Measuring Fairness in an Unfair World

Jonathan Herington
Department of Philosophy
University of Rochester
Rochester NY USA
Jonathan.Herington@rochester.edu

ABSTRACT

Computer scientists have made great strides in characterizing different measures of algorithmic fairness, and showing that certain measures of fairness cannot be jointly satisfied. In this paper, I argue that the three most popular families of measures – unconditional independence, target-conditional independence and classification-conditional independence – make assumptions that are unsustainable in the context of an unjust world. I begin by introducing the measures and the implicit idealizations they make about the underlying causal structure of the contexts in which they are deployed. I then discuss how these idealizations fall apart in the context of historical injustice, ongoing unmodeled oppression, and the permissibility of using sensitive attributes to rectify injustice. In the final section, I suggest an alternative framework for measuring fairness in the context of existing injustice: distributive fairness.

CCS CONCEPTS

• **Social and professional topics ~ Computing / technology policy** • Computing methodologies ~ Machine learning

KEYWORDS

Fairness; algorithmic decision-making; distributive justice; causal inference; discrimination

ACM Reference format:

Jonathan Herington. 2020. Measuring Fairness in an Unfair World. In *Proceedings of 2020 AAAI-ACM Conference on Artificial Intelligence, Ethics, and Society (AIES'20), February 7-8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3375627.3375854

AIES'20, February 7–8, 2020, New York, NY, USA.
© 2020 Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-7110-0/20/02…$15.00
DOI: https://doi.org/10.1145/3375627.3375854

1 Introduction

Developers of machine algorithms have started to grapple with the complex tradeoffs involved in detecting and eliminating bias in their creations. Some work has been done in computer science, statistics and criminology to generate measures of bias before, during or after deployment of an algorithm [7], Likewise, there has been work to show that biased algorithms may violate U.S. civil rights law because of their "disparate impact" upon minority groups [6]. Unfortunately, whilst computer scientists have made some strides in identifying appropriate measures of bias that operate in the abstract, there has been little engagement with the problem of diagnosing unfairness in the context of an already unjust world (c.f. [14]).

In what follows, I suggest that historical and contemporary oppression on the basis of race, sex and other sensitive attributes means that the diagnosis of bias cannot be divorced from the context in which an algorithm is deployed. Measures of algorithmic bias assume that an algorithm which is fair in the abstract will be fair in the world. This is a fatal mistake. As the long history of discrimination, inequality and oppression teaches us, policies that are just in ideal circumstances often contribute to oppression if deployed in the context of existing injustice. Measures of algorithmic bias are no exception.

2 Unfairness in Ideal Conditions

Roughly put, given the attributes of an object or individual, i, a subset, $\{M\}$, of those attributes will be attended to by the algorithmic model as it generates a classification, C, for i that reflects the probability that i will possess some target property, T. These classifications can then be used, by themselves or in conjunction with other factors, to make decisions (e.g. "approve loan" or "do not approve loan") which result in certain kinds of treatment (i.e. the disbursement of a loan, the serving of an advertisement, etc.). In this way, classification algorithms contribute, alongside many other factors, to the distribution of important benefits and burdens (e.g. loans, parole, medical care, etc.).

In part because algorithms control the distribution of important goods, algorithmic developers take great pains to try to avoid making classifications on the basis of various sensitive attributes (e.g. race, gender, sexual orientation, etc.). The underlying philosophical justification for these efforts is rarely articulated [8], but they all implicitly assume that algorithmic bias is wrongful when and because sensitive attributes *cause* classifications. Abstracting away from the messy reality of existing injustice, the computer science community has focused on measures and methods that assume that unfairness must flow through the algorithmic model and the classification.

Indeed, we can construct a model of the *ideal* set of causal relationships (or lack thereof) between the algorithmic model, target property, classification and sensitive attributes. Causal graphical models [25,29] allow us to represent this ideal causal structure, as well as explore deviations from them. Consider an algorithmic model that considers a set, {M}, of features to determine a classification, C, which represents the probability that the subject being evaluated will possesses the target property, T. We can represent {M}, C, and T as a set of nodes in a graph. Likewise, we can represent causal relationships between these nodes with directed edges. Thus, in so far as the model feature set causes the classification, we represent this with a directed edge between nodes {M} and C. Moreover, since we hope that the algorithmic classifications are accurate, we also suppose that the model variables cause, are caused by, or share an unknown common cause with, the target property, T. All three of these causal structures can be represented by a causal graph. For simplicity sake, I assume that the model variables are causes of the target property (as represented in Fig 1).

