Fairness

Kate Vredenburgh (K.Vredenburgh@lse.ac.uk)
Department of Philosophy, Logic, and Scientific Methods, the London School of Economics

Abstract: Despite widespread agreement that algorithmic bias is a problem, there is a lack of agreement about what to do about it. In this chapter, I argue that what should be done about algorithmic bias depends on whether the problem of algorithmic bias is conceptualized as a problem of fairness, or some other problem of justice. I substantiate this claim by examining the debate over different formal fairness metrics. One compelling metric for measuring whether a system is fair measures whether the system is *calibrated,* or whether a prediction has equal evidential value regardless of an individual's group membership. Calibration exemplifies a compelling notion of accuracy, and of fairness, in treating like cases alike. However, there can be a tradeoff between making systems fair, in this sense, and making them more just: to make more accurate predictions, a system may utilize social patterns that reinforce structures of unjust disadvantage. If one is concerned about racial or gender oppression, say, then fairness in the sense of calibration is beside the point, and may even be harmful. In response to this tradeoff, I argue that in situations of injustice, other values of justice ought to be privileged over fairness, as fairness has no value in the absence of just background institutions. I conclude by drawing out five proposals for better governance of AI for justice and fairness from the philosophical discussion of fairness and justice in AI. These are: a values-first approach to bias interventions; de-coupling decision processes; explicitly modeling structural injustice; interventions to increase data quality; and the use of (weighted) lotteries, rather than decision thresholds.

Keywords: fairness, justice, AI, algorithmic bias, thresholds

1. Introduction

For many, algorithmic decision-making mediates their access to credit, education, and employment, how they access information, the healthcare or state benefits they receive, and their coercive treatment by the state. If properly governed, AI could contribute to more evidence-based, consistent decision-making in the institutions that seriously and unavoidably shape people's lives. But, as AI is seriously re-shaping our individual and collective lives, academic research, journalistic investigation, and user reports have raised serious, repeated concerns of *algorithmic bias,* or the deviation of an algorithm's performance from what is required by some standard.[1] "Algorithmic bias" can mean statistical bias, e.g., an algorithm's predictions deviate from the true state of the world, where this is

[1] Danks and Fazelpour 2021a.

detected using previously observed testing data. Here, I will be using "algorithmic bias" in a moral sense. Algorithmic bias is the encoding of wrongfully discriminatory social patterns into an algorithm.[2] This encoding usually occurs through the statistical regularities that the algorithm uses for its predictive or classificatory task.[3] Such bias has been demonstrated in facial recognition systems,[4] risk assessment tools in criminal justice,[5] tools to prioritize scarce healthcare resources and that mediate access to jobs[6], education[7], and credit[8] – to name a few recent and startling examples.

The problem of algorithmic bias is a difficult technical and philosophical problem. Bias is required to make any prediction, yet it also plagues any prediction problem – in making an inference from the past to the future, the decision-maker needs to make simplifying assumptions about what the world is like, but those simplifying assumptions can introduce inaccuracy into her predictions.[9] Bias is just as serious a concern when a machine is making the inferences. Models developed through machine learning, for example, are a powerful decision aids because they use past data to discover predictively powerful correlations between available data and the target variable of interest. But, because of a history of discriminatory institutions and social practices, social identity properties are often powerful predictors of the outcome of interest. We thus need to be concerned about how bias can enter into the AI systems that are used as decision aids, or that automatically execute decisions on the basis of their predictions. In Section 2, I expand on the problem of algorithmic bias, focusing on different places that bias can enter into the data science pipeline.

---

[2] I take "discrimination" to be a morally neutral term. I may discriminate between wrestlers by assigning them to groups based on their weight, or discriminate between students by organizing them by last name. These examples motivate that it is not always wrong to draw distinctions among people based on certain properties that they have (Hellman 2011). For purported instances of differential outcomes based on group membership, it is important first to ask whether such discrimination is wrongful.

[3] Definition based on Johnson 2021: 9942.

[4] Buolamwini and Gebru 2018.

[5] Angwin et. al. 2016.

[6] Ali et al. 2019.

[7] Guardian Editorial 11 August 2020.

[8] Apple, for example, faced legal scrutiny after the Apple Card granted lower credit limits to women than men (AI Now 2019).

[9] Dotan 2020; Johnson forthcoming.

Despite research, political scrutiny, and activism over the last decade, algorithmic bias remains trenchant. One likely cause of this trenchancy is inaction that favors elite interests. Elites and members of privileged groups have an interest in maintaining power and privilege, and structural and individual discrimination can be a means to that end.[10] Discrimination in education[11] and in the private sector[12] shapes who develops and researches technology, leaving places with the most money and prestige disproportionately white, male, and WEIRD (Western, Educated, Industrialized, Rich, Democratic).

However, there is another problem: people do not actually agree on what the problem of algorithmic bias is. And, until there is mutual understanding between those who disagree about the problem, we will continue to lack regulatory or industry benchmarks to determine when AI-based decision-making is wrongfully biased.

Some take the problem to be a matter of *justice*, or the moral norms that govern our basic societal institutions, including norms of distribution and norms of respect. Others take the problem to be a matter of *fairness*, or whether alike individuals are treated equally. Fairness is one of the moral values that ought to shape our basic societal institution, but it is not the only value of justice. This disagreement has shaped debates over technical, regulatory, and governance interventions to reduce algorithmic bias.

In Section 2, I will lay out the problem of algorithmic bias, and different places where bias can enter in the development of an AI system. In Section 3, I argue that the problem of algorithmic bias is sometimes conceived as a problem of justice, and sometimes as a problem of fairness. I illustrate this claim with a discussion of fairness metrics in Section 4. In Section 5, I argue for the

---

[10] See Fischer et. al. (1996) on how policies have widened the wealth gap in America, or Mills (1999) on white supremacy and the so-called racial contract, where social institutions are set up to promote White equality and interests and subordinate people of color.

[11] Shoam et. al. (2018) found that 80% of AI Professors were men.

[12] For example, Brooke (2021) finds that gender bias determines knowledge sharing and recognition on Stack Overflow.

priority of justice over fairness. In Section 6, I conclude by examining five governance strategies for algorithmic fairness and justice.

2. Algorithmic bias

AI Now's 2019 Report lists inbuilt algorithmic bias as one of the emerging and urgent concerns raised by the deployment of AI systems into sensitive social domains.[13] AI systems raise such serious concerns about discrimination because, as we saw in the Introduction, AI has serious impacts on people's lives, in domains such as healthcare, criminal justice, finance, and education. It is important to consider potential discrimination throughout the process of building AI.[14]

Why is algorithmic bias such a pressing and pervasive social and political problem? One important explanation is the historical and current global severity and prevalence of identity-based oppression. Racism, sexism, classism, ableism, and other identity-based oppressions can be thought of as technologies that govern the distribution of advantage, and the imposition of harms and disadvantage, through institutional and social practices.[15] And, they are particularly powerful technologies, shaping where we live, where we work, with whom we associate, and so on.[16] Since AI systems are created by human beings and embedded in our institutions and everyday social practices, we should also expect that identity-based oppression also shapes these technologies.

