Proceedings of Machine Learning Research 279:68–95, 2024 Algorithmic Fairness Through the Lens of Metrics and Evaluation

The Intersectionality Problem for Algorithmic Fairness

Johannes Himmelreich JRHIMMEL@SYR.EDU
Maxwell School of Citizenship and Public Affairs
Syracuse University
900 S Crouse Ave
Syracuse, NY 13244, USA

Arbie Hsu WHSU10@DONS.USFCA.EDU
Ellen Veomett EVEOMETT@USFCA.EDU
University of San Francisco
2130 Fulton St
San Francisco, CA 94117, USA

Kristian Lum KRISTIANL@UCHICAGO.EDU
University of Chicago
5801 S Ellis Ave
Chicago, IL 60637, USA

Editors: Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, Golnoosh Farnadi

Abstract

A yet unmet challenge in algorithmic fairness is the problem of intersectionality, that is, achieving fairness across the intersection of multiple groups—and *verifying* that such fairness has been attained. Because intersectional groups tend to be small, verifying whether a model is fair raises statistical as well as moral-methodological challenges. This paper (1) elucidates the problem of intersectionality in algorithmic fairness, (2) develops desiderata to clarify the challenges underlying the problem and guide the search for potential solutions, (3) illustrates the desiderata and potential solutions by sketching a proposal using simple hypothesis testing, and (4) evaluates, partly empirically, this proposal against the proposed desiderata.

1. Introduction

That intersectionality matters is a point of consensus in the algorithmic fairness literature. A model's performance might be much worse for women of color than for women and people of color considered separately (Buolamwini and Gebru, 2018). In this paper, we elucidate a problem that intersectionality raises for algorithmic fairness in practice: Because data on intersectional groups is often severely limited, *verifying* that algorithmic fairness—under various definitions thereof—has been attained is difficult. Although this problem is recognized in the literature (Kearns et al., 2018; Foulds et al., 2020a; Morina et al., 2020; Molina and Loiseau, 2022), its challenges do not appear to be fully appreciated and many existing contributions violate minimal moral or methodological desiderata.

Our contribution is fourfold: We (1) elucidate the problem of intersectionality in algorithmic fairness, and (2) develop desiderata to clarify the challenges that underlie this problem of intersectionality and to guide the search for potential solutions. Moreover, we (3)

© 2024 J. Himmelreich, A. Hsu, E. Veomett & K. Lum.

illustrate the desiderata and potential solutions by presenting a statistical setup that uses simple hypothesis testing, and (4) evaluate this proposal, partly empirically, in light of the desiderata.

Our larger aim is to advance the literature on algorithmic fairness more broadly. The approach we propose in response to the problem of intersectionality differs fundamentally from the typical way of "measuring" algorithmic fairness.[1] We hence advance the debate by pointing out possibilities of approaching fairness differently: as accounting for uncertainty (instead of concentrating on point estimates) and as a matter of sufficiency (instead of equality).

2. Preliminaries

2.1. Algorithmic Fairness

In the literature on algorithmic fairness, "fairness" is typically defined as model performance (such as accuracy or false positive rate) that is roughly equal across all relevant groups. Many versions of algorithmic fairness consider fairness to have been achieved if

$$|m(G) - m(\cdot)| < \epsilon \quad \text{for some small } \epsilon, \ \forall G \tag{1}$$

Where G denotes a subgroup of the population, $m(G)$ a model's performance (however understood) on only the subset of the data that belongs to group G, and $m(\cdot)$ the model's performance calculated across the entire dataset, irrespective of group membership. Membership in G typically corresponds to a sensitive or protected attribute such as race, sex, age, disability or marital status but G may also be defined intersectionally as a *combination* of such attributes.

Equation (1) generalizes a large family of definitions or—when aggregating $|m(G) - m(\cdot)|$ for all groups—metrics of fairness. We thus take (1) to represent the *typical* way of understanding algorithmic fairness. This typical way of understanding fairness faces the problem of intersectionality.