The next step is to introduce a sensitive attribute variable, A. Ideally, the sensitive attribute will have a causal connection to neither the model variables, the classification, nor the target property. This is a compelling liberal-egalitarian ideal, but as we shall see, it can break down in important ways.

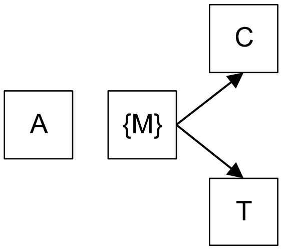

Figure 1: The ideal causal structure for algorithmic classification. A is the sensitive attribute, {M} a set of model features, C the classification score, and T the target property.

The computer science community have mostly been concerned with one particular failure of this ideal structure: where the sensitive attribute is a cause of some model feature (i.e. as in Fig 2, where A → M_A). Drawing on the legal theory of disparate impact, it is well recognized that seemingly benign model variables can skew classifications so that otherwise similar individuals of different races/genders receive wildly different scores [6]. Since sex, race and other sensitive attributes hold incredible sway over people's lives, many model variables may be causally influenced by sensitive attributes. In order to avoid unintentionally allowing A to cause C, computer scientists have therefore generated a suite of different measures of bias to aid in the diagnosis and elimination of A's causal influence over C.

Although there are at least twenty different measures proposed in the literature [7,10], they can be roughly categorized into three classes: (i) measures of unconditional independence, (ii) measures of target-conditional independence, and (iii) measures of classification-conditional independence [5,14]. Each class of measure is defined by the particular statistical associations that it tests between C, T and A. I briefly review each category of measure and their implicit assumptions below.

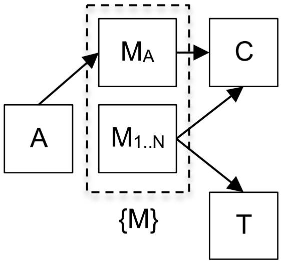

Figure 2: The standard deviation from the ideal. A causally influences some feature M_A which influences C but not T. This manifests as a statistical association between A and C, via the model set {M} (but no association between A and T).

2.1 Measures

Measures of unconditional independence require that classifications are statistically independent from sensitive attributes (race, sex, etc.) – e.g. that the distribution of scores for a particular race is roughly equivalent to the distribution of scores for other races. Examples of such measures include "demographic parity" [18], "statistical parity" [13] and "anti-classification" [10]. A failure of these measures is intuitively troubling because it suggests that there is an open causal pathway between A and C – e.g. race is partially causing the algorithm's predictions. Since, in ideal circumstances, it would be unfair for race (or other sensitive attributes) to influence our treatment of individuals, algorithms which fail to satisfy independence are also thought to be unfair.

Note that measures of independence make a key assumption: that *any* causal influence A wields over C is illegitimate. While this may accord with liberal egalitarian ideal, as we shall see, it is morally dubious in the context of existing injustice and attempts to rectify it.

Even letting this assumption stand, the technical literature is replete with papers that point out that measures of independence have unattractive features [12:8]. The most obvious of these is that independence is sensitive not only to bias introduced by the algorithm, but also to genuine differences in the distribution of the target property. For example, if the rate of malignancy in skin lesions is higher for women than men, then a perfectly accurate algorithm will classify women as higher risk for malignancy than men. Such an algorithm would fail to satisfy measures of independence, but only because it perfectly reflects the unequal distribution of malignancy in the actual world. This problem has motivated two kinds of conditional measures that seek to control for base rate differences.

Measures of target-conditional independence require that, conditional upon the actual value of the target property (e.g. malign or benign), classification scores are statistically independent from sensitive attributes. Examples of such measures include "balance" [19], and "predictive equality"[11]. By measuring the independence of A and C conditional on T, we control for differences in the distribution of the target property between groups. The hope, therefore, is that we can control for any statistical association between A and C that is due to the algorithm capturing any causal association between A and T. Any residual statistical association is thought to be illegitimate, in part because it gives rise to different rates of Type I and II errors [17].

