

A Systematic Approach to Group Fairness in Automated Decision Making

Corinna Hertweck

Zurich University of Applied Sciences,
University of Zurich
Zurich, Switzerland
corinna.hertweck@zhaw.ch

Christoph Heitz

Zurich University of Applied Sciences
Winterthur, Switzerland
christoph.heitz@zhaw.ch

Abstract—While the field of algorithmic fairness has brought forth many ways to measure and improve the fairness of machine learning models, these findings are still not widely used in practice. We suspect that one reason for this is that the field of algorithmic fairness came up with a lot of definitions of fairness, which are difficult to navigate. The goal of this paper is to provide data scientists with an accessible introduction to group fairness metrics and to give some insight into the philosophical reasoning for caring about these metrics. We will do this by considering in which sense socio-demographic groups are compared for making a statement on fairness.

Index Terms—algorithmic fairness, group fairness, statistical parity, independence, separation, sufficiency

I. INTRODUCTION

The field of algorithmic fairness is, to a large extent, concerned with the fairness of Automated Decision-Making (ADM) systems. ADM refers to the process of making decisions about individuals in an automated way based on the data that is available about these individuals. Examples are the automated approval of credit loan applications based on, e.g., the credit loan history and income of the applicant, the automated acceptance to universities based on grades and standardized test scores or the automated invitation of applicants to job interviews based on their CV. ADM typically involves a prediction \hat{Y} of a non-observable feature Y of individuals. An example of this would be the prediction \hat{Y} of whether or not a credit loan applicant would repay the loan they applied for. The automated decision is then based on this prediction. In many instances, it has been shown that such prediction-based decision systems lead to unintended unfairness and discrimination (see, e.g., [1], [2] for specific cases of unintended algorithmic discrimination or see [3] for an overview of the topic). These instances have made clear that the fairness properties of such decision systems have to be studied since fairness is not achieved automatically, but has to be consciously implemented.

Our paper provides data scientists with an entry point to the many metrics that fall into the category of *group fairness*. Measures in this category evaluate ADM system’s decisions with respect to different socio-demographic groups (e.g., different gender or racial groups). The basic question is: *Are there systematic differences in the decisions for the groups?* When evaluating fairness through group fairness metrics, we thus

always compare groups and look for differences in treatment. We analyze the metrics not by focusing on their mathematical definitions, but by looking at how exactly the decisions for socio-demographic groups are compared to each other by the given fairness metrics as this provides a basis for a moral interpretation. To help data scientists answer the question “What fairness measures should I care about, and why?”, we will provide a short (and necessarily incomplete) introduction to the philosophical arguments that have been made about the discussed fairness metrics.

II. RELATED WORK

The fairness discussion is not a new discussion that appeared with the increased usage of ADM systems. Quite the opposite is true: Questions of fairness and justice have been discussed for thousands of years, with a long history in philosophy, sociology, law etc. However, ADM systems require quantitative measures of fairness, and the algorithmic fairness literature of the last years has brought forward many different such measures: In [4], Arvind Narayanan discusses 21 fairness metrics, but notes that there are many more. The fairness auditing tool AIF360 claims to “[include] over 71 bias detection metrics” [5, p. 2]. However, it can be difficult to understand the specifics of these different fairness measures that have been proposed in the literature and are implemented in state-of-the-art tools. Some publications try to categorize fairness measures (see, e.g., [6], [7]), but these works usually lack intuitive explanations of the moral reasons to enforce the discussed fairness measures and instead focus on their mathematical definitions. This may create confusion on the side of practitioners about how to implement fairness in their data-based decision algorithms.

There is already a considerable amount of literature that discusses different fairness measures and in particular group fairness measures, which we focus on in this paper. Since most group fairness metrics are calculated from the values found in confusion matrices, existing literature oftentimes approaches the introduction of different group fairness measures from this angle: They show how these measures can be calculated from the confusion matrix (see, e.g., [6], [7], [8], [9], [10], [11]). Recent research (see [12], [13]), however, experimentally showed that confusion matrices are not intuitive and that it

takes time to fully understand them. From this, we conclude that it is hard to grasp the moral meaning of fairness metrics based on confusion matrices. Another explanatory approach is chosen in [14]: Here, the authors define common group fairness measures as (conditional) independence statements. While this approach is formally very elegant, it is difficult to interpret with respect to the concept of fairness and justice that is represented by each metric.