2.2. The Problem of Intersectionality

As the number of attributes that define subgroups grows, the amount of data available for each subgroup shrinks rapidly. After all, the number of subgroups grows *exponentially* with the number of protected attributes: For n binary attributes, there are 2^n intersectional groups. This, in turn, entails a data problem: When social identities are constituted by intersections of increasingly many attributes, and when these constituting attributes are not just binary, the data within each of the intersections can become very small. In Europe, where discrimination is highly intersectional and fairness audits are encouraged by legislation,[2]

1. We use fairness "measure," "metric" and their cognates with two caveats. First, the problem is one of estimation, not measurement. Second, fairness metrics are *meta-metrics* since they aggregate a higher-dimensional vector of model performance into a lower-dimensional summary (Lum et al., 2022).
2. Recital 49 of the EU Artificial Intelligence Act (2021a) encourages "the development of benchmarks and measurement methodologies for AI systems" (2021b). Yet the statistical problems of intersectional fairness are, in some way, greater in Europe. Since nationality groups are already comparatively small, intersectional groups are even smaller subgroups within already small nationality groups. For example,

fairness audits may need to account for several thousand subgroups.[3] Because gathering the data necessary for fairness audits is typically costly—e.g., the "ground truth" needs to be established to assess whether a prediction is correct—such data tend to be scarce.

In short, the intersectionality problem of algorithmic fairness is a problem of statistical uncertainty due to small data and, thus, raises problems for how "fairness" is typically defined.

Intersectionality renders fairness metrics, as they are typically defined, meaningless. These metrics, such as (1), rely on point estimates of model performance (e.g., whether this performance is roughly the same for all groups). But point estimates become nonsensical with small data (Kearns et al., 2018).[4] The challenge posed by intersectionality for algorithmic fairness is to define a fairness metric that provides meaningful estimates of fairness even when groups are very small and audit data are sparse.

Our discussion hence adds to the existing technical and critical objections against (intersectional) algorithmic fairness (Corbett-Davies et al., 2024; Kong, 2022), acknowledging that a commitment to intersectionality and fairness likely requires a broader set of actions than estimating certain properties of models (Stewart, 2022; Wang et al., 2022; Suresh et al., 2022; Klumbytė et al., 2022).

3. Existing Work

Various statistical methods have been proposed for intersectionality in algorithmic fairness.

3.1. Kearns et al.

An early identification and statement of the problem of intersectional fairness arising from small groups is due to Kearns et al. (2018). The approach of Kearns et al. involves an audit algorithm that learns to classify models as fair or unfair instead of defining a fairness metric. The process of learning this audit algorithm is subject to a fairness constraint that is weighted depending on the proportion of the population belonging to a particular group G.

Kearns et al. define $\alpha(G) = Pr(G)$ and reformulate fairness in (1) as

$$\alpha(G)|m(G) - m(\cdot)| < \epsilon \quad \forall\, G \tag{2}$$

Essentially, the addition of $\alpha(G)$ relaxes the original fairness metric of (1) depending on the proportion of G as a share of the overall population. The smaller G is, the more the condition is relaxed. As Kearns et al. explain, this addition is necessary to enable statistical estimation, given the increasing statistical uncertainty with decreasing group size.

Hungarian Roma face discrimination in the housing market, Maghrebi French in the labor market, whereas people of African descent in England and Wales face discrimination in the criminal justice system (Center for Intersectional Justice, 2020).

3. Assuming 3 binary attributes (e.g., non-white, cis-gender, same-sex orientation) 1 three-valued attribute (e.g., gender as 'male,' 'female,' and 'neither'), 9 different ethnic backgrounds (e.g., Roma, Chinese, Turkish), and 12 different nationalities or localities (e.g., Hungarian, German, French), yields 2,592 intersectional subgroups. And this number may be conservative since the number of discernible ethnic groups is larger than 9, of nationalities is larger than 12, and the legally protected attribute of age is not even included.
4. For example, in binary classification, an individual prediction is either 1 or 0; and the model accuracy for each singleton group is thus either 1 or 0.