Likewise, measures of classification-conditional independence require that the target property is statistically independent of the sensitive attributes (race, sex, etc.), conditional upon their classification (e.g. positive or negative). Examples of such measures include "test-fairness" [9], and "calibration" [19]. Measures of classification-conditional independence eliminate base rate errors by controlling for differences in the distribution of scores between groups. The hope is that we can show that C is tightly *calibrated* with T, such that learning about A would not increase (or decrease) the probability of possessing the target property. In other words, all groups enjoy classifications of roughly the same degree of accuracy.

Note that both kinds of conditional measures make two key assumptions.

i. they assume that the causal influence of A over T is directed through the model variables, and hence also causes C.
ii. they assume that whatever causal influence A wields over T is legitimate, and hence it is fair to allow A to have the same causal influence over C.

As I argue below, in the context of an unjust world, neither assumption is sustainable.

Importantly, these three kinds of measures cannot be satisfied simultaneously when the prevalence of the target property differs between groups [5,9,19]. Moreover, they each rely upon different normative justifications and background assumptions. In so far as we must choose a definition of fairness to apply, we ought to understand whether those background assumptions apply in the actual, unjust world where algorithms are deployed.

3 Unfairness in Non-Ideal Conditions

Much of the work done by statisticians and computer scientists to define these measures of bias has assumed that ideal, or relatively benign, causal structures operate in the background. But this assumption cannot be sustained. Building on prior work [14], I recount three ways in which existing injustice causes key assumptions of these measures of fairness fail: (a) historical injustice, (b) unmodelled injustice, and (c) the legitimacy of using sensitive attributes to rectify both kinds of injustice.

3.1 Historical Injustice

First, historical injustice means that we cannot assume that associations between A and T are legitimate. Recall that conditional measures, by controlling for S or T, implicitly assume that it is legitimate for A to cause S up to the degree that A causes T. A's causal influence over T is therefore taken to be appropriate (or at best, a background condition outside the scope of study).

In the absence of oppression, this may be a safe assumption: since we might hypothesize that any association between A and T is due to benign differences in group preferences or accidental regularities in behavior. In an unjust context though, these differences are anything but benign or accidental. Race has been a site of immense structural injustice, that has fundamentally constrained almost every black American's life (the same, *mutatis mutandis*, for women under patriarchy) [2]. In the context of historical injustice, we thus ought to be skeptical of claims that differences in the distribution of benefits and burdens between sexes or races are the result of "natural" differences in capacities or interests. Since historical injustice is likely responsible for race- and gender-based inequality, these inequalities are likely to be illegitimate [2,23:55–60].

By assuming that fairness consists in measuring only that quantum of causal influence between A and C that is not due to the influence of A over T, these measures implicitly ask us to endorse the historical influence of A over T. In the context of injustice this amounts to endorsing (or assuming away) historical oppression.

3.2 Unmodeled Injustice

Second, the unjust circumstances in which we live can result in the sensitive attribute having causal influence over the target property in ways that do not affect any model variables (Fig. 3.). The most obvious example where this may be true is in the context of explicit racial bias. If explicitly racist police officers are more likely to arrest an individual *simply* because they are black, then this may not be fully captured by changes in any algorithmic model variable, since the method by which race causes arrest is direct. It might also occur in the context of implicit attitudes. If employees at Company A have implicit attitudes of disgust towards the disabled, then they may be less likely to collaborate with them, advocate for their work, or boost their morale. While the absence of these things might cause poor job performance at Company A, it may *not* cause changes in the model variables that are used to predict the job performance of applicants to Company A (i.e. education level, psychometric tests, number of prior jobs, etc.).

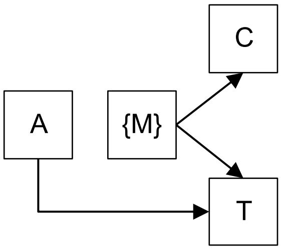

Figure 3: An example of unmodelled injustice. Note that the causal influence of A over T is not captured by {M}.

Note that if the influence of the sensitive attribute on the target is fully mediated by model variables (e.g. family postcode, history of drug use, education level) then this edge will disappear into an edge between A → {M}. But this happy case is rare. The model variables available to a classification algorithm are unlikely to fully capture all the different mechanisms of contemporary oppression.

