In comparison, the false negative rate for Asian patients is less than 0.1%. As above, making the optimal screening decision for each patient would violate the principle of equalizing error rates.

Also as above, we could equalize false negative rates by setting group-specific screening thresholds. For example, if we screen White patients above a 0.9% threshold, and Asian patients above a 2.5% threshold, then both groups would have a false negative rate of 0.7%. But such a policy would mean that we recommend screening for some relatively low-risk White patients and do not recommend screening for some relatively high-risk Asian patients, harming some individuals in both groups.



(a) *Suffolk County, MA residential distribution by race and ethnicity.*

(b) *Trade-offs between maximizing court appearances and the demographic allocation of vouchers.*

Figure 3: *The map in (a) shows the geographic distribution of Suffolk County (Boston) residents, with the star marking the location of the main county courthouse, where most individuals would be required to appear for court appointments. In Suffolk County, White residents tend to live closer to the courthouse than Black residents. In (b), we illustrate the range of possible policy options to provide a hypothetical population of residents with a free ride to court, where each option maximizes appearances for a given distribution of vouchers. For example, if one wants 50% of vouchers to go to Black residents, given the existing budget, 710 additional people would get to court. Alternatively, under the same budget, a policy that aims to maximize appearances overall would allow 730 additional people to get to court and would have approximately 30% of rides offered to Black residents.*

Trade-offs in resource-constrained settings

Risk-based screening for diabetes is a setting where policy choices are not constrained by resources—it is feasible to offer regular screening to every adult in the United States if that were determined to be medically advisable. In other scenarios, however, policy decisions have to be made under significant resource constraints, leading to inherent trade-offs that complicate the design of equitable algorithms. To illustrate, we transition from healthcare to the criminal-legal system, and consider the problem of increasing court appearance rates among individuals with upcoming court dates. In the United States, missed court dates typically prompt a judge to issue a "bench warrant"—mandating the individual be arrested when they next encounter law enforcement—which in turn can lead to days or even weeks in jail (40; 72). Such incarceration imposes high costs on individuals and their communities, including job loss and social stigma (33; 34; 39; 49; 69). Many people report missing their court appointments due to transportation barriers, and so one promising

proposal to improve appearance rates and reduce the resulting incarceration is to provide individuals with transportation vouchers (e.g., for public transit or ride-share services) (2; 11; 23). But these vouchers can be costly, and so such programs may not be able to provide transportation assistance to every individual who might benefit from it.

Given the budget constraint, policymakers implementing transportation-assistance programs face a difficult trade-off: on one hand, they will want to spend their budget strategically, to increase appearances as much as possible (and, accordingly, maximally reduce incarceration); on the other, they might also be interested in achieving a certain racial or ethnic balance among those who benefit from a voucher (23). Consider the case of Boston, Massachusetts, where Black individuals tend to live farther away from the courthouse than White individuals, as shown in Figure 3(a). Because the costs of ride-share vouchers increase with the distance traveled, the demographic distribution of residents across the Boston area implies that, all else being equal, it would be more expensive to provide rides to Black individuals than to White individuals. As a result, a program solely focused on maximizing appearance rates would see more of its funds go to White clients. If instead one were to insist on equalized decision rates (i.e., offering transportation assistance to an equal proportion of White and Black individuals), this would necessarily mean that fewer appearances can be achieved.

To make this trade-off more concrete, we describe the results of a simple, stylized simulation. Suppose that 5,000 White and 5,000 Black individuals in a fictional city have upcoming court dates. We imagine that it costs \$5 per mile to transport each individual from their home to the courthouse and back. But, as in Boston, our hypothetical Black individuals on average live farther from the courthouse than our hypothetical White individuals. Finally, we suppose that for each individual i, they would successfully make it to court if provided a ride-share voucher but, if not provided a voucher, would appear with a known probability p_i—estimated, for example, with a model trained on historical court appearance data (23). Assuming we have a transportation budget of \$10,000, Figure 3(b) shows how the number of additional court appearances varies with the demographic allocation of vouchers, where each point on the curve corresponds to an allocation strategy that maximizes the number of appearances while ensuring a certain demographic composition of recipients. Among these policy options, there is no one "correct" choice. The *best* choice will depend on one's preference for trading off the total number of court appearances with the distribution of vouchers across Black and White individuals, an idea we discuss more below. For now, we note that certain formal fairness constraints—e.g., requiring an equal proportion of White and Black individuals receive vouchers—represent but one among several options to make that trade-off.