We discuss the implications in Section 4, and give the results of an empirical study regarding this formulation in Appendix E.

3.2. Foulds et al. and Morina et al.

Foulds et al. (2020a) provide an alternative approach based on ratios of model performance metrics. An expanded version of which is, in turn, given by Morina et al. (2020).[5]

These definitions require that the ratio of some metric value between two groups be within a fixed interval. For example, suppose $m(G)$ measures the true positive rate (TPR) for subgroup G. Then the ϵ-differential intersectional definition of TPR parity (equal opportunity), given by Morina et al. (2020), is that

$$e^{-\epsilon} \leq \frac{m(G)}{m(G')} \leq e^{\epsilon} \quad \forall\, G, G' \tag{3}$$

Morina et al. (2020) note that $\epsilon = 0$ corresponds to "perfect fairness" ($m(G) = m(G')$).

3.3. Molina and Loiseau

Molina and Loiseau (2022) use a statistical approach to addressing intersectionality and fairness. They call a classifier (ϵ, δ)-probably intersectionally fair if "the expected number of people that faces a discrimination more than ϵ is less than $n\delta$" (n is the population size).[6]

3.4. Cherian and Candès

Cherian and Candés (2023) address fairness auditing for many subpopulations within the framework of hypothesis testing, as we do here. They use a bootstrap process to provide statistical performance bounds for many subpopulations at once. Our addition to this study is the illumination and discussion of desiderata (in Section 4), a clear description of how one can design fairness metrics using hypothesis testing (Section 5), and an empirical study showing that these metrics encourage (rather than discourage) the gathering of additional data to improve model performance (Section 6).

3.5. Khan et al., Agrawal et al., Herlihy et al.

Khan et al. (2023) consider metrics of fairness, accuracy, and variance for model estimators. They empirically show that there tends to be a tradeoff between these three values. In a similar vein, Agrawal et al. (2021) study debiasing methods, and in doing so show both theoretically and empirically that estimation variance tends to be higher in small subgroups.

5. We note that Foulds et al. (2020b) (the same group as in Foulds et al. (2020a)) also study the usage of Bayesian modeling to more accurately measure fairness metrics than point estimates. Although these Bayesian models for measuring fairness metrics may give more accurate estimates than point estimates, they do not allow for the same kind of statistical analysis and ethical evaluation as a confidence interval (which we propose in Sections 4 and 5).
6. Molina and Loiseau moreover highlight the issue of estimating fairness of a model on subgroups for whom the set of predicted values on that subgroup is a proper subset of the set of all predicted values. This becomes an issue because they use a ratio similar to that in Equation (3), which is undefined if $m(g') = 0$ for some group g'. Our models do not suffer from this issue; however, extremely tiny subgroups do come with their own statistical uncertainty issues, as we highlight in Section 5.

Additionally, they prove results suggesting that partial debiasing results in both less variance and better fairness properties. Herlihy et al. (2024) use a structured regression approach in an effort to optimize the bias-variance trade-off.

4. Desiderata

Although the problem of intersectionality is recognized in the literature, how difficult this problem is may not have been fully appreciated. At least some of the existing contributions violate minimal moral or methodological desiderata, as we shall see in Sections 4.1, 4.2, and 4.3 (and Appendix E).

A core tenet of building ethical algorithms is that machine-learned models need to be consistent with "human values," which can be formulated as desiderata. We see the following desiderata for intersectional fairness metrics.

4.1. Minimal Justice

A first desideratum we call "minimal justice." The idea is, roughly, that a standard of fairness should not be lower for certain groups, such as those historically targeted for discrimination or facing structural injustice. Intuitively, minimal justice is a form of minority protection that says "don't disadvantage the disadvantaged."