This expectation has been borne out, as the examples in the Introduction show. But, we also have special reasons to be concerned about AI systems, beyond the pervasiveness of discrimination. AI systems, especially those developed through machine learning, are an especially powerful mirror

---

[13] Crawford et. al. 2019.

[14] Philosophers of science have argued that the potential impacts of scientific choices mean that scientists ought to consider social values throughout the scientific process (Douglas 2007). This literature focuses on the costs of false positives and false negatives, in line with the algorithmic fairness literature (Section 4).

[15] Benjamin 2019; Gabriel forthcoming.

[16] This insight is a central and longstanding one in feminist philosophy and the philosophy of race. As James Baldwin (1998: 723, from Benjamin 2019: 5) observed: "The great force of history comes from the fact that we carry it within us, are unconsciously controlled by it in many ways, and history is literally *present* in all that we do."

of the past, as has been the focus of much recent scholarship.[17] AI systems, especially those generated through machine learning, are powerful decision-making tools for solving classification problems because they learn predictively useful patterns relevant to variable *X* from past data. But, those data are generated by human activity within oppressive social and political structures and attendant ideologies, which make particular social identities highly relevant to a variety of outcomes.[18] And, in cases where someone's social identity is relevant to the predictive task, it is impossible to separate out the effects of discrimination and have a model suited for the predictive task.[19]

And, AI technologies are not only a mirror of the social world. They are also a powerful *shaper* of our institutions and social practices.[20] Hiring processes, for example, incorporate algorithmic decision-making to find candidates, sort applications, and conduct initial interviews. Algorithms used in hiring have been shown to reproduce societal biases. But, increased data sharing[21] and feedback loops between algorithmically-mediated decisions, the data, and algorithms can also increase unjust disadvantage for members of oppressed or marginalized groups. For example, the data that determines one's credit score often appears in a credit report, which about half of employers in the US use in the hiring process.[22] Biased data can influence both employment outcomes and outcomes that require access to credit, such as education and housing. And,

---

[17] E.g., Eubanks 2018; Noble 2018; Mayson 2019.
[18] As Frye (1983: 19) says: "It is not accurate to say that what is going on in cases of sexism is that distinctions are made on the basis of sex when sex is irrelevant; what is wrong in cases of sexism is, in the first place, that sex is relevant; and then that the making of distinctions on the basis of sex reinforces the patterns which make it relevant."
[19] One strategy is to remove the social identity variables as training data for the learning process or inputs to the prediction model. But, if the social identity is a significant predictor for the target, then many other predictively useful inputs will be correlated with social identity as well (the so-called "proxy problem") (Corbett-Davies and Goel 2018).
[20] Gabriel forthcoming; Noble 2018.
[21] Gandy 2009, Hellman 2021, Herzog 2021.
[22] Kiviat 2019. It is often used as a proxy for responsibility (O'Neill 2016: 147).

algorithmic decision-making can create new social identities that are the subject of discrimination, such as the algorithmically left out[23] or the commercially unprofitable.[24]

The moral concern about algorithmic bias is the concern that algorithmic decision-making can be *wrongfully discriminatory*.[25] Wrongful discrimination is unjustifiably differential treatment. To further flesh out this idea, I will borrow two standards for wrongful discrimination from US federal law: *disparate treatment* and *disparate impact*. Disparate treatment involves treating someone differently because of a social identity characteristic, in a manner that is inconsistent with their equal moral worth. The wrongness of the treatment is often understood to be a matter of the decision-maker's intentions: treatment is wrongfully discriminatory when a decision-maker is motivated by negative attitudes about members of that group.[26] However, here I follow Hellman (2008) in thinking that no intention is required to treat someone in a discriminatory way. Treatment can be disparate when someone is *demeaned*, or treated as having lesser moral worth, in virtue of a socially salient characteristic. Disparate impact, by contrast, does not locate discrimination in an interpersonal interaction, be it one where one party intends to discriminate, or where one party demeans another. Instead, disparate impact occurs when there are unjustified inequalities between marginalized or oppressed groups and privileged groups.[27] While the law is an imperfect guide to morality, these two standards capture important moral concerns about discrimination: how we treat people, and how our institutions tend to allocate benefits and burdens among different groups. AI can be discriminatory in both of these senses, as we will see.

---

[23] O'Neill 2016.

[24] Fourcade and Healy (2013) discuss the ways in which companies can use AI and big data to predict who will be a profitable customers, creating new, economically and socially salient categories of the profitable and unprofitable.

[25] Hellman 2008: 2-4.

[26] Alexander 1992. For criticism of such accounts, see Hellman 2008; Lippert-Rasmussen 2013: Chapter 4.

[27] Selmi 2013: Chapter 12. A theory of disparate impact must specify when inequalities qualify as discrimination, lest, for example, it condemns affirmative action, or harmless inequalities that happen to come about through group member's choices. More analysis is needed to state when an inequality is unjustified. But, since disparate impact is not the focus of this chapter, I will set such issues aside.

Let's now turn to our examination of the process of developing and deploying an AI system for decision-making. I will focus on major points that bias can enter into AI systems. The first step to decide on the task that the system is to perform. Say that a company wants to invest in an algorithm to streamline the hiring process. They want this algorithm to identify the best job candidates from a pool. But, "find good employees" is not a task that a machine can accomplish. The task must be the right kind of task, namely, a problem that can be solved by predicting the value of some target variable. And, the target variable must be one that can be measured, and the model must be built, tested, and deployed using available data. "Predict who would make the most sales in their first year on the job," is, by contrast, a task that an algorithm can accomplish.[28]

But, certain ways of formulating tasks can be discriminatory. Disparate impact is a particular worry regarding task choice, especially because it is not always obvious when a certain problem specification will lead to unjustifiable differential impacts on protected groups.[29] Disparate impact can come about because the task specification leads to differently accurate predictions for members of different groups. For example, Obermeyer et. al. (2019) show that a common healthcare algorithm is less accurate in predicting Black patients' comorbidity score, i.e., Black patients have a higher comorbidity score than White patients at the same level of assigned algorithmic score. They hypothesize that the bias arises because the algorithm does not predict comorbidity score; instead, predicts total medical expenditure in a year. But, total medical expenditure is not a good operationalization of health need for Black patients, because Black patients generate fewer healthcare costs at a given level of health.