3 A path forward

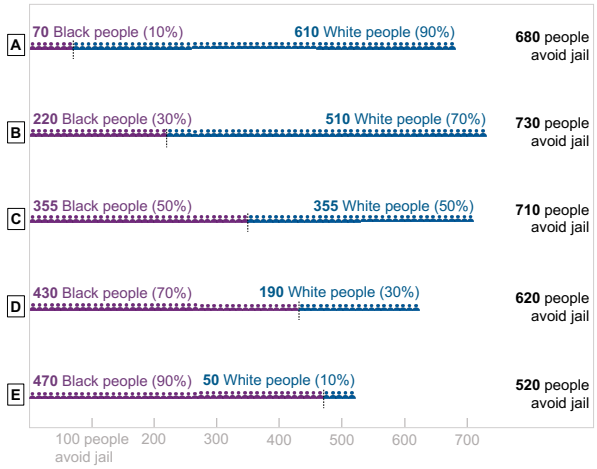
Our diabetes example suggests that adherence to popular fairness constraints often comes at the cost of inflicting additional burden on individuals, including those in marginalized groups. In our ride-share example, where resources are limited, imposing formal fairness constraints can likewise result in allocation policies that do not reflect the preferences of stakeholders. This tension highlights the need for a robust discussion about the specific way in which proposed fairness constraints are connected to the inherently normative concept of equity.

On one hand, the constraints could be understood as an *end* in and of itself. This conceptualization of fairness constraints is consistent with a deontological account of ethical decision-making, which postulates that an ethical decision is one that adheres to universally applicable, moral rules. Under this view, fairness criteria establish an outcome-independent set of constraints that should be imposed for their own sake, without regard to the specific results in a particular context.

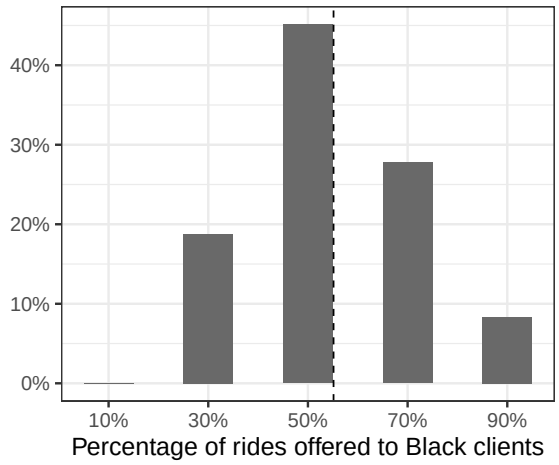
A competing understanding of fairness constraints treats them not as an *end*, but as a *means* to achieve equitable, algorithmic decisions. Conceiving of constraints in this way is consistent with a *consequentialist* account of ethical decision-making, whereby the morality of a decision is defined not by its dogmatic adherence to a set of rules, but by the outcomes it achieves. Under a consequentialist view, popular fairness constraints merely act as potentially useful heuristics to achieve desirable outcomes. However, if it can be demonstrated that a particular constraint imposes net burdens on marginalized groups, or society more generally, a consequentialist conception would counsel against dogmatic adherence to the constraint, as it would not achieve its desired goal of furthering equity in that case.

Although fairness constraints are frequently promoted and implemented, deeper discussions of their normative underpinnings are almost entirely absent from the literature (for a rare example, see Card and Smith (15)). In principle, treating fairness constraints as a dogmatic principle or as a useful heuristic appear defensible. In our example of estimating diabetes risk, advocates in favor of race-blind tools may, for instance, argue that it is inherently wrong to make decisions for an individual based on their (immutable) group membership, or that race-based decision-making reinforces damaging beliefs about inherent differences between individuals of different racial groups. And perhaps some believe that these concerns outweigh the negative health effects that patients—particularly Asian patients—might experience under the race-blind algorithm. But irrespective of one's particular preferences, we believe that a more considered discussion and explicit acknowledgement of the potential cost of these constraints is necessary to avoid inflicting accidental harms.