Our paper attempts to introduce group fairness metrics in a more intuitive way by showing that each metric corresponds to an analysis of the decisions made about a specific part ("subpopulation") of the population on which the system operates. For example, we might consider all members of the population, or only those who are negatively affected by the decision algorithm. We argue that it is exactly this choice of the subpopulation which establishes which theory of fairness a specific metric aligns with.

III. WHAT IS A "FAIR ADM SYSTEM"?

The question of what constitutes fairness is a heavily debated philosophical question. In this section, we will first show that the intuitive interpretation of fairness as non-discrimination cannot be easily applied to the field of algorithmic fairness. We then introduce another philosophical concept, egalitarianism, which can be applied to understand when ADM systems are considered to be unfair.

A. Non-discrimination

Often, fairness is equated with non-discrimination, which is legally defined in anti-discrimination laws. Article 8 of the Swiss federal constitution, for example, states that "no person may be discriminated against, in particular on grounds of origin, race, gender, age, language, social position, way of life, religious, ideological, or political convictions, or because of a physical, mental or psychological disability." The law thus defines several properties on which discrimination is specifically prohibited. We will refer to these as *sensitive attributes* (e.g. origin, race and gender). While equating fairness with non-discrimination seems intuitive, it turns out that the *philosophical* concept of discrimination is actually not applicable to "unfair" machine learning (see [15] for a detailed analysis). One of the reasons that Reuben Binns [15] gives for this is that one may see discrimination as morally wrong only when it is intentional. However, it is difficult to speak of intention in the case of ADM systems which we would typically not portray as conscious, autonomous actors. The questions that this finding leaves us with are: What is fairness in ADM systems if the philosophical concept of non-discrimination cannot be applied? On what philosophical grounds should an ADM system be assessed?

B. Egalitarianism

A philosophical concept that might better help us understand when ADM systems are unfair is *egalitarianism*. Binns describes egalitarianism as "the idea that people should be treated equally, and (sometimes) that certain valuable things should

be equally distributed" [15, p. 5]. As he further argues, this does not prohibit inequalities. Distributive inequalities might sometimes be justified and considered to be fair. What fair inequalities are cannot be objectively determined though – instead it depends on the philosophical theory of justice one holds. One may, for example, argue that differences between people are only just if they are not caused by brute luck, such as being born into a wealthy family (this view is known as luck egalitarianism and defended by philosophers such as Ronald Dworkin and G. A. Cohen, see, e.g., [16], [17]). On the other side, one could claim that such inequalities are fair under certain circumstances (a definition of such circumstances is given and defended by the American philosopher John Rawls [18]). To summarize, fairness in ADM systems can be understood as some form of equal treatment or equal distribution between people. People, however, disagree about what constitutes equal treatment, what justifies inequalities and how this should be measured – which is how we end up with the many fairness metrics we find in the literature [19].

IV. MEASURING FAIRNESS

Data-based decision algorithms typically operate in a context where only partial information is available at the time of decision making: The missing information Y is not available and thus replaced by a prediction \hat{Y} , which is then used to take a decision. However, this means that decisions might be wrong on the individual level: a bank might give a loan to someone who does not pay this loan back; a student might be accepted to a college even though they do not succeed in their studies there. When assessing the fairness of a decision system, we may thus not be interested in random mistakes, but in systematic patterns in the decision. In other words: We may call a decision algorithm unfair if it *systematically* disadvantages one group with respect to the other. This concept is known as *group fairness*.

Since the decisions in the end still affect individuals, both the philosophical theories of justice as well as the law tend to discuss justice with respect to individuals. The algorithmic fairness literature actually provides approaches for measuring fairness on an individual level. The concept of *individual fairness*, for example, asks whether "similar" individuals receive "similar" decisions (it is up to the decision-maker to define what individuals and what decisions are considered to be "similar") [20]. *Counterfactual fairness* asks whether individuals would have received the same decision if they were (born as) a member of another socio-demographic group [21].

In this paper, we focus on group fairness measures which assess whether a certain property of the decision algorithm, averaged over the members within different socio-demographic groups, is equal across these groups. The averaging reveals the systematic effects of the decision system. In the context of this paper, we therefore define fairness as socio-demographic groups having to be equal with respect to a certain property of the ADM system.