In such circumstances, the algorithm will fail to satisfy target conditional independence. This is because there is an "open" causal pathway between A and C (i.e. A → T ← {M} → C).[1] The same is obviously true for measures of classification-conditional independence (since A and T are directly associated, A → T). It is not true, however, for measures of unconditional independence (i.e. T is a collider that "blocks" the open path A → T ← {M} → C).

[1] Recall that a pathway is open if every node is "on", and a node is "on" if (i) it is *not* in the conditioning set and it is an initiator, a mediator, or a terminator, or (ii) it is in the conditioning set and is a collider. [29]

Note, however, that the failure to satisfy conditional measures is not because the algorithm's *classification* is caused by the sensitive attribute. Rather, any association between A and C is wholly induced by conditionalizing on the target property. In this respect, the existence of oppression in the societies where these algorithms are deployed mean conditionalized measures may simply be diagnosing existing bias that is uncaused and unaffected by the algorithm's classification.

Coupled with the problem of historical injustice, this makes it hard to defend the continued use of conditional measures of fairness. Consider that the only time conditional measures are sensitive to historical oppression in the distribution of a target property is when that oppression is not captured by model variables. If we can "capture" the existence of historical or ongoing oppression in the set of model variables, then it ceases to appear as unfair according to conditional measures. This inconsistency should be fatal to the use of conditional measures in the context of injustice.

3.3 Rectification of Injustice

Finally, this confluence of unjust circumstances sometimes makes it legitimate for sensitive attributes to cause classifications. The classic example is affirmative action policies that make hiring or admission decisions explicitly based on race in order to compensate for the historical disenfranchisement and exclusion of blacks from education and professional employment. Rectifying historical injustices, in so far as it involves the redistribution of resources, will necessarily require statistical associations between attributes and classifications. More subtly, in the context of injustice we might think unintended associations between sensitive attributes and classifications are legitimate if they prioritize aid to the most vulnerable. Consider a simple algorithm used to predict an elderly population's need for health support services. It would be unsurprising to find an association between those scores and race. What matters here is the direction of the association (i.e. is it biased towards whites or blacks), and its effect on all things considered equality between races, genders and other sensitive attributes.

This is important because it undermines the main assumption of measures of independence: that *any* causal association between A and C is illegitimate. While we might endorse this assumption in the context of ideal background conditions – where oppression on the basis of race and sex never occurred – it is indefensible in the context of historical injustice.

4 Distributive Fairness

If the idealizations required by the most popular measures of fairness do not hold in an unjust world, then how should we measure fairness in the context of existing injustice? In what follows, I propose a family of measures that seeks to capture a particular kind of fairness especially relevant in the context of historical injustice: "distributive fairness".

Distributive fairness is the idea that members of a just society ought to enjoy a fair share of the benefits and burdens produced by that society [20]. Like the measures of fairness popular in the machine learning literature, it captures the intuition that one's race, gender, religious commitments, etc. should not determine how one is treated. But instead of focusing on how individuals are classified by the algorithm, it focuses on how the use of the algorithm changes the distribution of benefits and burdens in society more broadly.

Thinking about fairness in this way allows us to draw on a rich philosophical literature on the nature of distributive justice. In particular, there is a deep debate over the kinds of benefits and burdens we ought to distribute equally. Some people have the intuition that, all other things being equal, we ought to minimize inequality of income, wealth, resources or wellbeing [30]. Others, skeptical that we can eliminate inequalities in wellbeing, want to eliminate inequalities in opportunities or luck [4,26]. Still others want to eliminate inequalities in our status or power as citizens [3]. This debate over the nature of distributive justice is an underutilized resource in the discussion over fair ML (c.f. [8,16]), and suggests a new family of measures under the broad classification of distributive fairness.

Importantly, measures of distributive fairness will be structurally related to measures of unconditional independence. Since algorithmic classifications often distribute important benefits and burdens, associations between A and C may also contribute to inequalities in those outcomes. We can model this connection by simply extending our basic representations of algorithmic causal structures to include these outcomes, O.