The choice of task can also lead AI systems to be discriminatory, in the disparate treatment sense. Consider research by Latayna Sweeney (2013) that found significant discrimination in ad

[28] Passi and Barocas 2019.
[29] Barocas and Selbst 2016.

delivery by Google AdSense. Sweeney found that a greater percentage of ads containing "arrest" in the title were delivered for searches of names that are racially associated with being Black than for searches of names that are racially associated with being White. One plausible mechanism that produces this difference is the target metric – optimizing for clicks. By optimizing for clicks, Google and other ad systems create the conditions for discriminatory user behavior to determine ad relevance. By clicking ads with "arrest" in the title more often for names that are associated with being Black than for names that are associated with being White, users make ads with "arrest" in the title more relevant to Google searches for names associated with being Black. Thus, the task, coupled with data generated by the users, create associations between Blackness and criminality. AI systems may not have intentions to discriminate or discriminatory attitudes. But, an ad system that associates Black names with criminality is surely demeaning, qualifying as disparate treatment.

The next step of the AI pipeline is to gather data and extract features. Discrimination can arise here as well. One will often hear the adage "garbage in, garbage out" to describe the idea that one's model is only so good as the data that it is built on. One way that data can lead to discrimination is by being inaccurate. Measurement error includes over- or under-sampling from a population or missing features that better predict outcomes for some group in the feature extraction process. Inaccuracy can lead to disparate impact because the algorithm will make more mistakes about one group. The remedy is usually to gather better data.

However, in some cases, the problem is not measurement error. In those cases, the bias is due to discrimination in the institutions and other social contexts that generate the data.[30] For example, Richardson et. al. (2019) argue that police corruption and other unlawful practices have seriously degraded the quality of data in many places in the United States, making accurate feature

[30] Mayson 2018.

choice to train predictive policing models incredibly difficult.[31] In such cases, collecting more and more accurate data will not address the problem of discrimination.

The next step in the process of building an AI system is to build a model using the data. Here is a third place where bias can enter, in the form of model bias. Models, especially those learned through machine learning techniques, can learn biased statistical regularities from the human-generated data they are trained on. A major source of model bias is biased data, but other modeling decisions can introduce bias. Machine learning, for example, aims to find the optimally performing model for some task, relative to a particular standard.

The choice of which metric to optimize for can have discriminatory impacts. For example, optimizing for predictive accuracy can worsen disparate impact discrimination. Predictive success is often achieved by using *proxy attributes*, or features that correlate with some socially salient characteristic, to the target of interest.[32] Even if proxy attributes have only a small correlation with a sensitive attribute, access to many such features will produce a classifier that implicitly uses sensitive attributes for prediction.[33] Prediction based on proxy attributes can thus induce a tradeoff between accuracy and the demand not to discriminate. Because social identities pervasively structure our social lives, reducing a reliance on proxy attributes can reduce the model's accuracy. But, if those predictions have negative, differential impact on protected groups, then we have discrimination-based reasons to reduce the reliance on those features in decision-making.[34]

The final step where bias may enter is when an AI system is used for decision-making. To use AI for decision-making, decision-makers must convert predictions to decisions. *Decision thresholds* are often used to convert a continuous prediction into a decision. A decision threshold is a function

---

[31] See also Knox, Lowe, and Mummolo (2020), as well as Gabler et. al. (2020) for pushback.
[32] Dwork et. al. 2012; Johnson 2021. Discrimination that arises from decision-making based on proxy attributes has long been a concern to philosophy of discrimination and the law (see, e.g., Alexander 1992).
[33] Barocas, Hardt, Narayanan 2019.
[34] Johnson 2021.

from a prediction to a decision. Decision thresholds are useful because they take a more complex prediction, often in the form of a probability distribution, and assign individuals into a "yes" or a "no" set, based on whether they are over or under the threshold.

Features of the context in which the algorithm is deployed may introduce bias. The decision-maker's goal may be to identify members of disadvantaged or marginalized groups in order to further disadvantage them.[35] The performance of the model may also decrease if the model is used in a context where the data is very different from the data that the model was trained on.[36] If the model is systematically less accurate for members of disadvantaged or marginalized groups, then the model can create disparate impact, as discussed in the context of healthcare above. Furthermore, systematic deployment of a model across many different contexts can create feedback loops, further entrenching disadvantage. For example, if employers tend to use a small number of hiring algorithms that disfavor older workers, future algorithms will not have much data about successful older workers, leading to even fewer that are hired.[37] This could also lead to disparate treatment, as being older becomes associated with being unemployable.

Accurately predicting an individual's risk of being caught committing a crime, and then making a bail decision on that basis, can further disparate impact discrimination. And, the decision-maker may have multiple goals, some of which are not represented in the target metric.[38] But, if a decision-maker is concerned about disparate impact discrimination, algorithmic predictions can still be very useful. In response to the accuracy-anti-discrimination tradeoff discussed earlier, some have suggested that some algorithmic predictions should be considered as diagnostic for the presence of injustice.[39] Prediction still guides action, in such cases, but it is more likely to point the decision-

---

[35] Eubanks 2018 chronicles how AI can be used by government officials to further disadvantage the poor.
[36] This is the so-called problem of external validity: under what conditions does a model predicts or explains well in new contexts, and when should decision-makers be confident that the model will do so? (Rodrick 2009)
[37] Herzog 2021.
[38] Kleinberg et. al. 2018 call this "omitted payoff bias."
[39] Mayson 2018.

maker towards remedying background injustice that produces disparate impact discrimination, rather than bearing on the decision at hand. This allows decision-makers to use accurate predictions to achieve anti-discrimination goals, and thereby to mitigate the tradeoff between the two.

AI thus raises serious concerns about wrongful discrimination, due to algorithmic bias that can enter at different points in the design and deployment of AI systems. Since AI is a powerful mirror and shaper of the social world, we urgently need to better understand the problem of algorithmic bias, and what to do about it. The rest of this Article takes on that task.

3. A problem of fairness, or a problem of justice?

Researchers, civil society activists, politicians, and businesses have generated a wide range of policy proposals, technical tools, and governance strategies to make algorithms fairer. For any of these, however, there is trenchant disagreement. Consider ProPublica's 2016 reporting on COMPAS, an algorithm that informed judicial decisions in the United States criminal justice system. They alleged that the risk assessment algorithm, which predicted a defendant's risk of future crime, was biased against African-Americans. The developers of COMPAS, however, contended that the algorithm was indeed fair. What can explain these different judgments?

In this section, I will argue that research, governance, and public debate about algorithmic bias contain a key unarticulated disagreement over whether it is more pressing to govern AI to be *just* or *fair*.[40] "Justice" is a term that encapsulates a number of moral commitments about how society's key institutions ought to treat people. Fairness is one type of standard of justice.

There have been increasing demands for *justice* in AI. "Justice" refers to the moral standards that ought to structure major societal institutions, such as legal, political, and economic institutions.[41]

---

[40] This diagnosis is a more general version of Barocas and Selbst's (2016) argument that attempts to use US anti-discrimination law to tackle algorithmic bias suffer from an ambiguity in the law, as to whether anti-discrimination is an anti-classification or an anti-subordination project.

[41] Rawls 1999.