We now conclude our discussion by considering four technical and policy-related aspects of algorithm design that we believe are critical to building more equitable tools: (1) grappling with the inherent trade-offs that we highlighted above; (2) checking the calibration of predictive models; (3) judiciously selecting the target of prediction; and (4) appropriately collecting training data.



(a) *Options presented to survey participants, which involve a trade-off between minimizing overall incarceration and incarceration for Black people.*



(b) *Survey responses corresponding to the options in Figure 4(a), with the dashed vertical line at 55% indicating the average response.*

Figure 4: *Preferences for allocating ride-share vouchers to a hypothetical population of residents depicted in Figure 3(b). In (a), we show the options presented to survey respondents. Responses to this question are shown in (b).*

Grappling with inherent trade-offs

One way to make salient and arbitrate between the competing aspects of equity we describe above is to elicit the preferences of stakeholders. To illustrate this approach in our example of allocating ride-share vouchers, we designed and administered a poll to a diverse sample of Americans (23; 65).[2] Mirroring our simulation above, survey respondents were introduced to a hypothetical jurisdiction with an equal proportion of Black and White residents, with Black residents, on average, living farther from the courthouse than White residents. We then asked respondents to state how they would balance appearance rates (and, accordingly, incarceration for missed court appoitments) with the demographic distribution of transportation assistance. To aid in their decision, participants were shown the graphic depicted in Figure 4(a).

The results of the survey are shown in Figure 4(b), and reveal that the respondents have highly heterogenous preferences. Most frequently, respondents prefer a policy that mirrors the demographics of the underlying population (Option C), but many respondents prefer a different balance of "efficiency" and "demographic balance", favoring policies that shift more resources towards Black individuals. Only 19% of respondents prefer the efficiency-maximizing allocation policy that minimizes overall incarceration (Option B).

[2]We ran our survey on the Prolific platform. By selecting the platform's "representative sample" option, the distribution of self-identified sex, age, and ethnicity in our sample matched the distribution in the U.S. Census. The survey resulted in 144 respondents.

Preferences elicited in this way are but one input into complex policy decisions. Further, while we surveyed a diverse sample of Americans, identifying the relevant stakeholders is itself a difficult problem, defying general prescriptions. We hope, though, that this simple exercise demonstrates the feasibility of productively grappling with the thorny trade-offs at the heart of many policy design problems.

Checking calibration across groups

As its name suggests, the primary objective of a risk assessment algorithm is to accurately estimate risk. Without care, however, it is common for statistical algorithms to systematically overestimate risk for some groups and underestimate risk for others—a problem that is also referred to as "miscalibration". For example, in Figure 1(a) (left), a 1% estimated risk of diabetes corresponds to an observed diabetes rate of about 0.5% for White individuals but about 1.6% for Asian individuals. These miscalibrated risk scores can lead to over-screening White patients and under-screening Asian patients, resulting in worse outcomes for individuals in both groups. Similarly, gender-blind risk assessment tools commonly used in the criminal justice system to predict recidivism tend to systematically overestimate risk for women and underestimate risk for men (91). These miscalibrated estimates can in turn lead to incarcerating women who are much less likely to recidivate than their risk scores suggest.

Miscalibrated risk scores can typically be corrected by including group membership (e.g., race or gender) as a risk factor in the predictive model. Doing so, however, can run afoul of legal restrictions—for example, explicit considerations of race often receive heightened legal scrutiny in the United States, a standard that is notoriously difficult to satisfy. And even when legally permissible, race/ethnicity- and gender-aware algorithms may not be socially or politically acceptable. As we discussed in the context of our running diabetes example, the use of race or ethnicity could reinforce pernicious attitudes about inherent differences between racial and ethnic groups, potentially outweighing the benefits of including these features in the models. Regardless of whether one ultimately decides to include race, ethnicity, gender, or other sensitive attributes in risk assessment algorithms, we believe it is important to assess whether the risk estimates are calibrated (like we do in Figure 1(a)) in order to better understand the costs and benefits of algorithm design choices.