This desideratum is a weak form of prioritarianism. Recent work in algorithmic fairness has identified a similar prioritarian idea in "predictive justice" (Lazar and Stone, 2024). Whereas prioritarianism, a theory of distributive justice for well-being, demands that "benefitting people matters more the worse off these people are" (Parfit, 1997), minimal justice requires only that those "worse off" should be given *at least the same* weight in aggregating a fairness metric. The desideratum does *not* require that greater weight be given to any group, and is hence met when a standard of fairness is identical for all groups.

To illustrate the desideratum, consider an example. Notwithstanding its merits, the proposal of Kearns et al. (2018) may violate minimal justice. As noted above, the addition of $\alpha(G)$ in (2) relaxes the fairness constraint proportional to the size of a group. The smaller a group is (as a share of the data), the worse a model performance can be and still certify the model as fair. The fairness standard is hence lowered for small groups. On the assumption that these small groups include historically disadvantaged or oppressed groups, (2) violates minimal justice.

And drastically so: For a group G' that is c times smaller than group G (i.e. $\frac{\alpha(G)}{\alpha(G')} = c$), a model can be certified as "fair" if the disparity between the average performance and the performance for group G' is as much as c-times worse than it is for group G. Furthermore, for some value of ϵ there are groups that are proportionally so small that there is no model performance poor enough to certify the model as unfair. For example, if $\epsilon = .01$, for a binary classifier, any group whose proportion of the total population is less than ϵ is protected by essentially no fairness constraint at all.[7]

The ethical impact can be immense. A group might *look* relatively small in the data but be, in fact, large in absolute numbers in the population. Indeed, disadvantaged groups

7. The average model performance $m(G)$ and $m(\cdot)$ is constrained to be less or equal to 1. But the maximum deviation of accuracy in the binary setting is 1. Thus, even if the model is entirely inaccurate for this population and perfectly accurate for the rest of the population, the constraint is still satisfied.

tend to be under-represented in data (Lerman, 2013; Giest and Samuels, 2020). Thus, the approach of Kearns et al. may lower the standard of fairness for precisely those groups that fairness is meant to protect.

4.2. Consistent Conceptualization

Any fairness metric operationalizes a certain idea, or concept, of fairness. A second desideratum is that fairness metrics should operationalize a concept of fairness consistently.

This desideratum may resemble that of Minimal Justice. But whereas Minimal Justice is a moral desideratum, Consistent Conceptualization is a methodological one. Minimal Justice is a requirement on the *content* of a standard of fairness, on how a standard of fairness treats certain groups. Consistent Conceptualization, by contrast, requires a form of coherence between the informal "intuition" and the formal explication of a standard of fairness, which is known as minimal "construct validity." The importance of construct validity for fairness is already established in the literature (Jacobs and Wallach, 2021).

Typically, fairness metrics in algorithmic fairness operationalize the idea of *equality*. This is particularly evident in (1) which, for each group G, restricts the absolute disparity of $m(G)$ from overall mean performance $m(\cdot)$. This is one—albeit a very simple—way of operationalizing inequality (for alternatives see Sen (1997)). Likewise, (3) operationalizes fairness as equality (Foulds et al., 2020a; Morina et al., 2020).[8]

Moreover, (1) and the definition by Foulds et al. and Morina et al. operationalize equality *consistently*. The fairness metrics apply an equality condition without bounds or exceptions.

Not so the proposal by Molina and Loiseau (2022), which explicitly *bounds* equality. Effectively, the fairness measure permits that some small number of people faces severe discrimination, as long as the likelihood of discrimination or their relative size as a share of the overall population is small.[9] This fairness metric thus fails the desideratum of operationalizing the concept of equality consistently.

Typically fairness metrics, and all instances of (1), operationalize fairness as equality. Alternatives, well-known from distributive justice, include *prioritarianism*, stating that more of some good, such as model performance, should be given to those in greater need (Parfit, 1997), and *sufficientarianism*, requiring that everyone has *enough* of some good (instead of the same) (Frankfurt, 1987; Slote, 1989).