V. UNDERSTANDING GROUP FAIRNESS MEASURES

Group fairness in the context of data-based decision making means comparing how different socio-demographic groups are treated by the ADM system. In this section, we analyze how exactly socio-demographic groups can be compared in the fairness-context. As it turns out, there are different options, and each one constitutes its own "fairness metric" (or "fairness measure"). We also note that there are similarities between some metrics which lead to the broader category of "fairness criteria" [14].

Table I provides an overview of the fairness metrics we discuss in this section. Each fairness metric that we consider is based on three elements:

- The definition of socio-demographic groups to be compared: The choice of which socio-demographic groups are compared is crucial. By, for example, comparing only women and men, unfair treatment of non-binary people cannot be detected. By comparing different gender groups and, in a separate evaluation, different racial groups, we might not notice that black women are particularly harmed by a system [22].
- The relevant subpopulation on which those groups are compared: This could, for example, be the subpopulation of people negatively affected by the decision.
- The relevant comparison property: The definition of *what* should be equal when comparing these socio-demographic groups.

We will see that the metrics differ in the second and third element, which provides an intuitive, hands-on understanding of these metrics. This understanding of the metrics helps us in taking a step towards answering the arguably biggest question in implementing group fairness metrics: Which metric should one aim to fulfill?

Our discussion of the fairness metrics will be based on the example of credit lending. We assume that a machine learning model predicts whether or not an applicant pays back their loan (\hat{Y}). The model has been trained on a training set consisting of the sensitive attribute A , which in our case is the gender of the applicant (to simplify the explanation, we assume that we only compare female to male applicants), additional attributes \vec{X} , such as their credit history and income, and the binary indicator Y of whether or not the applicant repaid their loan. A subsequent decision algorithm then takes the decision D (whether a given applicant should receive a loan) based on the prediction \hat{Y} (calculated from the features \vec{X} and A), combined with business related parameters such as current interest rates.¹ This decision model is tested with a representative pool of credit loan applicants consisting of women and men by checking for systematic differences in the decision pattern. Hereinafter, we will refer to this testing data as the "full population" and to parts of it as "subpopulations".

¹A large part of the algorithmic fairness literature is focused on cases where D , Y and A are binary. In the interest of making this paper as accessible as possible, we will follow this convention and do not discuss more complex settings.

The most obvious way to compare women and men is to look at the decision outcomes for the full population, i.e., to compare the decisions for *all men of the population* to the decisions for *all women of the population*. This leads to the fairness criterion called "independence" (see [14] and Table I), which is also known as "statistical parity" or "demographic parity" in the case of classification. It answers the question: Are women and men equally likely to get a loan?² Note that this does not require equal numbers of women and men to be accepted, but equal shares of women and men to be accepted.

Instead of considering the full population, we might also restrict the fairness analysis to subpopulations. For example, we might ask for the probability of getting a loan in the subpopulation of applicants with a stellar credit history and high income (\rightarrow "conditional statistical parity") or in the subpopulation of applicants who actually repay their loan (i.e., "good debtors" \rightarrow "true positive rate"). For a systematic overview of the possible subpopulations, we recall the basic structure of the decision problem, which is characterized by the vector of variables (A, \vec{X}, Y, D) . While the variable A defines the groups that are being compared (in our case: men and women), all other variables might be used to define subpopulations.

First, we might restrict the analysis to a subpopulation characterized by a specific range of \vec{X} . This leads to "conditional statistical parity" (see Table I).³

Another option is to restrict the fairness analysis to individuals according to their value of Y , e.g., to individuals with $Y = 1$. These individuals still differ in their value of D , i.e. in whether their loan application is approved or not, and a natural approach to check for group differences is to compare the acceptance rate of men to that of women. However, this is done only within the subpopulation of individuals actually repaying their loan – it thus represents a fairness assessment on the group of "good debtors". In the case of binary classifiers, the combination of both resulting fairness metrics (for the subpopulations of either $Y = 0$ or $Y = 1$) is called "separation" [14] (see Table I).