Selecting the target of prediction

Even when including information about group membership, it is still possible to have miscalibrated estimates if the labels available in the data are an imperfect representation of the true target of the prediction. This occurrence is known as *label bias* (28; 100). For our running diabetes example, the NHANES data provides two pieces of information that we can use to construct a label for whether a patient has diabetes: (1) the results of a blood test administered to the entire survey population;

and (2) whether the patient has ever received a diabetes diagnosis from a doctor. A diabetes label based solely on receiving a doctor's diagnosis is often tied to how regularly a patient is seen by a doctor. But one can imagine that the frequency of seeing a doctor varies considerably across a number of dimensions. For instance, given the same health-related attributes, Black patients tend to go to primary care physicians less often than White patients (5). And because the probability of detecting diabetes increases with the frequency with which a patient goes to the doctor, a dataset that records the presence of diabetes using a doctor's assessment might systematically under-record the presence of diabetes in racial minorities. Consequently, a model trained on such a dataset would systematically underestimate the *true* diabetes risk of racial minorities, even if it is well calibrated to predicting a doctor's diabetes diagnosis.

Figure 5 lends empirical support to these theoretical concerns by illustrating the negative effects of using a doctor's diagnosis as a proxy for the presence of diabetes. It shows the risk scores of a model trained to predict the proxy (doctor's diabetes diagnosis) using a patient's age, race/ethnicity, and BMI. Yet, in spite of supplying the model with data on group membership—race and ethnicity, in this case—the resulting risk scores are still miscalibrated when compared against the true label (blood test *or* doctor diagnosis) for each group. The risk scores produced using the proxy underestimate risk for all groups, but miscalibration is especially pronounced for Black, Asian, and Hispanic patients. This pattern may in part stem from known racial disparities in healthcare access, with patients from minority racial groups less likely to have received a diabetes diagnosis from their doctor. The miscalibrated risk scores could lead to under-screening of patients from *all* groups, but most severely for racial and ethnic minorities.

Unfortunately, one's ability to mitigate the effects of label bias is often heavily constrained by the data collection process and availability. One possible solution is to adjust the target of prediction by focusing on or leveraging other outcomes that are less likely to exhibit bias. In our running diabetes example, we accomplish this by constructing a diabetes label using *both* the blood test results provided in the survey *and* whether the patient had ever received a diabetes diagnosis from a doctor. Because the blood test was administered to the entire population, using this information to construct the diabetes label fills in informational gaps that arise from relying solely on past diabetes diagnoses from a doctor. Countering label bias is challenging and there is rarely a perfect solution, but when designing algorithmic tools for decision-making, it is imperative to scrutinize the data variables used— especially that of the target of interest—for potential sources of bias and mitigate downstream effects to the greatest extent possible.

Collecting training data

Finally, we consider the role of training data in equitable algorithm design. As a general heuristic, we advise to train algorithms on data that are representative of the population to which the algorithms will be applied. Failure to do so can lead to starkly inequitable outcomes. For example,

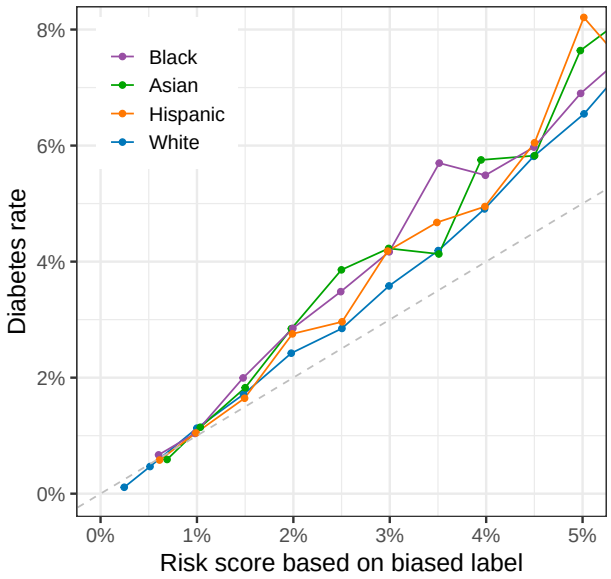


Figure 5: *In contrast to the risk scores presented in Figure 1(a) which predict the prevalence of diabetes from blood tests and doctors' diabetes diagnoses, here the models are only trained to predict a doctor's diabetes diagnosis. The model inputs are age, race/ethnicity and BMI. Likely due in part to racial and ethnic disparities in healthcare access, predicting a doctor's diabetes diagnosis introduces bias into the model when compared against the results from the combined label. We observe that Asian, Black, and Hispanic Americans have higher true diabetes risk than White Americans with the same nominal risk under the model.*