These standards determine people's obligations, entitlements, opportunities, and burdens. In doing so, they unavoidably shape people's life trajectories, as well as their ideas about justice. One example is a country's laws about income tax. Tax law determines how much everyone has to pay in income tax, which in turn determines the level of inequality in a country, how much individuals have to work to sustain themselves, incentives to work, and so on. And, the moral standards behind income tax law also determine people's ideas about the justness of different income tax regimes. If, for example, the tax regime is based on the idea that people deserve their pre-tax income, then citizens may not be willing to support higher income taxes.

Taking the perspective of justice allows us to better interrogate issues around AI and inequalities of material resources, opportunity, and power, individual autonomy and political self-determination, representation, and respect and recognition. In other words, it allows us to take a structural perspective on institutions and the moral standards they ought to live up to, but usually fail to.[42] We can ask questions of *distributive justice*: how basic rights, liberties, opportunities, and material resources ought to be distributed. Or, we can ask questions about *productive justice*: how work ought to be organized, such that the burdens and benefits of work are justly apportioned. We can also ask questions of *corrective* justice – how should institutions right past wrongs, especially in cases of historic injustice – or *racial and gender* justice – how structures perpetuate the exercise of power by some groups over others. Thinking about justice opens up a wide swathe of important structural questions (see the chapter on AI and structural injustice in this volume).

Standards of justice are grounded in a plurality of different types of moral reasons: reasons of equality, autonomy, gratitude, or dessert. One important category of such reasons are reasons of *fairness*, which embody a valuable type of equality. Fairness is about *respecting equal claims,* as well as

[42] Le Bui and Noble (2020) call for a moral framework of justice to be integrated into research and governance on AI. They, following Mills 2017, push for an explicitly non-liberal moral framework of justice.

respecting claims in proportion to their strength (Broome 1991). Reasons of fairness are grounded in the moral equality of people. All else equal, people have an equal claim to important resources to make their life go well, or to equality of respect and standing in their political community. Of course, all else may not be equal, with regard to material resources or other goods. Someone may have a claim to more of a good, because of a prior agreement that establishes a greater claim, or for the reason that they deserve or need more of the good. In those cases, fairness requires respecting people's claims in proportion to the strength of those claims. Thus, fairness is a matter of *proportional equality,* or giving people their due relative to what is owed to them.

Fair decision procedures are those that treat like cases alike, in terms of people's claims.[43] If two people have a claim to the same level of worker's compensation based on injury, then a fair decision procedure gives them both that level of worker compensation. A fair outcome is a distribution of resources, material or not, that respects people's claims. And, alongside examining distributions, we can also interrogate whether decisions using AI satisfies people's equal claim to respect, or to not be discriminated against in the disparate treatment sense.

Fairness is a central moral concept in our thinking about justice and AI. However, it is not the only moral concern one might have. Some of the disagreement over algorithmic bias comes from a disagreement over whether to pursue fairness, or whether to pursue some of the other values of justice. Consider debates over the use of AI for predictive policing in contexts there are racial disparities in the past crime data. Let's say that those disparities come about due to different base rates in the crimes committed by members of each group.[44] In such contexts, calls for fair models often motivate calls for *accurate* predictive policing. Accurate predictive policing, one might claim,

---

[43] Hart 2012. Zimmerman and Lee-Stronach (forthcoming) call this the "Like Cases Maxim."

[44] As previously discussed, disparities in law enforcement data often come about through racially biased policing (Richardson 2019; Mayson 2018). This Article does not assume that any actual disparities in past crime data are due to a difference in base rates.

respects people's equal claims to protection from law enforcement. This can be done through predictive models that utilize accurate statistical generalizations about the actual populations the model is trained on (call this a "bare statistic") to make predictions. But, those bare statistics may only be true of the arbitrary populations the model was trained on; they may not be *projectable,* or true in new populations. One important reason that bare statistics may not be projectable is that statistical generalizations are true against a background of particular social structures and practices, i.e., certain economic and social determinants.[45] A concern for racial justice, by contrast, more often recognizes that these statistics are not projectable. Calls for racial justice in policing are often calls to ban algorithmically-driven predictive policing, or to use AI for other interventions to reduce the difference in base rates.[46] Such calls are concerned that a focus on accurate prediction of crime may, perversely, increase the unjust subordination of one group, by keeping in place the social conditions that underwrite the generalization and entrenching the pattern by using it as an evidential basis for policy. Calls for racial justice are calls to increase the justice of the background structures that produce crime, rather than targeting an equal application of current legal standards.

The disagreement about fairness and other values of justice inflects a number of disagreements over algorithmic bias. In the next section, I will examine one such disagreement in more detail: the disagreement over fairness metrics.

4. Illustration: fairness metrics

There has been an explosion of research in computer science on algorithmic fairness, e.g., how to develop less discriminatory algorithms from a technical standpoint. *Statistical* approaches are post-processing techniques that take a learned model and aim to measure and mitigate discrimination based on observable inequalities between groups. *Causal* approaches are often taken to correct

[45] Munton 2019.
[46] Mayson 2018.

shortcomings in statistical approaches.[47] I will not attempt an exhaustive survey of this fast-paced, interdisciplinary area of research. Instead, I will use the example of statistical fairness criteria to illustrate the divide between proponents of fairness and proponents of other values of justice.[48]

Different statistical criteria for fairness have been proposed. These fairness criteria each formalize a notion of fair prediction as prediction where the output prediction is non-discriminatory, or does not depend on individuals' social identities. Thus, most fairness criteria are expressible in terms of a relationship of (conditional) independence between the predictor *R*, the sensitive attribute *A,* and the target *Y* (i.e., the algorithm's task).[49]

Statistical fairness criteria exploded onto public consciousness because of ProPublica's (2016) allegation that COMPAS, an algorithm that predicts a defendant's risk of future crime, was biased against African-Americans. ProPublica made this allegation on the basis of a comparison of the predicted risk score against later data about who committed crimes. Their analysis found that the algorithm wrongly labeled black defendants as high risk at about twice the rate as white defendants; conversely, it wrongly labeled white defendants as low risk at almost twice the rate as black defendants. They concluded that the algorithm was discriminatory. That conclusion assumes that *false-positive or false-negative equality* is the best criterion for detecting algorithmic bias. These criteria measure the proportion of false positives or false negatives within each group. Such false-positive or false-negative equality metrics are motivated by the idea that people who are the same with respect to outcomes ought to be treated the same in terms of having true (or false) decisions made about

[47] Kusner and Loftus 2020.
[48] Statistical fairness criteria are also worth focusing on because they are, arguably, still the standard approach in computer science and in industry to quantify the amount of bias in an algorithm (Danks and Fazelpour 2021a).
[49] Barocas, Selbst, and Narayanan 2019: Chapter 2.

them at similar rates. More formally, conditional on the target variable *Y*, the score *R* should be independent of protected attribute *A*.[50]

The developers of COMPAS, however, contended that false-positive or false-negative equality is not the best measure of whether the algorithm is accurate. This point has been compellingly argued by a number of philosophers, computer scientists, and social scientists. Here is one example argument as to why a difference in false-positive or false-negative rates is not always an indicator of unfairness. Recall that AI systems often employ a threshold to convert a prediction into a decision. There would be fewer false positives for individuals that are predicted to be far from the threshold (more colloquially, the clear yeses or a clear noes), and more false positives and negatives closer to the threshold. If members of one group cluster closer to the threshold, and members of another group are on the high and lower ends of the thresholds, then there will be more false positives and negatives for the former group. This difference in false-positives and false-negative rates between groups is not in and of itself unfair; to put it another way, the difference is not unfair in all circumstances. So, if the best criterion is one that an unbiased system must always fulfill, then false-positive and false-negative equality is not such a criterion.