Finally, we might also restrict the analysis to individuals with the same decision D , e.g. to individuals with $D = 1$. In this case, assessing the average approval rate makes no sense. What can be evaluated though is the average value of Y for this subpopulation, i.e., the average repayment rate of people with $D = 1$. Equal repayment rates among accepted female and male credit applicants might be seen as a proof that the decision does not lead to systematic bias. The combination of both resulting fairness metrics (for the subpopulations of either

²Alternatively, we might ask whether women and men are equally likely to be denied their credit loan application. In the case of statistical parity, equality (inequality) in the acceptance rate implies equality (inequality) in the rejection rate. Thus, assessing the rejection rate does not provide new information about the fairness of the model – it just frames it in a different way. This is also the case with the following metrics for which we compare the credit approval rate. For better readability, we restrict the discussion to the approval rate.

³Even though conditional statistical parity is only a relaxation of statistical parity [14] and as such might not demand another detailed explanation, we add it to this list as it seems to hold intuitive value for many people. We see this, for example, in discussions about the gender pay gap [23].

positive or negative decisions) is called "sufficiency" [14] (see Table I).

The important realization here is that there is a limited number of ways to compare socio-demographic groups (here, men and women) in the case of the decision problem with the four variables A, \vec{X}, Y, D . In fact, most of the group fairness measures proposed in the literature fall into one of the three categories of either independence, separation, or sufficiency [14].⁴

VI. JUSTIFICATIONS FOR GROUP FAIRNESS MEASURES

Creating a fair decision system means enforcing one or several fairness criteria by either manipulating the training data, the prediction model or the decision rule. Note that enforcing *all* fairness metrics at the same time is a mathematical impossibility in most cases (and thus essentially always in practice) [14], [24], [25]. Therefore, we usually have to make a choice between these fairness metrics. This choice needs a justification, and this is a moral, not a technical, justification. In the following we will discuss some justifications for the different fairness criteria. Since the application context is highly relevant for which justification is deemed the most fitting one, it is not surprising that, in different contexts, different fairness metrics might be suggested. Also, people might differ in their views on which justification is valid. However, the goal of this section is to make clear which justifications are coupled with the different fairness criteria.

Note that we focus on fairness *with respect to the decision subjects*, i.e., the people directly affected by the decision. What is considered to be morally adequate for decision-makers or wider society is a different discussion. One might, for example, find that statistical parity would be fair towards the decision subjects, but would mean a loss in performance of the ADM system, which in turn would lead to an indefensible loss in profit to the decision-maker. In this case, the moral assessment of fairness metrics might differ for different stakeholders. However, we will focus on what it means to be fair towards the decision subjects. This discussion is by no means meant to be exhaustive, since the philosophical literature on this topic is still very much a work in progress.

a) Statistical parity: Several works (see, e.g., [26], [27], [28], [29]) argue that if differences in predictive features \vec{X} are due to unjust historical injustices or unjust societal structures (such as male students being more often encouraged to go into STEM fields than their equally talented female counterparts), then these differences should be corrected for by enforcing statistical parity. [29] argues that past and current injustices alone are, however, not sufficient for justifying enforcing statistical parity: It is also crucial to consider the effects of the intervention. For a case of medical treatment, they show

⁴Note that, formally, one might also assess the average Y (instead of the average D) in the case of the full population. However, this metric does not reflect a property of the decision algorithm, but rather a statistical property of the population and is typically referred to as the "base rate". In our example, it is the rate at which women and men in the dataset repay their credit loans.

that enforcing statistical parity might actually harm the already underprivileged group.

Therefore, enforcing statistical parity might make sense if (1) there are differences in \vec{X} that are caused by unjust circumstances and that would therefore, without intervention, lead to unjustifiable and unjust differences in D , and if (2) this intervention at least does not harm the already underprivileged group.

b) Conditional statistical parity: As a relaxation of statistical parity, conditional statistical parity also assumes that fairness means equality in the (average) decision for the two groups. The difference is that equality of decisions should only hold when controlling for some elements of \vec{X} , called "legitimate" attributes [30], [31]. Focusing on *conditional* statistical parity is thus motivated by the assertion that differences in certain attributes justify differences in the decision D : statistical parity should hold, but only for people with, e.g., a stellar credit history. However, the choice of legitimate attributes has to be well-justified. When we, for example, argue that differences in the credit history justify differences in treatment, we might reason that people are fully responsible for their credit history – which is an assumption that does not necessarily hold. To see this, let us consider an example provided in [26]: Assume that bank clerks are more lenient with the repayment deadlines of men compared to women. In this case, the credit history is unfit as a legitimate attribute as it is biased. In fact, if we compared the credit repayment rates of women and men with similar credit histories, we would find that women repay their loans at a higher rate than men. One may thus argue that the credit approval rate of women in this subpopulation of individuals with similar credit scores should be higher than that of men. Since conditional statistical parity demands that female and male applicants *with the same credit history* are accepted at the same rate, the metric would (in this case) disadvantage female applicants.