in an analysis of automated speech recognition systems, Koenecke et al. (64) found that state-of-the-art models made twice as many errors transcribing Black speakers than they did for White speakers, a disparity that likely stemmed from a relative sparsity of speech data from Black speakers used to train the models. Similarly, in an analysis of image analysis tools, Buolamwini and Gebru (13) found that popular algorithms performed significantly worse at classifying the gender of dark-skinned individuals compared to light-skinned individuals, likely due to a lack of dark-skinned faces in the training data—and the performance was worst for dark-skinned women.[3]

Training algorithms with representative data is a useful starting point, but, like in many other contexts we discuss here, there are subtleties to consider. For example, given a limited budget, the optimal data collection strategy depends on the statistical structure of the underlying population, the cost of collecting data from different subgroups, and the relative value of model performance across subgroups (14). In particular, if the connection between features and outcomes is similar across subgroups, one might trade representativeness for more data from the subgroups for whom

[3]In both of these examples, error rates by race and skin tone are a useful metric for auditing the algorithms because one would not generally expect the difficulty of transcribing speech or identifying gender to vary substantially across these categories. In contrast, in many risk assessment settings like our diabetes example, we expect risk distributions to differ across racial/ethnic and gender groups, limiting the diagnostic value of comparing error rates.

data acquisition is less costly. Conversely, if certain groups have idiosyncratic statistical properties, one might choose to oversample from them. In short, and in line with our general philosophy, it is important to carefully consider the trade-offs inherent to different data collection strategies.

4 Conclusion

With the proliferation of algorithmically guided decision-making in healthcare, the criminal-legal system, banking, and beyond, there is increasing need to ensure that algorithms are fair. A plethora of formal fairness metrics and design principles have been proposed in recent years, particularly in the computer science community. But, as we have argued here, popular approaches to fairness often lead to worse outcomes for individuals, including those from marginalized communities. In some cases, the conflict between formal fairness constraints and equitable outcomes suggests shortcomings of the constraints themselves. For instance, in our running diabetes example, it seems difficult to justify equalizing error rates on consequentialist grounds. In other cases, though, the tension is harder to allay. Diabetes risk estimates that consider race and ethnicity may lead to more accurate screening decisions, but such non-blind algorithms might also validate and encourage insidious beliefs in inherent differences across groups, with possible negative repercussions for marginalized patients, many of whom already have significant distrust in the U.S. healthcare system (5; 10). There are often no easy answers to these difficult trade-offs, but we hope our discussion equips researchers, practitioners, and policymakers to make more informed choices.

Acknowledgements

We thank Sam Corbett-Davies, Johann Gaebler, Avi Feller, David Kent, Keren Ladin, Hamed Nilforoshan, and Ravi Shroff for helpful conversations. We draw in this paper from a more technical exposition of algorithmic fairness by Corbett-Davies et al. (28). Our work was supported by grants from the Harvard Data Science Institute, the Stanford Impact Labs, and Stanford Law School. Code to reproduce our analysis is available at: https://github.com/madisoncoots/equitable-algorithms.

References

[1] Rahul Aggarwal, Kirsten Bibbins-Domingo, Robert W. Yeh, Yang Song, Nicholas Chiu, Rishi K. Wadhera, Changyu Shen, and Dhruv S. Kazi. Diabetes screening by race and ethnicity in the United States: Equivalent body mass index and age thresholds. *Annals of Internal Medicine*, 175(6):765–773, 2022.

[2] Sophie Allen. Illegible courts and failure to appear. Working paper, 2023.

[3] Maxwell Allman, Itai Ashlagi, Irene Lo, Juliette Love, Katherine Mentzer, Lulabel Ruiz-Setz, and Henry O'Connell. Designing school choice for diversity in the San Francisco Unified School District. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 290–291, 2022.