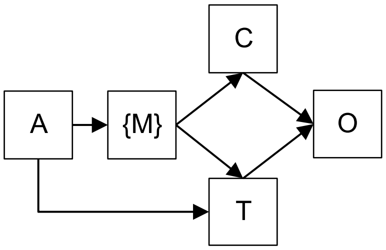

Figure 4: A complex causal association between a sensitive attribute, A and an outcome, O. Note that both classification, C, and target, property, T, mediate the causal influence of A over O

Importantly, the kinds of outcomes relevant to justice often have complex causal relationships with the model variables and the target property (see Fig. 4). For instance, individual wealth is not only dependent on whether one's loan is approved, but also whether or not one defaults on the loan (i.e. the target property). Likewise, one's wealth can also be independently determined by some of the variables that the model uses to generate the classification (e.g. current income or employment history could influence one's future wealth, regardless of the loan approval process, or whether or not one defaults). Or all three of the classification, model variables and target property may influence the outcome.

The virtue of measures of distributive fairness is that they do not require careful screening of these causal structures. Precisely how an attribute is causally associated with a benefit or burden is irrelevant. Instead, what matters is that there is such an association, and that it results in a widening rather than a narrowing of inequality on the basis of the attribute. In place of complex causal analysis, there are two value judgements that must be made to generate a fully-formed measure of distributive fairness: (i) the currency of justice, and (ii) the demandingness of justice.

4.1 Currency of Justice

First, developers and policymakers must make a value-laden judgement about the particular outcome, O, that is relevant. As we noted above, there is a long-standing philosophical literature about the appropriate "currency" of benefits and burdens that must be fairly distributed. In the simplest case, one might want to equalize the distribution of some pertinent primary resource (e.g. income, wealth) or all-things-considered wellbeing (e.g. health, preference satisfaction). But there are well established reasons – including the so-called "conversion" problem [27] – that one might be skeptical of pure egalitarianism with respect to material resources or wellbeing. One could, alternatively, focus on inequalities in opportunity: that the probability of realizing wellbeing should be dependent on a function of one's talents and effort [16,26]. More radically, others have claimed that we owed individuals equality of luck: that their probability of realizing wellbeing should be independent of all those features outside of their control (including their "natural" talents) [4,15]. Others argue that neither luck nor opportunity for wellbeing are appropriate, and we should instead focus on inequalities in individuals *capability* to realize different ways of living a life [1,22,28]. Still others, argue for equality of *citizenship*: that we owe individuals those resources necessary to be "equal democratic citizens" in relation to others [3].

Importantly, almost all of these theories of justice are concerned with inequality in wide-scope outcomes (i.e. the distribution of wealth) rather than narrow-scope outcome (i.e. the allocation of credit) [21]. If the currency of justice is a broad social indicator – e.g. income, wealth, "primary goods", wellbeing or capabilities – then it is the effect of the algorithm on those goods which is relevant. Moreover, these wide scope outcomes are the result of very complex social processes, which may weaken (or even reverse) the effect of an inequality in a narrow scope outcome. For instance, the effect of being denied a loan on your lifetime wealth is mediated by a host of factors – your prior wealth, your income, your family status, etc. – that may weaken or strengthen the effect of the denial of the loan. In this respect the influence of an algorithmic classification on the possession of a wide-scope outcome is complex in the extreme. As we noted above, however, the precise causal structure which results in these wide scope outcomes is, in general, not relevant to an evaluation of distributive fairness. What matters is that the pattern of distribution of the relevant wide-scope outcome – i.e. income – is fair.

Of course, it will sometimes be the case that normative significance attaches to inequality in a narrow-scope outcome. For instance, if an algorithm which predicts which voters in a registration list are most likely to be ineligible is not carefully invigilated, it may create an inequality in a narrow scope outcome – i.e. expungement of a voter registration – which is unjust, regardless of its effect on other wide scope outcomes. But even when the currency of justice is more tightly defined – e.g. inequalities in voting rights – the causal influence of the classification on the outcome may be swamped by other causal factors.

4.2 Demandingness

Second, developers and policymakers face a value judgement about the demandingness of distributive fairness. To understand why, consider that distributive fairness is determined by analyzing the effect of the algorithm on the pre- and post-deployment distribution of the relevant outcome for different values of the sensitive attribute. To test this consideration, developers must engage in a two-step analysis. In the first stage, they must identify the distribution of the outcome prior to the deployment of the algorithm. In the simple case where O is a binary outcome, this is captured by the difference in the conditional probability of realizing the outcome before deployment, O_{PRE}, for different values of the sensitive attribute (i.e. $\Delta(O_{PRE}) = Pr[O_{PRE}|A=w] - Pr[O_{PRE}|A=b]$). In the second stage, they discern whether the deployment of the algorithm would widen or close any inequality that existed in the first stage (i.e. $\Delta(O_{PRE}) - \Delta(O_{POST})$).[2] In this respect, measuring distributive fairness requires developers to estimate the degree of dependency between A and O both before and after the algorithm is deployed.