4.3. Incentive Compatibility

The final desideratum starts with the recognition that metrics specify incentives. Anyone who wants to increase their models' fairness may want to maximize a fairness metric. The final desideratum thus requires that a fairness metric not have "perverse" incentives of two kinds: discouraging data collection and allowing "gaming."

First, a fairness metric should not discourage data collection. Any fairness metric that indicates greater *un*fairness only because further data are sampled from some group would fail

8. Compared to (1), (3) aims for equality between groups (as opposed to minimizing disparity with $m(\cdot)$), and measures *relative* disparity (a performance ratio instead of performance difference).
9. Molina and Loiseau (2022) write: "It can be seen for some given ϵ as a statement on the expected size of the population that is not being discriminated too much against."

to be incentive compatible. Likewise, inversely, any fairness metric would fail the desideratum that indicates greater *fair*ness only because data based on group identity are dropped.

The fairness metric (2), of Kearns et al., likely violates this desideratum of incentive compatibility. This is because collecting more data on a minority population G tightens the constraint by increasing $\alpha(G)$, thus making a certification of "fairness" at a given level of ϵ more difficult. Specifically, suppose that $m(G) = .15$ and $m(\cdot) = .85$. If $\alpha(G) = .01$, then the performance would be deemed "fair" for all $\epsilon > 0.7 \times 0.01 = .007$. However, if we collect more data for group G such that $\alpha(G) = .2$, then the model would be "fair" only for $\epsilon > 0.7 \times 0.2 = .014$. Unless the additional data results in material improvements to $m(G)$, for any ϵ such that $.007 < \epsilon < 0.014$, the fairness metric (2) would certify a given model as fair prior to further data collection, but as unfair afterwards. In short, under (2), fairness for hard-to-predict groups could be attained simply by under-representing them in the training data. We see this effect in our empirical study, described in Appendix E. We leave the details of this study to Appendix E, but the results show that the fairness metric suggested by Kearns et al. appears to indeed disincentivize additional data collection, violating *Incentive Compatibility*.

This is a "perverse" effect because, in practice, additional data collection about a minority group will help improve the model performance for that group. In other words, the metric gives an incentive to do the opposite of what it is meant to achieve.[10]

Whether other metrics (such as those by Morina et al. (2020); Foulds et al. (2020a); Molina and Loiseau (2022)) violate this desideratum depends on whether the estimated performance disparity is greater than the true disparity (which further data would likely help approximate). Fairness metrics that operationalize fairness as *equality* (e.g., as model performance disparity across groups), incentivize $m(G)$ to be nearly the same for all subgroups G. If the true model performance is nearly equal among groups, then these metrics incentive further data collection in order to have more accurate estimates of $m(G)$.

Second, a fairness metric should not encourage knowingly erroneous predictions. But some metrics (e.g., statistical or demographic parity) have exactly this property: Even if the label that we want to predict is known (which it generally, of course, isn't), "fairness" as these metrics define it can be improved by erroneous predictions. This is an undesirable property of fairness metrics (Dwork et al., 2012).

5. Two Alternative Metrics

We now illustrate how these desiderata can be met. We propose two alternative models, which we call the "optimist's" and "pessimist's model" respectively. Both define the problem using hypothesis testing. The optimist has the null hypothesis that the model is fair, and we have to prove it is not (similar to "innocent until proven guilty"); the pessimist inverts the "burden of proof" and has the null hypothesis that the model is unfair.[11]

10. We do *not* contend that more data should be collected. Privacy considerations are important. Our point is instead that maintaining the appearance of a good fairness metric is a bad reason to not collect more data.