[4] Shamena Anwar and Hanming Fang. An alternative test of racial prejudice in motor vehicle searches: Theory and evidence. *The American Economic Review*, 2006.

[5] MJ Arnett, Roland J Thorpe, DJ Gaskin, Janice V Bowie, and Thomas A LaVeist. Race, medical mistrust, and segregation in primary care as usual source of care: findings from the exploring health disparities in integrated communities study. *Journal of Urban Health*, 93: 456–467, 2016.

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

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

[8] Jason R Bent. Is algorithmic affirmative action legal. *Georgetown Law Journal*, 108:803, 2019.

[9] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50 (1):3–44, 2021.

[10] L Ebony Boulware, Lisa A Cooper, Lloyd E Ratner, Thomas A LaVeist, and Neil R Powe. Race and trust in the health care system. *Public health reports*, 118(4):358, 2003.

[11] Rebecca Brough, Matthew Freedman, Daniel E Ho, and David C Phillips. Can transportation subsidies reduce failures to appear in criminal court? evidence from a pilot randomized controlled trial. *Economics Letters*, 216:110540, 2022.

[12] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In

Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2019.

[13] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[14] William Cai, Ro Encarnacion, Bobbie Chern, Sam Corbett-Davies, Miranda Bogen, Stevie Bergman, and Sharad Goel. Adaptive sampling strategies to construct equitable training datasets. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2022.

[15] Dallas Card and Noah A Smith. On consequentialism and fairness. *Frontiers in Artificial Intelligence*, 3:34, 2020.

[16] Alycia N Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in Big Data*, 5, 2022.

[17] James H Carr, Isaac F Megbolugbe, et al. *The Federal Reserve Bank of Boston study on mortgage lending revisited*. Fannie Mae Office of Housing Policy Research, 1993.

[18] Lindsay Cattell and Julie Bruch. Identifying students at risk using prior performance versus a machine learning algorithm. Technical report, U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Mid-Atlantic, 2021. (REL 2021–126).

[19] Anupam Chander. The racist algorithm. *Michigan Law Review*, 115:1023, 2016.

[20] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[21] Alex Chohlas-Wood and ES Levine. A recommendation engine to aid in identifying crime patterns. *INFORMS Journal on Applied Analytics*, 49(2):154–166, 2019.

[22] Alex Chohlas-Wood, Joe Nudell, Keniel Yao, Zhiyuan Lin, Julian Nyarko, and Sharad Goel. Blind justice: Algorithmically masking race in charging decisions. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 35–45, 2021.

[23] Alex Chohlas-Wood, Madison Coots, Emma Brunskill, and Sharad Goel. Learning to be fair: A consequentialist approach to equitable decision-making. *arXiv preprint arXiv:2109.08792*, 2023.

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

[25] Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89, 2020.

[26] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, 2018.

[27] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 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, 2017.

[28] Sam Corbett-Davies, Johann Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *arXiv preprint arXiv:1808.00023*, 2023.

[29] Amanda Coston, Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 582–593, 2020.

[30] Bo Cowgill and Catherine E Tucker. Economics, fairness and algorithmic bias. *In preparation for: Journal of Economic Perspectives*, 2019.

[31] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[32] Matthew DeMichele, Peter Baumgartner, Michael Wenger, Kelle Barrick, Megan Comfort, and Shilpi Misra. The Public Safety Assessment: A re-validation and assessment of predictive utility and differential prediction by race and gender in Kentucky, 2018. URL `https://papers.ssrn.com/abstract=3168452`.

[33] Stephanie Holmes Didwania. The immediate consequences of federal pretrial detention. *American Law and Economics Review*, 22(1):24–74, 2020.

[34] Will Dobbie, Jacob Goldin, and Crystal S. Yang. The effects of pretrial detention on conviction, future crime, and employment: Evidence from randomly assigned judges. *American Economic Review*, 108(2):201–40, February 2018. doi: 10.1257/aer.20161503. URL `http://www.aeaweb.org/articles?id=10.1257/aer.20161503`.

[35] Mitchell L Doucette, Christa Green, Jennifer Necci Dineen, David Shapiro, and Kerri M Raissian. Impact of shotspotter technology on firearm homicides and arrests among large metropolitan counties: a longitudinal analysis, 1999–2016. *Journal of urban health*, 98(5): 609–621, 2021.