Many computer scientists, social scientists, and philosophers favor *calibration* as the best metric for fairness. A well-calibrated system is one in which, conditional on a particular score *R*, an individual's group membership *A* and the outcome variable *Y* are independent. The intuitive idea behind calibration measures is that the standards that the decision system uses work equally well to identify who meets that standard, independent of their group membership. It is often taken to formalize a notion of "fair testing" or "fair application of a standard": a test or application of a standard is fair to the extent that a decision is an equally good predictor of the actual outcome of

---

[50] Where R is a binary classifier and *a* and *b* are two groups, this can be expressed in terms of the conditional probability that P{R = 1 | Y = 1, A = a} = P{R = 1 | Y = 1, A = b} for false positive equality, and P{R = 1 | Y = 0, A = a} = P{R = 1 | Y = 0, A = b} for false negative equality.

interest. For example, a calibrated university entrance exam is one whose classification of students to admit and not admit on the basis of their test scores is an equally good predictor of their success at university across different demographic groups.

Calibration thus seems to better embody a requirement of fairness, understood as treating like cases alike. Consider the example of a miscalibrated entrance exam, where a score of 90, say, is associated with an 85% chance of success at university for members of group 1, but only a 65% chance of success for members of group 2. If a decision-maker admitted all students with a score of 90% or more, members of group 1 have a 15% chance of a false positive, whereas members of group 2 have a 35% chance of getting a false positive. Since members of group 2 have a systematically higher chance of being subjected to false positives, the system is unfair.

Calibration is often satisfied without any explicit intervention, in cases where a sensitive attribute can be predicted from other attributes.[51] This fact is not surprising: if there is enough data about individuals from different groups that can be used to build an accurate predictor for a target that is correlated with group membership, then the scores that the model assigns to individuals are equally informative for the outcome. Interventions to ensure that systems are calibrated aim to produce accurate models, holding fixed background social structures. However, as discussed above, accurate modeling within unjust institutions can replicate and further injustice. Thus, while calibration may be a good metric to promote fairness, relying on calibration alone to mitigate algorithmic bias can set back other values of justice.

[51] And, calibration does not ensure that ither design choices – the target goal, or the decision function to translate predictions into scores – are just. Say that a bank wishes to discriminate against black loan applicants, and knows that black applicants live in zip codes with relatively high default rates, and that white and black applicants have similar default rates within each zip code. The bank could develop a calibrated algorithmic system that predicted default rates by zip code alone. But, one could build a calibrated alternative system that uses more information and divides individuals into finer risk buckets (Corbett-Davies and Goel 2018: 16).

If one is concerned about these other values of justice, then unequal false positive or negative rates may be a better metric. That is because differential rates of false positives may be evidence of injustice, even if they are not unjust or unfair in themselves. For example, if more members of a privileged group tend to cluster away from a threshold, it could be because of past injustice in that group's access to resources. Or, more false positives may have a disproportionate impact on the welfare of members of a marginalized group. However, it is important to be clear about the claim here. A difference in false positive or false negative rates is evidence of injustice, not unjust itself. Thus, a decision maker does not have reason to reduce false positive or false negative rates, unless they have evidence that doing so is a good means to increase welfare or reduce injustice. For example, one could reduce the disparity between white and black individuals in COMPAS by prosecuting more black individuals who are low risk. But, to do so would be unjust.

5. Justice over fairness

I've argued that different technical interventions can promote different values of justice, such as fairness and racial justice. Furthermore, I've also argued that there can be tradeoffs between promoting fairness and other values of justice. These leads us to an important moral question: which of those values should be prioritized?

There is not always a trade-off between fairness and other values of justice, however. That is because fairness has no value in situations of serious and pervasive injustice. Fairness depends on other standards of justice, because these other standards are necessary to determine which cases are alike, from a moral perspective. On its own, the value of equality does not speak to the question of which cases are alike; to put it another way, standards of fairness are "empty" unless other norms of justice determine which properties of agents count as relevant to the decision at hand.[52] Ought individuals be provided with unemployment benefits because they cannot work but are owed the

[52] Hart 2012; Westen 1982.

means to live a decent life, because they have made valuable contributions to society but can't find a job at the moment, or just in virtue of being a member of the community? These different ideas of what justice requires back different claims; it is then a further question of fairness whether institutions respect those claims, *given what other standards of justice have fixed them to be*.

This point leads us to my claim of this section: If the contexts in which AI is designed and deployed are seriously unjust, then one must prioritize other values of justice over fairness.

The first argument for this claim is that without just institutions, fairness is of little moral worth. The value of fairness depends on the existence of just institutions in the background. Institutions determine at least some of people's claims, e.g., which individuals count as relevantly similar, as discussed above. However, mere equality of treatment, in the face of any possible set of claims, is not morally valuable. If the standards that determine individuals' claims are unjust, then applying rules in an even-handed manner does not have moral value. Say that I make a rule that the first ten children to arrive at a birthday party get a slice of cake, and the rest don't get any cake. I faithfully implement this rule at the birthday party, giving the first ten children a slice of cake, and the rest no cake. For the children who didn't get any cake, is it morally valuable that I fairly applied the rule? Arguably, not. Implementing a rule consistently or impartially is not in itself valuable. We value consistent rule application when the rules are just in their distribution of benefits and harms. In other words, fairness is an important value of justice because people have *legitimate* claims, that ought to be respected. Fairness then ensures that those rules are used consistently and impartially in decision-making, respecting the claims of each.

One reason, then, to prioritize other values justice is that without justice, there is no fairness. The second reason to prioritize other values of justice is one that I have already discussed in some detail throughout. The reason is that fair decision-making can compound injustice. Say that there is enough good housing for everyone in a city at affordable rates to rent or own, but that higher-quality

housing is more expensive, and lower-quality housing less expensive. And, say that the political and social institutions of this society have the norm that everyone deserves quality housing, and that housing of different quality should be distributed according to willingness to pay. Landlords decide between everyone who can afford the housing by lottery. However, this society is also marked by historic and current access to credit along racial and gender lines. The landlord's decision lottery is fair, as it respects everyone's equal claims to a house. But, it compounds injustice, in that ability to secure a mortgage or loan for a rent shortfall on fair terms is determined in party by race and gender. If a fair decision procedure relies on inputs that are the outputs of historic or current structural injustice, then the decision procedure will compound injustice. And, since not compounding injustice is more important than fairness, we should focus on whether AI compounds injustice, if there is a tradeoff between that value and fairness.