11. Throughout, we assume that for the metric $m(\cdot)$ larger values are better (think accuracy, not error rates). Specifically, and without loss of generality, we use accuracy as our sample metric. This choice is for simplicity only; one could replace accuracy with any other metric for which higher values are preferred. For the hypothesis tests we describe, we use a z-score of 1.64, which corresponds to a 95% confidence

5.1. Optimist's Model

We could formulate the problem of fairness for small groups as testing the joint hypothesis that

$$H_0 : m(G) > c \quad \forall\, G$$
$$H_1 : m(G) \leq c \quad \exists\, G$$

Consider a group G of size n_G. Suppose $m(G)$ is accuracy. As a sample proportion, the standard error for our estimate of $m(G)$ is $\sqrt{\frac{m(G)(1-m(G))}{n_G}}$. Then, we would reject the null if the upper end of its confidence interval is less than c, i.e., if $m(G) + 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}} < c$ (ignoring multiple testing).[12] Under this formulation, we reject H_0 if $m(G)$ is sufficiently less than c, where "sufficiently less" has to do with our statistical power to detect that it is less. We would declare the model fair, if at given level c we cannot statistically reject that the model performs at least c well for all groups.

A minority population which is sufficient in number would easily reject the null if $m(G)$ is truly below c. Indeed, even with a population size of $n_G = 1000$, if $c = 0.7$, then a value of $m(G) < 0.67$ would reject the hypothesis that the model is fair.

5.2. Pessimist's Model

Depending on a model's deployment context, the optimistic approach might be problematic.[13] Consider instead the following pessimistic hypothesis test.

$$H_0 : m(G) < c \quad \exists\, G$$
$$H_1 : m(G) \geq c \quad \forall\, G$$

We would declare the model fair, if at a given level c we know with statistical certainty that the model performs at least c-well for all groups. In this case, (ignoring multiple testing again) we would require that $m(G) - 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}} > c$ for all G.

5.3. Fairness Metrics

The formulations can be extended from a hypothesis test to a fairness metric by finding the maximal c for which the respective null hypothesis cannot be rejected (for the optimist) or can be rejected (for the pessimist). In the optimist's model, choose the maximal c such that

$$c \leq m(G) + 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}} \tag{4}$$

for all relevant groups G. The fairness metric is the maximal c such that we cannot reject the hypothesis that the model performs at least c-well for all groups.

interval for a one-sided hypothesis test. This is a conventional parameter choice and nothing in our argument depends on it.

12. This ignores multiple hypothesis testing, which we address in Appendix B.
13. Depending on the ethical risks involved in how a model is used, the more precautionary assumptions behind the pessimist's model might be more appropriate.

This metric can be read as saying that a model is "fair up to c." Intuitively, this means that, for all we know, the model performance $m(G)$ (say, accuracy) is likely as high as c for each group.

On the pessimist's model, we instead choose the maximal c such that

$$c \leq m(G) - 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}} \tag{5}$$

for all relevant groups G. This fairness metric is the maximal c such that we reject the hypothesis that the model is *un*fair, that is, we reject that it does not perform at least c-well for each group.

This metric can be read as saying that a model is "unfair above c." The model likely performs at least c-well for each group; but for values above c, there likely is at least one group for which the model does not perform at least c-well—and we hence can't rule out that the model is unfair.

In summary, the fairness metrics are defined as bounds of the interval

$$\left(m(G) - 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}}, m(G) + 1.64\sqrt{\frac{m(G)(1-m(G))}{n_G}}\right)$$

This interval, of course, now has two interpretations. For one, it is the 90% confidence interval for the value of $m(G)$ for each G. Moreover, across all groups, it is also the interval in which we cannot reject the hypothesis that the model is unfair, nor can we reject the hypothesis that the model is fair.[14]

5.4. Discussion: Desiderata

Both metrics satisfy *Minimal Justice*. The bound c encodes a standard of fairness that is identical for all groups. Moreover, the relative size of groups doesn't matter. Whether a null hypothesis can be rejected changes with the absolute size of the group n_G (rather than the proportion $\frac{n_G}{n}$).

On the optimist's metric, for a small group, the difference between the actual (lower) model performance and the level up to which a model can be certified as fair might be large. But both of our metrics base their certification of "fairness up to c" on an aggregation that gives all groups the same weight. In fact, the pessimist's metric can be called "epistemically risk averse" insofar as it picks the *highest lower* bound out of all groups' confidence intervals (and hence is similar to the maximin decision rule).