[36] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[37] Robin S Engel and Rob Tillyer. Searching for equilibrium: The tenuous nature of the outcome test. *Justice Quarterly*, 25(1):54–71, 2008.

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

[39] Keith Finlay, Michael Mueller-Smith, and Brittany Street. Children's indirect exposure to the us justice system: Evidence from longitudinal links between survey and administrative data. Working paper, 2023.

[40] Alissa Fishbane, Aurelie Ouss, and Anuj K Shah. Behavioral nudges reduce failure to appear for court. *Science*, 370(6517):eabb6591, 2020.

[41] John J Friedewald, Ciara J Samana, Bertram L Kasiske, Ajay K Israni, Darren Stewart, Wida Cherikh, and Richard N Formica. The kidney allocation system. *Surgical Clinics*, 93 (6):1395–1406, 2013.

[42] Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. A causal framework for observational studies of discrimination. *Statistics and Public Policy*, 2022.

[43] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Causal feature selection for algorithmic fairness. *Proceedings of the 2022 International Conference on Management of Data (SIGMOD)*, 2022.

[44] George C Galster. The facts of lending discrimination cannot be argued away by examining default rates. *Housing Policy Debate*, 4(1):141–146, 1993.

[45] Sharad Goel, Ravi Shroff, Jennifer Skeem, and Christopher Slobogin. The accuracy, equity, and jurisprudence of criminal risk assessment. In *Research handbook on big data law*, pages 9–28. Edward Elgar Publishing, 2021.

[46] Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 169(12):883–884, 2018.

[47] D James Greiner and Donald B Rubin. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785, 2011.

[48] Nina Grgić-Hlača, Gabriel Lima, Adrian Weller, and Elissa M. Redmiles. Dimensions of diversity in human perceptions of algorithmic fairness. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22, New York, NY, USA,

2022. Association for Computing Machinery. doi: 10.1145/3551624.3555306. URL https://doi-org.stanford.idm.oclc.org/10.1145/3551624.3555306.

[49] Arpit Gupta, Christopher Hansman, and Ethan Frenchman. The heavy costs of high bail: Evidence from judge randomization. *The Journal of Legal Studies*, 45(2):471–505, 2016. doi: 10.1086/688907. URL https://doi.org/10.1086/688907.

[50] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29:3315–3323, 2016.

[51] Deborah Hellman. Measuring algorithmic fairness. *Virginia Law Review*, 106(4):811–866, 2020.

[52] Daniel E Ho and Alice Xiang. Affirmative algorithms: The legal grounds for fairness as awareness. *University of Chicago Law Review Online*, page 134, 2020.

[53] Paul W Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[54] Lily Hu and Issa Kohler-Hausmann. What's sex got to do with machine learning? In *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 2020.

[55] Aziz Huq. Racial equity in algorithmic criminal justice. *Duke Law Journal*, 68, 2019.

[56] Kosuke Imai and Zhichao Jiang. Principal fairness for human and algorithmic decision-making. *arXiv preprint arXiv:2005.10400*, 2020.

[57] Kosuke Imai, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of algorithm-assisted human decision-making: Application to pretrial public safety assessment. *arXiv preprint arXiv:2012.02845*, 2020.

[58] Maximilian Kasy and Rediet Abebe. Fairness, equality, and power in algorithmic decision-making. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 576–586, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445919. URL https://doi.org/10.1145/3442188.3445919.

[59] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 656–666, 2017.

[60] Pauline T Kim. Race-aware algorithms: Fairness, nondiscrimination and affirmative action. *California Law Review*, 110:1539, 2022.

[61] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*, 2017.

[62] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10:113–174, 2018.

[63] John Knowles, Nicola Persico, and Petra Todd. Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109(1), 2001.

[64] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14): 7684–7689, 2020.

[65] Allison Koenecke, Eric Giannella, Robb Willer, and Sharad Goel. Popular support for equity in algorithmic decision-making. Working paper, 2023.

[66] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[67] Anja Lambrecht and Catherine Tucker. Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads. *Management science*, 65(7): 2966–2981, 2019.