Having established the effect of deployment, developers and policymakers face a value judgement about the demands of distributive fairness. Three views predominate. On the least demanding view, an algorithm is distributively unfair only if it *increases* inequality in the distribution of the relevant outcome. For instance, taking into account existing racial disparities in health, we might constrain a health services provision algorithm so that it does not widen those racial disparities [c.f. 24]. This account of the demandingness of distributive fairness is similar to current measures of fairness, which seek only to avoid creating new inequalities or dependencies. Accordingly, this view is vulnerable to the charge that it implicitly treats existing inequalities as legitimate and just. In this respect, while any algorithm that satisfies this constraint would avoid the problem of inadvertently widening or deepening inequality, it treats the status quo as morally legitimate in a way that is philosophically dubious.

Alternatively, on the most demanding view, an algorithm is distributively unfair if it fails to *minimize* inequality in the distribution of the relevant outcome. For instance, taking into account existing racial disparities in health, we might tune a health services provision algorithm to maximize its impact in reducing those racial disparities. While this view is more philosophically defensible, it is also more likely to force developers and policymakers to make a tradeoff between accuracy and fairness. Consider that in some cases, owing to structural or data availability constraints, we might be able to increase the accuracy of health resource targeting for one group (i.e. men) but not other groups (i.e. women, non-binary persons, etc.). If we hold that fairness demands that we minimize gaps in health outcomes between groups, then we must forego improvements in the accuracy of the algorithm for the sake of fairness. This is known as the "levelling-down objection" to strict equality [30:247]. Given the intuitive power of this objection, more work is needed to assess the precise boundaries and justifiability of this tradeoff given different currencies of justice and deployment contexts.

On a final view, an algorithm is distributively unfair if it fails to *maximise* the enjoyment of the relevant outcome for the worst-off group (and the next best-off group if there are multiple versions of the algorithm that would equally benefit the worst-off). For instance, taking into account existing racial disparities in health, we might tune a health services provision algorithm to maximize its impact in improving the health outcomes of the worst-off racial group. This view avoids the levelling-down objection, since it allows us to accept inequalities so long as we have done as well as we could for the least well-off. Nonetheless, while this kind of

[2] For non-binary outcomes, which includes most outcomes of interest, more complicated expectations and comparisons must be made. I leave these extensions for later work.

prioritarian principle has a long history and is intuitively attractive [26], it may be inappropriate for some kinds of currencies of justice. Consider that we should demand strict equality for certain kinds of outcomes – e.g. voting access. More work is therefore needed to assess the special cases where strict equality is demanded by the particular currency of justice.

5 Conclusion

In this paper I have cast doubt on the utility of standard measures of algorithmic fairness. I have argued that the idealization and assumptions which make these measures intuitively attractive in the abstract, fatally undermine their plausibility in the context of existing injustice. To remedy this problem, I suggest that we abandon accounts of fairness that rely upon diagnosing certain causal pathways between attributes and classifications. Instead, we ought to assess the fairness of algorithms according to their impact on overall distributive justice. By doing so our measures of algorithmic fairness will avoid committing to erroneous causal assumptions, be better able to deal with historical injustice, and more clearly specify the value judgements that underpin the choice of measure.

ACKNOWLEDGMENTS

Special thanks to Bruce Glymour for collaborating on earlier work [14] that informs this piece and for introducing me to thinking about discrimination with causal graphs. All errors of understanding are my own.

REFERENCES

[1] Sabina Alkire. 2002. *Valuing Freedoms: Sen's capability approach and poverty reduction*. Oxford University Press, Oxford.

[2] Elizabeth Anderson. 2010. *The Imperative of Integration*. Princeton University Press, Princeton, NJ.

[3] Elizabeth S. Anderson. 1999. What Is the Point of Equality? *Ethics* 109, 2 (January 1999), 287–337. DOI:https://doi.org/10.1086/et.1999.109.issue-2

[4] Richard J. Arneson. 1989. Equality and Equal Opportunity for Welfare. *Philos. Stud. Int. J. Philos. Anal. Tradit.* 56, 1 (May 1989), 77–93.