We conclude that conditional statistical parity is appropriate when the "equal treatment" paradigm can justifiably be applied to subpopulations defined by legitimate attributes (rather than to the full population).

c) Separation: Separation is fundamentally different from independence in that the "equal treatment" paradigm is conditioned on Y : Equal treatment should hold for individuals with the same value of Y , but not necessarily when comparing individuals with different Y . For example, one may argue that it is not necessary that a credit lending system treats women and men, on average, equally. Instead, it might be sufficient if women and men, who repay their loan, are treated equally and if women and men, who default on their loan, are also treated equally. Thus, separation explicitly accepts treatment differences based on Y and assume that they do not constitute a source of unfairness per se. This is, for example, the case when D is about granting a benefit or access to resources, such as college admissions or job promotions, where $Y = 1$ constitutes the main reason for $D = 1$, and this is morally right. Note that a necessary condition for focusing on separation is that Y is not biased with respect to A (as it would,

TABLE I
STRUCTURED OVERVIEW OF FAIRNESS METRICS

Fairness criterion	Fairness metric	Subpopulation	Property	Equation
Independence	Statistical parity	-	$P(D = 1)$	$P(D = 1 A = a) = P(D = 1 A \neq a)$
Relaxation of independence	Conditional statistical parity	$\vec{X} = x$	$P(D = 1)$	$P(D = 1 \vec{X} = x, A = a) = P(D = 1 \vec{X} = x, A \neq a)$
Separation	Parity of true positive rates	$Y = 1$	$P(D = 1)$	$P(D = 1 Y = 1, A = a) = P(D = 1 Y = 1, A \neq a)$
	Parity of false positive rates	$Y = 0$	$P(D = 1)$	$P(D = 1 Y = 0, A = a) = P(D = 1 Y = 0, A \neq a)$
Sufficiency	Parity of positive predictive values	$D = 1$	$P(Y = 1)$	$P(Y = 1 D = 1, A = a) = P(Y = 1 D = 1, A \neq a)$
	Parity of false omission rates	$D = 0$	$P(Y = 1)$	$P(Y = 1 D = 0, A = a) = P(Y = 1 D = 0, A \neq a)$

for example, be the case if the bank was more lenient with the credit repayment deadlines of men compared to women, so that Y does not actually represent timely repayments, but a biased view of it). [32] therefore argues that separation is appropriate to use if Y corresponds to what *morally* justifies a beneficial or harmful outcome. Sometimes, the moral justification is given only if the individuals can be made *responsible* for Y . However, this is not always the case. Sometimes, treating people differently based on Y is morally justified even if they are not responsible for their Y . For example, in health-care, it may be morally legitimate to decide on giving a therapy to people based on whether or not they will be cured through the therapy (where whether or not they get cured corresponds to Y), irrespective of whether they are responsible for their current health issues; in credit lending it may be considered morally legitimate to deny a loan to someone who cannot repay it, even if the individual is not morally responsible for the failure to repay, etc.

Separation is therefore justified if we believe that inequalities in treatment between groups are acceptable as long as people with the same Y are treated equally.

d) Sufficiency: Focusing on sufficiency means looking at the subpopulations defined by D and assessing them with respect to Y . The decision system is considered fair if the average propensity for $Y = 1$ is equal for both groups. For the credit lending example, we assess whether the accepted men have the same repayment rate as the accepted women and whether the rejected men have the same repayment rate as the rejected women. We first of all note that experimental evidence has shown that sufficiency is oftentimes (close to) fulfilled in unconstrained learning (i.e., whenever there is no specific intervention to achieve a certain kind of fairness) [33].