[68] Martin Leo, Suneel Sharma, and Koilakuntla Maddulety. Machine learning in banking risk management: A literature review. *Risks*, 7(1):29, 2019.

[69] Emily Leslie and Nolan G. Pope. The unintended impact of pretrial detention on case outcomes: Evidence from new york city arraignments. *The Journal of Law and Economics*, 60 (3):529–557, 2017. doi: 10.1086/695285. URL `https://doi.org/10.1086/695285`.

[70] Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 58–59, 2022.

[71] Joshua R Loftus, Chris Russell, Matt J Kusner, and Ricardo Silva. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*, 2018.

[72] Barry Mahoney, Bruce D Beaudin, John A Carver III, Daniel B Ryan, and Richard B Hoffman. Pretrial services programs: Responsibilities and potential. *National Institute of Justice*, 2001.

[73] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[74] Sandra G Mayson. Bias in, bias out. *The Yale Law Journal*, 128(8):2218–2300, 2019.

[75] Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, and James A Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2(5):e221–e223, 2020.

[76] Vishwali Mhasawade and Rumi Chunara. Causal multi-level fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 784–794, 2021.

[77] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8(1):141–163, 2021. doi: 10.1146/annurev-statistics-042720-125902.

[78] George O Mohler, Martin B Short, Sean Malinowski, Mark Johnson, George E Tita, Andrea L Bertozzi, and P Jeffrey Brantingham. Randomized controlled field trials of predictive policing. *Journal of the American statistical association*, 110(512):1399–1411, 2015.

[79] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[80] Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pages 16848–16887. PMLR, 2022.

[81] Julian Nyarko, Sharad Goel, and Roseanna Sommers. Breaking taboos in fair machine learning: An experimental study. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021.

[82] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[83] Cathy O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

[84] James O'Neill. How facial recognition makes you safer. *New York Times*, 9, 2019.

[85] Jessica K Paulus and David M Kent. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digital Medicine*, 3(1):1–8, 2020.

[86] Stephen R Pfohl, Agata Foryciarz, and Nigam H Shah. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of biomedical informatics*, 113:103621, 2021.

[87] Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In *The 21st International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[88] Maya Sen and Omar Wasow. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19(1):499–522, 2016.

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

[90] Camelia Simoiu, Sam Corbett-Davies, and Sharad Goel. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

[91] Jennifer Skeem, John Monahan, and Christopher Lowenkamp. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior*, 40(5):580, 2016.

[92] Till Speicher, Muhammad Ali, Giridhari Venkatadri, Filipe Nunes Ribeiro, George Arvanitakis, Fabrício Benevenuto, Krishna P. Gummadi, Patrick Loiseau, and Alan Mislove. Potential for discrimination in online targeted advertising. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 5–19. PMLR, 23–24 Feb 2018. URL `https://proceedings.mlr.press/v81/speicher18a.html`.

[93] Yixin Wang, Dhanya Sridhar, and David M Blei. Equal opportunity and affirmative action via counterfactual predictions. *arXiv preprint arXiv:1905.10870*, 2019.

[94] Bryan Wilder, Laura Onasch-Vera, Graham Diguiseppi, Robin Petering, Chyna Hill, Amulya Yadav, Eric Rice, and Milind Tambe. Clinical trial of an ai-augmented intervention for hiv prevention in youth experiencing homelessness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14948–14956, May 2021. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17754`.

[95] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.

[96] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-fairness: A unified framework for measuring causality-based fairness. *Advances in Neural Information Processing Systems*, 32, 2019.

[97] Crystal S Yang and Will Dobbie. Equal protection under algorithms: A new statistical and legal framework. *Mich. L. Rev.*, 119:291, 2020.

[98] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180, 2017.

[99] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 228–238, 2017.

[100] Michael Zanger-Tishler, Julian Nyarko, and Sharad Goel. Risk scores, label bias, and everything but the kitchen sink. Working paper, 2023.

[101] Junzhe Zhang and Elias Bareinboim. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[102] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3929–3935, 2017.

[103] Yan Zhang and Peter Trubey. Machine learning and sampling scheme: An empirical study of money laundering detection. *Computational Economics*, 54:1043–1063, 2019.