On *Consistent Conceptualiztion*, both of our metrics conceptualize fairness as sufficiency. They understand fairness not as a matter of whether everyone has the same (as equality does), but whether everyone has *enough* (Frankfurt, 1987; Slote, 1989). This idea is operationalized in (4) and (5) in a transparent and natural way: with an inequality. Moreover, the threshold c, what counts as "enough," is determined absolutely in the terms of model performance measure, and not depending on, e.g., how well the model performs on other groups. Thus, both of our metrics operationalize sufficiency consistently across all groups.

14. The reader may go to Appendix A for an exploration on the impacts of changing m and n on these models.

For *Incentive Compatibility* the picture is mixed: Both of our metrics discourage gaming (and thus satisfy Incentive Compatibility in this respect). This is because both fairness metrics determine (un)fairness as the highest (or lowest) expectable model performance across all groups. As such, improving model performance will never increase unfairness; and decreasing model performance will never increase fairness. In fact, decreasing model performance may lead to a decrease in fairness. It appears that operationalizing the idea of fairness as sufficiency is what makes our fairness metrics less susceptible to gaming—in particular, that the minimum level of model performance is defined in absolute terms and equally enforced for all groups.

But one of our metrics, namely the optimist's, may discourage further data collection (and thus violate Incentive Compatibility in its first respect). Because the optimist's model starts with the null hypothesis that a model is fair at a given c, gathering more data can make things "worse"; that is, with more data, we might come to reject the optimistic null hypothesis of fairness at a given c. A model might perform very poorly for certain groups, but we cannot reject the null hypothesis that the model is fair up to c, thanks to sparse data—and the metric thus results in an incentive to not sample more data but to instead "look the other way."

6. Fairness Datasets Analysis

We evaluate empirically whether our metrics meet the desideratum of incentive compatibility. The question is: Do our metrics incentivize or disincentivize additional data collection?

To answer this question, we "simulate" additional data collection by experiment. We train models on increasingly larger subsamples of benchmark datasets and observe how metrics behave as the size of the training data increases. The behavior that we want to see is that the fairness metrics increase with the size of the training data sampled from the dataset. If, instead, a fairness metric *decreased* as greater shares of the dataset are sampled, the metric would *disincentivize* further data collection.

We seek to observe our metrics' behavior across the largest feasible range of benchmarks. To achieve this, we use lale, a Python library created by IBM (Baudart et al., 2020). Lale allows for the creation of consistent automated machine learning models across 20 well-known "fairness datasets" that can easily be fetched, modeled, and evaluated (Hirzel and Feffer, 2023). These datasets are all tabular with a categorical target variable. They each come with "fairness metadata," which includes protected attributes, along with ranges/values of those attributes that correspond to the privileged group.[15] Details on the methods of our analysis are in Appendix C. Here we only discuss the main result on testing whether our metrics incentivize against data collection.

For each of the datasets, we observe model performance $m(G)$, as well the optimist's c_1^g and the pessimist's fairness metric c_2^g respectively. For ease of interpretation we use accuracy as model performance; neither our results nor their interpretation depend on this.

We ran two versions of this experiment. In one version, we subsample the entire dataset of each benchmark; whereas in another, we subsample only on the *critical subgroup*, which

15. While there have been critiques of the usage of some of these datasets (Ding et al., 2021; Bao et al., 2021), they are still appropriate for the purpose of testing whether our proposals incentivize or disincentivize the collection of additional data.

is the group that is right on the c threshold. The first version simulates additional data collection for *all* groups, whereas the latter for those groups that "drag down" the fairness metric. Here we concentrate on results from subsampling on the critical subgroup only, shown in Figure 1.[16] Full results for both versions are in Appendix C.4.