Let us now think about what assumptions would morally justify enforcing sufficiency. For this, we consider the motivation behind a relaxation of sufficiency which is often used in practice, e.g., in measuring bias in policing [34]: the "outcome test" (which aligns with testing for equality in positive predictive values (PPVs)). The test was proposed by the American economist Gary Becker [35] to measure discrimination in credit lending. As an example, assume that a bank estimates each credit loan applicant's repayment probability and gives a loan to men if they have a repayment probability above 30%, but to women only if their repayment probability is above 70%. In this case, it seems clear that women, who receive a loan, repay it at a much higher rate than men. Since in real life the thresholds above which the bank grants loans to

women and men are not known, the observation of different repayment rates might be used to infer a systematic difference in how women and men are treated. However, differences in repayment rates do not mathematically guarantee differences in thresholds because many different combinations of *individuals who apply for a loan* and *thresholds* could lead to the same results in the outcome test. This is what is known as the *infra-marginality problem*.⁵ Therefore, the intuition that different threshold have been applied when sufficiency is violated does not hold. [34] shows that this is not only a theoretical issue, but also a practical one. Since sufficiency does not measure differences in thresholds, it is still unclear why sufficiency would be morally relevant in the fairness discussion.

A possible argument for focusing on sufficiency is given in [32]. According to the generalized theory in [32], sufficiency is appropriate when a different decision D justifies different expectations for the individuals of obtaining a given (beneficial or harmful) outcome Y . This might be plausible when the decision D corresponds to the decision to give a recommendation or warning message such as "it is dangerous *for you* to rent this Ferrari". In this case, it is morally appropriate for people to have different expectations of getting hurt in driving the car depending on whether they got this warning message or not. People who receive the message, for example, rightly expect that they are more likely to get hurt than people who do not receive the message. This affects the kind of moral complaint they may have against the rental company if they still decide to drive the car.

Sufficiency therefore appears to be appropriate if we believe that the individuals about whom decisions are made justly have different expectations of their Y depending on their D .

VII. CONCLUSION

We provided an introduction to group fairness for data scientists by considering how different socio-demographic groups may be compared in the context of prediction-based decision systems. In particular, we focused on the question of what parts of the full population are pertinent for assessing the question of fairness. Starting from the formal definition of a prediction-based decision problem with the variables A, \vec{X}, Y, D , we described six fairness metrics that cover different perspectives of how to compare socio-demographic groups with respect to decision fairness. In line with existing literature, we pointed to three broad categories of group fairness under which these

⁵For more detailed information on this problem and a way to infer the thresholds from the data, see [34].

metrics fall: independence, separation and sufficiency. This approach to group fairness can help practitioners understand the choices they have if they want to enforce a group fairness metric. Since the literature on philosophical guidance is not evolved enough yet to provide concise rules for when to apply each criterion, we provided some pointers to aspects that have to be considered when choosing a fairness metric.

ACKNOWLEDGMENT

This work was supported by the National Research Programme “Digital Transformation” (NRP 77) of the Swiss National Science Foundation (SNSF), grant number 187473. We thank Michele Loi as well as our anonymous reviewers for their helpful feedback.