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. *Fairness and Machine Learning*. Retrieved from fairmlbook.org

[6] Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *Calif. Law Rev.* 104, (2016), 671–732.

[7] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. *ArXiv170309207 Stat* (March 2017). Retrieved November 2, 2017 from http://arxiv.org/abs/1703.09207

[8] Reuben Binns. 2017. Fairness in Machine Learning: Lessons from Political Philosophy. *ArXiv171203586 Cs* (December 2017). Retrieved July 12, 2018 from http://arxiv.org/abs/1712.03586

[9] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *arXiv:1610.07524 [cs, stat]*. Retrieved November 7, 2017 from http://arxiv.org/abs/1610.07524

[10] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *ArXiv180800023 Cs* (July 2018). Retrieved August 22, 2018 from http://arxiv.org/abs/1808.00023

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. *ArXiv170108230 Cs Stat* (January 2017). DOI:https://doi.org/10.1145/3097983.309809

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. *ArXiv11043913 Cs* (April 2011). Retrieved November 16, 2018 from http://arxiv.org/abs/1104.3913

[13] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '15), ACM, New York, NY, USA, 259–268. DOI:https://doi.org/10.1145/2783258.2783311

[14] Bruce Glymour and Jonathan Herington. 2019. Measuring the Biases That Matter: The Ethical and Causal Foundations for Measures of Fairness in Algorithms. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT* '19), ACM, New York, NY, USA, 269–278. DOI:https://doi.org/10.1145/3287560.3287573

[15] Vivek Gupta, Pegah Nokhiz, Chitradeep Dutta Roy, and Suresh Venkatasubramanian. 2019. Equalizing Recourse across Groups. *ArXiv190903166 Cs Stat* (September 2019). Retrieved December 18, 2019 from http://arxiv.org/abs/1909.03166

[16] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, ACM Press, Atlanta, GA, USA, 181–190. DOI:https://doi.org/10.1145/3287560.3287584

[17] Deborah Hellman. 2019. *Measuring Algorithmic Fairness*. Social Science Research Network, Rochester, NY. Retrieved July 22, 2019 from https://papers.ssrn.com/abstract=3418528

[18] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.). Curran Associates, Inc., 656–666. Retrieved November 16, 2018 from http://papers.nips.cc/paper/6668-avoiding-discrimination-through-causal-reasoning.pdf

[19] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Proceedings of Innovations in Theoretical Computer Science (ITCS)*. Retrieved November 7, 2017 from http://arxiv.org/abs/1609.05807

[20] Julian Lamont and Christi Favor. 2017. Distributive Justice. In *The Stanford Encyclopedia of Philosophy* (Winter 2017), Edward N. Zalta (ed.). Metaphysics Research Lab, Stanford University. Retrieved December 17, 2019 from https://plato.stanford.edu/archives/win2017/entries/justice-distributive/

[21] Kasper Lippert-Rasmussen. 2006. The badness of discrimination. *Ethical Theory Moral Pract.* 9, 2 (April 2006), 167–185. DOI:https://doi.org/10.1007/s10677-006-9014-x

[22] Martha C. Nussbaum. 2000. *Women and Human Development: The Capabilities Approach*. Cambridge University Press, Cambridge.

[23] Martha C. Nussbaum. 2000. *Sex and Social Justice* (First Edition edition ed.). Oxford University Press, Oxford New York Athens.

[24] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (October 2019), 447. DOI:https://doi.org/10.1126/science.aax2342

[25] Judea Pearl. 2009. *Causality: Models, Reasoning and Inference* (2nd ed.). Cambridge University Press, New York.

[26] John Rawls. 2001. *Justice as Fairness: A Restatement*. Belknap Press, Cambridge, MA.

[27] Amartya Sen. 1983. *Poverty and Famines: An Essay on Entitlement and Deprivation*. Oxford University Press, Oxford. Retrieved from http://books.google.com.au/books?id=FVC9eqGkMr8C

[28] Amartya Sen. 1993. Capability and Wellbeing. In *The Quality of Life*, Amartya Sen and Martha C. Nussbaum (eds.). Clarendon Press, Oxford, 31–53.

[29] Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search* (2nd ed.). MIT Press, Cambridge, Mass.

[30] Larry S. Temkin. 1993. *Inequality*. Oxford University Press, New York, NY, USA.