6. Policy recommendations for just AI

I want to close by discussing policy strategies to govern AI for justice, including fairness. These policy strategies come out of the preceding discussion of algorithmic bias, justice, and fairness. They are not intended as a complete survey of the governance options.

6.1 Take a values-first approach to bias interventions

It can be tempting to introduce AI into existing decision processes, without reflection on the new moral problems that AI raises, or how it might exacerbate existing problems. Furthermore, companies and others who benefit from easily available data have tried to inculcate the attitude that the use of any data is fair, as long as it is predictively useful.[53] This can further blind us to the ways in which algorithmic choices are value-laden. An especially gaping gap in algorithmic governance is a moral examination of the predictive task, as interventions to measure and reduce algorithmic bias

[53] Zuboff 2019.

tend to focus on modeling choices and model-based interventions, as well as, to a lesser extent, the quality of the data.

The first governance suggestion is to take a values-first approach to algorithmic bias. By "values-first," I mean that clear statements of values and their tradeoffs ought to determine the choice of a policy goal for AI regulation or the measurement and mitigation of bias in a particular algorithmic system. Sometimes, in combatting bias, it seems as if the goal is clear – say, to increase diversity. But, we can better understand what the goal is, and what interventions can achieve it, if we recognize the values behind increasing diversity. In hiring, for example, a decision-maker may be concerned that equality of opportunity is violated because qualified candidates from one race are not noticed as often by recruiters (perhaps the company is racially homogenous and employees often recommend their friends for jobs at the company). If the value is equality of opportunity, then steps to increase diversity in sourcing – such as auditing job ads for potential biased language – may be called for. In university admissions, by contrast, a decision-maker may be concerned that students from disadvantaged backgrounds are subject to further disadvantage because they did not have the same opportunities to build their resume that other students had. Here, an affirmative action system may be warranted by the value of reducing educational injustice.

A values-first approach is especially important for algorithmic bias, because different values of justice support different standards to measure bias in the system, and support different interventions. Regulatory and governance efforts must be clear about what the important moral values are in that domain, and how the algorithm's task, data, modeling choices, and design connect with those values. Take risk assessment tools in criminal justice. What kind of risk ought they to be predicting, in contexts with historical injustice and current structural injustice? Say that legislators aim to avoid racial injustice, and keep people safe from violent crime. These two values support decision-making tools to predict the risk of violent crime arrest. Of course, there may not be enough

data to predict violent crime, since it is rarer than other kinds of crime; in that case, decision-makers should not substitute another target that is easier to predict. AI development and deployment should be driven by values, not by data availability.[54]

6.2 De-couple decision processes

The second governance suggestion is to de-couple decision processes. This point holds across decision process, and for a single type of decision, such as loans, or hiring. As argued above, AI can compound injustice across decision processes. Furthermore, because AI systems can operate at a much greater scale than well-coordinated human decision makers, and can create salient social identities that may be the target of injustice, there are strong reasons of justice to avoid using the same AI system for a large class of decisions.[55]

Governance efforts can target AI-generated decision aids that are commonly used across different types of decisions. Credit reports, for example, are used to make decisions not only about loans, but also about jobs, housing, and insurance, even though there is not much evidence that credit reports are a good predictor of productivity.[56] But, because personal data about individuals is cheap and easily available, this move will be of limited utility on its own. Governance efforts will also need to reduce the sale and use of personal data.[57]

A concern about this strategy is that it reduces the accuracy of decisions, by limiting the input data for building AI systems. In many cases, such as the use of credit reports in hiring, data are used to make efficient decisions that are better than choosing at random, not highly accurate decisions. Furthermore, governance strategies can promote accuracy in the long term. One such governance strategy promotes the use of lotteries early in a decision process, say, a lottery among

[54] Mayson 2018: 2269.
[55] Hellman and Creel forthcoming.
[56] Weaver 2015.
[57] See Veliz's contribution to this volume.

interested law students to hire summer interns at a law firm, to avoid compounding injustice. After a certain amount of time, employers will have enough data about the actual job performance of those interns to make a data-driven decision.[58] Another kind of governance strategy introduces multiple decision processes, based on different criteria. A diversity of criteria can avoid over-indexing on a single set of criteria, which are likely to arbitrarily privilege those who have been the recipients of past advantage.[59]

6.3 Model structural injustice

A third governance suggestion is to include information about structural injustice in the decision process, especially for decision-makers that ought to advance justice.[60] How computer and social scientists model the social world depend on their assumptions about the prevalence and severity of racial injustice.[61] If racial injustice is prevalent, then, per the discussion above, scientists ought to be cautious about including proxy attributes for race, limiting what they can predict. Scientists ought to make these empirical assumptions explicit, so that decision-makers are in a better position to judge whether the algorithm's assumptions hold in her context.

Furthermore, decision-makers that want to promote gender and racial justice need to understand where and how they can intervene, and what the downstream effects of using an AI system will be. Predictive algorithms ten to take institutional background structures as given, which can sideline possibilities for intervening on those institutions to promote more just outcomes.[62] And, the moral permissibility of using a particular decision system depends on its long-term impacts in a

[58] Hu and Chen 2017.
[59] Fishkin 2013.
[60] Zimmerman and Lee-Stronach 2021; Mayson 2018. See also Gabriel forthcoming, Herzog 2021, and Ferretti 2021 for arguments that the duty to advance justice applies to private companies developing and deploying technology.
[61] Hu 2021.
[62] Zimmerman and Lee-Stronach 2021.

particular context. Without modeling the dynamics of that social context, the decision-maker does not have a proper evidence base to judge whether to deploy the AI system.[63]

6.4 Better data

The fourth suggestion is more regulation to improve data quality. Data is often taken as given and as fact; furthermore, there is a tendency to use data that is easily available. But, as we saw, biased data is one of the key causes of model bias. Increasing data quality will produce more accurate and robust models.

There are a number of scientific, social, and political interventions required to produce better data. Some regulations that would improve data quality are more scientific or technical.[64] Other regulations would instead correct incentives in the data economy that produce poor or limited data. For example, as Véliz discusses in this volume, data brokers have incentives to collect massive quantities of data cheaply, and a disincentive to ensure that any piece of information is accurate. Banning targeting advertising would address data quality issues. In its place, governments should consider funding public organizations to create accurate and inclusive data sets. They should also mandate data sharing by private companies with researchers and auditors.

6.5 Replace decision thresholds with more (weighted) lotteries

The previous governance strategies focused on other values of justice. The final suggestion focuses on fairness. If one is concerned about designing for fairness, then one should not use decision thresholds. Instead, fairness requires more decision-making by lottery. Alongside calibration, randomness offers another technical lever to increase fairness.