REFERENCES

- [1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine bias,” *ProPublica*, 2016. [Online]. Available: <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
- [3] C. Orwat, “Diskriminierungsrisiken durch verwendung von algorithmen,” *Eine Studie, erstellt mit einer Zuwendung der Antidiskriminierungsstelle des Bundes*, 2019.
- [4] A. Narayanan, “Translation tutorial: 21 fairness definitions and their politics,” in *Conference on Fairness, Accountability and Transparency*, 2018.
- [5] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, “AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias,” Oct. 2018. [Online]. Available: <https://arxiv.org/abs/1810.01943>
- [6] S. Verma and J. Rubin, “Fairness definitions explained,” in *2018 IEEE/ACM International Workshop on Software Fairness (Fairware)*. IEEE, 2018, pp. 1–7.
- [7] S. Mitchell, E. Potash, S. Barocas, A. D’Amour, and K. Lum, “Algorithmic fairness: Choices, assumptions, and definitions,” *Annual Review of Statistics and Its Application*, vol. 8, pp. 12.1–12.23, 2021.
- [8] J. S. Kim, J. Chen, and A. Talwalkar, “FACT: A diagnostic for group fairness trade-offs,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 5264–5274.
- [9] M. Miron, S. Tolan, E. Gómez, and C. Castillo, “Addressing multiple metrics of group fairness in data-driven decision making,” *arXiv preprint arXiv:2003.04794*, 2020.
- [10] D. Hellman, “Measuring algorithmic fairness,” *Va. L. Rev.*, vol. 106, p. 811, 2020.
- [11] K. Makhlof, S. Zhioua, and C. Palamidessi, “On the applicability of ml fairness notions,” *arXiv preprint arXiv:2006.16745*, 2020.
- [12] J. Nyarko, S. Goel, and R. Sommers, “Breaking taboos in fair machine learning: An experimental study,” 2020, Working Paper.
- [13] S. Goel, J. Nyarko, and R. Sommers, “Predictions – whether algorithmic or human – may not be fair,” *The Boston Globe*, 2020. [Online]. Available: <https://www.bostonglobe.com/2020/11/02/opinion/predictions-whether-algorithmic-or-human-may-not-be-fair/>
- [14] S. Barocas, M. Hardt, and A. Narayanan, “Fairness and machine learning,” 2020, Incomplete Working Draft. [Online]. Available: <http://fairmlbook.org>
- [15] R. Binns, “Fairness in machine learning: Lessons from political philosophy,” in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, ser. Proceedings of Machine Learning Research, S. A. Friedler and C. Wilson, Eds., vol. 81. New York, NY, USA: PMLR, 23–24 Feb 2018, pp. 149–159. [Online]. Available: <http://proceedings.mlr.press/v81/binns18a.html>
- [16] R. Dworkin, “What is Equality? Part 2: Equality of Resources,” *Philosophy and Public Affairs*, vol. 10, no. 4, pp. 283–345, 1981.
- [17] G. A. Cohen, “On the Currency of Egalitarian Justice,” *Ethics*, vol. 99, no. 4, pp. 906–944, 1989.
- [18] J. Rawls, *Justice as fairness: A restatement*. Harvard University Press, 2001.
- [19] A. Z. Jacobs and H. Wallach, “Measurement and fairness,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 375–385.
- [20] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, “Fairness through awareness,” in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.
- [21] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, “Counterfactual fairness,” *arXiv preprint arXiv:1703.06856*, 2017.
- [22] J. Buolamwini and T. Gebru, “Gender shades: Intersectional accuracy disparities in commercial gender classification,” in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.
- [23] D. Grimshaw and J. Rubery, “The adjusted gender pay gap: a critical appraisal of standard decomposition techniques,” *Manchester School of Management, UMIST, Manchester, UK*, 2002.
- [24] J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent trade-offs in the fair determination of risk scores,” *arXiv preprint arXiv:1609.05807*, 2016.
- [25] A. Chouldechova, “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” *Big data*, vol. 5, no. 2, pp. 153–163, 2017.
- [26] R. Binns, “On the apparent conflict between individual and group fairness,” in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 514–524.
- [27] T. Ráz, “Group fairness: Independence revisited,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 129–137. [Online]. Available: <https://doi.org/10.1145/3442188.3445876>
- [28] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, “On the (im)possibility of fairness,” *arXiv preprint arXiv:1609.07236*, 2016.
- [29] C. Hertweck, C. Heitz, and M. Loi, “On the moral justification of statistical parity,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT ’21. New York, NY, USA: Association for Computing Machinery, 2021, p. 747–757. [Online]. Available: <https://doi.org/10.1145/3442188.3445936>
- [30] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, “Algorithmic decision making and the cost of fairness,” in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.
- [31] I. Žliobaite, F. Kamiran, and T. Calders, “Handling conditional discrimination,” in *2011 IEEE 11th International Conference on Data Mining*. IEEE, 2011, pp. 992–1001.
- [32] M. Loi, A. Herlitz, and H. Heidari, “Fair equality of chances,” in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES ’21. New York, NY, USA: Association for Computing Machinery, 2021, forthcoming. Available at SSRN: <https://ssrn.com/abstract=3450300>.
- [33] L. T. Liu, M. Simchowitz, and M. Hardt, “The implicit fairness criterion of unconstrained learning,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 4051–4060.
- [34] C. Simoiu, S. Corbett-Davies, S. Goel *et al.*, “The problem of infra-marginality in outcome tests for discrimination,” *The Annals of Applied Statistics*, vol. 11, no. 3, pp. 1193–1216, 2017.
- [35] G. S. Becker, *The economics of discrimination*. University of Chicago press, 2010.