It is important to be clear upfront about what these arguments purport to show. They show that decision-makers have reasons of fairness to use lotteries instead of decision thresholds. They do

[63] Danks and Fazelpour 2021b.
[64] See Gebru et. al. (2018) on data sheets for data sets.

not show that decision makers have decisive reason to introduce more randomness into algorithmic decision-making. There may be other, weightier reasons to use decision thresholds.

If decision-makers know what individual's claims are, as well as their strength, then they should use weighted lotteries, rather than decision thresholds. The argument for this claim relies on the definition of fairness as the satisfaction of claims in proportion to their relative strength. Say that Carl has lent me $100, and Dani has lent me $200, but I only have $60 to pay them back. In such a case, fairness requires that I ought to give Dani $40 and Carl $20, since her claim is twice as strong as Carl's. In the case of divisible goods like money, the good should be allocated proportionally, where the proportions are determined by the relative strength of people's claims.

And, when goods are indivisible, they should also be allocated proportional to the strength of someone's claim, by weighed lotter. Let's now change the example to a kidney transfer. Dani is four times as sick as Carl, and thus has a claim to the kidney that is four times as strong as Carl's. The fairest way to allocate the kidney is by a weighted lottery, where the weights reflect the proportional strengths of various claims. A kidney lottery, for example, should be set up so that Dani is four times more likely to win the lottery as Carl.

This argument may strike you as objectionable – shouldn't the person with the strongest claim get the kidney? At first glance, a policy that always distributes a good to someone with the strongest claim seems to be the most fair policy.[65] Such a challenge seems more challenging for a single decision, say, whether to give the kidney to Carl or Dani.

However, such a policy ignores the weaker claims of others. The moral strength of this point is more easily appreciated when one zooms out to reason about the fairness of an allocation procedure for a population over time. Imagine that this kidney allocation lottery was used across a country to allocate donated kidneys to prospective patients. Fairness requires that the decision-

[65] Hooker 2013.

maker can give the losers of an allocation procedure – those who don't win the kidney lottery, say – a reason why their claims were respected. And, in the case of the kidney allocation, the reason must be that they had a real chance at getting a kidney.[66] It does not seem reasonable to ask them to completely sacrifice their claims to the someone with a stronger claim, which would be required by an allocation procedure that always allocated indivisible goods to those with the strongest claims. A weighted lottery, by contrast, respects everyone's claims, as individuals have a chance of getting the good that is proportional to the strength of their claim. Thus, fairness requires weighted lotteries for indivisible goods – and, more generally, that claims be satisfied in proportion to their strength.

This point raises a serious challenge to the allocative fairness of most algorithms. For most algorithms, those below the decision threshold have some claim to the desirable outcome (or to avoid an undesirable one). However, anyone below the decision threshold has no *ex ante* real chance at the outcome. So, any algorithm that uses a decision threshold to separate individuals with greater and lesser claims is unfair.

The argument for weighted lotteries assumes that individuals have well-established claims of differing strengths, and that decision-makers can gather enough information about those claims to design a weighted lottery. In situations of injustice, it may be that individuals have an equal claim to a good, rather than claims of differing strength. For example, in a society where people have unjust differential access to educational opportunities and material resources, it may be that better qualified individuals do not have a greater claim to the job. So, initial lottery for all qualified individuals would be fairer, rather than allocating the job to the most qualified individual. Furthermore, individuals may have claims of differing strengths, but decision-makers may not be able to gather enough information to identify which individuals have stronger claims. In such cases, a lottery with equal

[66] Spiekermann 2021.

weights would be fairest, as each individual has the same *ex ante* chance of their claim being disregarded, in light of what the decision-maker knows.

7. Conclusion

"Fairness" is an ambiguous and messy term, one that is central to politics but can be more obfuscating than clarifying. This chapter distinguished between fairness and other reasons of justice, and explained disagreements over how to address algorithmic bias as disagreements over whether to prioritize fairness or those other reasons of justice. One major lesson of the chapter is that more accurate decision-making can contribute to injustice. However, less accurate decision-making is not the solution, as it may not advance justice, and can come at too high a cost to other values. The chapter ended by promoting a number of governance strategies to promote fairness, such as reducing the use of decision thresholds, and to reduce the injustice that can arise from AI systems that exploit predictively powerful correlations that increase inequalities or otherwise entrench disadvantage, such as modeling structural injustice and better data.

References

Alexander, L. 1992. "What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies." *University of Pennsylvania Law Review* 141(1): 149.

Ali, M., P. Sapiezynski, M. Bogen, A. Korolova, A. Mislove, A. and Rieke. 2019. "Discrimination through optimization: How Facebook's ad delivery can lead to skewed outcomes." Retrieved from https://arxiv.org/pdf/1904.02095.pdf.

Angwin, J, J. Larson, S. Mattu, and L. Kirchner. 23 May 2016. "Machine Bias. There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks." *Pro Publica.*

Baldwin, J. 1998. *James Baldwin: Collected Essays*. New York: Library of America.

Barocas, S. and A. Selbst. 2016. "Big Data's Disparate Impact." *104 California Law Review* 671.

Barocas, S., M. Hardt, A. Narayanan. 2019. *Fairness and Machine Learning.* fairmlbook.org. Available at https://fairmlbook.org/.

Benjamin, R. *Race after Technology:Abolitionist Tools for the New Jim Code*. Polity Press.

Buolamwini, J. and T. Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1–15.

Brooke, S.J. 2021. "Trouble in programmer's paradise: gender-biases in sharing and recognising technical knowledge on Stack Overflow." *Information, Communication & Society* 24:14: 2091-2112, DOI: 10.1080/1369118X.2021.1962943

Broome, J. 1991. "Fairness." *Proceedings of the Aristotelian Society* 91: 87-101.

Crawford, K., R. Dobbe, T. Dryer, G. Fried, B. Green, E. Kaziunas, A. Kak, V. Mathur, E. McElroy, A.Sánchez, D. Raji, J. Rankin, R. Richardson, J. Schultz, S. Myers West, and M. Whittaker. 2019. AI Now 2019 Report. New York: AI Now Institute, https://ainowinstitute.org/AI_Now_2019_Report.html.

Corbett-Davis, S. and S. Goel. 2018. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." arXiv:1808.00023v2 [cs.CY].

Datta, A., Tschantz, M. C., & Datta, A. 2015. "Automated Experiments on Ad Privacy Settings." *Proceedings on privacy enhancing technologies* 2015(1).

Danks, D. and S. Fazelpour. 2021a. "Algorithmic Bias: Senses, Sources, Solutions." *Philosophy Compass*. https://doi.org/10.1111/phc3.12760.

Danks, D. and S. Fazelpour. 2021b. "Algorithmic Fairness and the Situated Dynamics of Justice." *Canadian Journal of Philosophy*.

Dotan, R. 2020. "Theory choice, non-epistemic values, and machine learning." *Synthese*: 1–21. https://doi.org/10.1007/s11229-020-02773-2

Douglas, H. E. 2007. "Rejecting the Ideal of Value-Free Science". *Value Free Science? Ideals and Illusions*, ed Kincaid, Dupré, and Wiley. Oxford: Oxford University Press.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, R. Zemel. 2012. "Fairness Through Awareness." arXiv:1104.3913 [cs.CC].

Dworkin, R. 1977. *Taking Rights Seriously*. Cambridge: Harvard University Press.
Eubanks, V. 2018. *Automating Inequality*. MacMillan Publishers.

Ferretti, T. 2021. "An Institutionalist Approach to AI Ethics: Justifying the Priority of Government Regulation over Self-Regulation." *Moral Philosophy and Politics*.

Fisher, A., J. Medaglia, B. Jeronimus. 2018. "Lack of group-to-individual generalizability is a threat to human subjects research." *Proceedings of the National Academy of Sciences* 115 (27): 6106-6115. DOI: 10.1073/pnas.1711978115.

Fourcade, M. and K. Healy. 2013. "Classification Situations: Life-Chances in the Neoliberal Era." *Accounting, Organizations and Society* 38(8):559–72. https://doi.org/ 10.1016/j.aos.2013.11.002.

Gabriel, I. forthcoming. "Towards a Theory of Justice for Artificial Intelligence." *Daedalus*. arXiv:2110.14419 [cs.CY].

Gaebler, J., W. Cai, G. Basse, R. Shroff, S. Goel, J. Hill. 2020. "Deconstructing Claims of Post-Treatment Bias in Observational Studies of Discrimination." arXiv:2006.12460 [stat.ME]

Gandy, O. 2009. *Coming to Terms with Chance: Engaging Rational Discrimination and Cumulative Disadvantage*. Burlington, VT: Ashgate Publishing Company.

Gebru, T., J. Morgenstern, B. Vecchione, J. Wortman Vaughan, H. Wallach, H. Daumeé III, and K.

Crawford. 2018. "Datasheets for Datasets," arXiv:1803.09010.

Guardian Editorial. 11 August 2020. "The Guardian view on A-level algorithms: failing the test of fairness." *The Guardian*. https://www.theguardian.com/commentisfree/2020/aug/11/the-guardian-view-on-a-level-algorithms-failing-the-test-of-fairness.

Hart, H.L.A. 2012. *The Concept of Law*, third edition. Oxford: Oxford University Press.

Hellman, D. 2008. *When is Discrimination Wrong?* Cambridge: Harvard University Press.

Hellman, D. 2020. "Measuring Algorithmic Fairness," 106 Va. L. Rev. 811.

Hellman, D. forthcoming. "Big Data and Compounding Injustice." *Journal of Moral Philosophy*.

Hill, K. December 29, 2020. "Another Arrest, and Jail Time, Due to a Bad Facial Recognition Match." New York Times.

Hooker, B. 2013. "Fairness." *Ethical Theory and Moral Practice* 8(4): 329-352.

Hu, L. and Y. Chen. 2017. "Fairness at Equilibrium in the Labor Market." *FATML*. *arXiv:1707.01590* ***[cs.GT]***

Hu, L. May 6, 2021. "Race, Policing, and the Limits of Social Science." *The Boston Review*. https://bostonreview.net/articles/race-policing-and-the-limits-of-social-science/.

Johnson, G. forthcoming. "Are Algorithms Value-Free? Feminist Theoretical Virtues in Machine Learning." *Journal of Moral Philosophy*.

Johnson. G. 2020. "Algorithmic Bias: On the Implicit Bias of Social Technology." *Synthese* 198: 9941-9961.

Kiviat, B. 2019. "The art of deciding with data: evidence from how employers translate credit reports into hiring decisions." *Socio-Economic Review* 17(2): 283–309.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. 2018. "Human decisions and machine predictions." *Quarterly Journal of Economics* 133(1): 237–293.

Knox, D., W. Loew, and J. Mummolo. 2020. "Administrative Records Mask Racially Biased Policing." *American Political Science Review* 114(3): 619 – 637. https://doi.org/10.1017/S0003055420000039.

Kusner, M. and J. Loftus. 2020. "The Long Road to Fairer Algorithm." *Nature* 578: 34-37.

Le Bui, M. and S. Noble. 2020. "We're Missing a Moral Framework of Justice in Artificial Intelligence." *The Oxford Handbook of Ethics of AI*.

Lippert-Rasmussen, K. 2013. *Born Free and Equal?: A Philosophical Inquiry into the Nature of Discrimination*. Oxford: Oxford University Press.

Long, R. 2021. "Fairness in machine learning: Against false positive rate equality as a measure of fairness." *Journal of Moral Philosophy*. doi: https://doi.org/10.1163/17455243-20213439

Mayson, S. 2019. "Bias In, Bias Out." *Yale Law Journal* 128(8): 2122-2473.

Mills, C. 1999. *The Racial Contract*. Ithaca: Cornell University Press.

Mills, C. 2017. *Black Rights/White Wrongs: The Critique of Racial Liberalism*. New York: Oxford University Press.

Munton, J. 2019. "Beyond Accuracy: Epistemic Flaws with Statistical Generalizations." *Philosophical Issues* 29(1): 228-240.

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

O'Neill, Cathy. 2016. *Weapons of Math Destruction*. New York: Penguin.

Passi, S. and S. Barocas. 2019. "Problem Formulation and Fairness." Conference on Fairness, Accountability, and Transparency (FAT* '19), January 29-31, 2019, Atlanta, GA, USA.

Rodrik, D. 2009. "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In *What Works in Development?* Cohen and Easterly (eds.). Washington DC: The Brookings Institute, 24-48.

Richardson, R., J. Schultz, and K. Crawford. 2019. "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice."

Rawls, J. 1999. A Theory of Justice. Revised edition. Cambridge: Harvard University Press.

Yoav Shoam et al. 2018. "The AI Index 2018 Annual Report," AI Index Steering Committee, Human-Centered AI Initiative, Stanford University.

Selmi, M. 2013. "Indirect Discrimination and the Anti-discrimination Mandate." In *Philosophical Foundations of Discrimination Law*, ed. D. Hellman and S. Moreau, 250-268. Oxford: Oxford University Press.

Sweeney, L. 2013. "Discrimination in Online Ad Delivery." arXiv:1301.6822v1 [cs.IR].

Spiekermann, K. 2021. "Good Reasons for Losers: Lottery Justification and Social Risk." *Economics and Philosophy*.

Obermeyer, Z., B. Powers, C. Vogeli, S. Mullainathan. 2019. "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 6464: 447-453.

Westen, P. 1982. "The empty idea of equality." *Harvard Law Review*, 537–596.

Weaver, A. 2015. "Is Credit Status a Good Signal of Productivity?", *Industrial and Labor Relations Review* 68: 742–770.

Zimmerman, A. and C. Lee-Stronach. 2021. "Proceed with Caution." *Canadian Journal of Philosophy*.

Zuboff, S. 2019. *Surveillance Capitalism*. Profile Books.