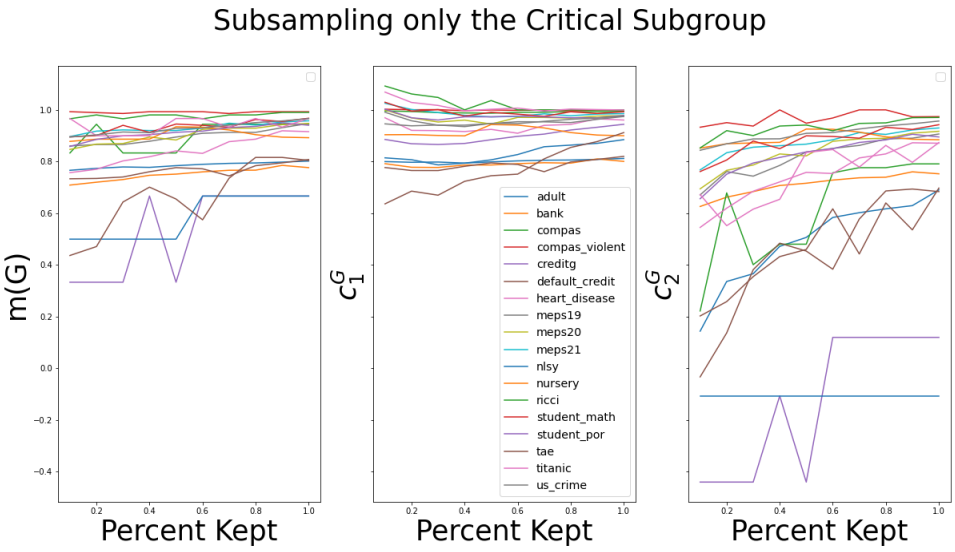

Figure 1: Plots of accuracy $m(G)$, optimist's metric c_1^g, and pessimist's metric c_2^g of critical subgroups G for each dataset. The x-axis corresponds to the percentage of the critical subgroup that is kept. Legend lists the dataset name.

The thing to note here is that there is a trend upwards in each of these plots. Most notably, the middle plot, on the optimist's metric c_1^g shows this upward trend.[17] This suggests that our optimist's metric—at least for the datasets tested—does *not* pose perverse incentives. The further we go on the x-axis (representing more data being "collected"), the model performance as well as the fairness metrics tend to improve.

Consider for example the behavior of the optimist's metric for the model trained on increasing amounts of data from the `tae` dataset (brown line that "starts" lowest in middle figure). Although the metric does not strictly increase as the training is based on greater data (the metric decreases slightly from 20% to 30% of data used), it shows a very strong upward trend.

16. Some of the plots are disconnected. This is because sometimes the subsampling of the dataset did not include any members of the critical subgroup; in those cases, the model could not predict for that subgroup, so no accuracy measurement could be taken. The most erratic curves (curves of $m(G)$ and c_2 for the `creditg` and `nlsy` datasets) correspond to either a subgroup of size 1 or 2.
17. Some datapoints "overshoot" on the y-axis with values > 1, suggesting a negative trend, e.g., for the `compas` dataset (green line). But this behavior is an artifact of the standard way of calculating the confidence interval.

7. Conclusion

Although the general idea of intersectionality seems easy to state, putting intersectionality to work in quantitative social science is, generally, far from straight-forward (Bright et al., 2016). Likewise, intersectionality presents a problem for algorithmic fairness: Intersectionality requires estimating statistical properties across subgroups that are increasingly small, which gives rise to statistical as well as moral-methodological challenges.

Statistically, small groups are a challenge for estimation. As statistical uncertainty increases (due to more and smaller groups), the point estimates of model performance for these groups become meaningless. Any approach of intersectional fairness needs to account for statistical uncertainty. But some existing metrics do not seem to fully appreciate the moral-methodological challenges that underlie this problem and "lower the fairness bar" for smaller groups, i.e., the metrics violate desiderata such as Minimal Justice or Consistent Conceptualization.

With this paper, we elucidate this intersectionality problem for algorithmic fairness: We develop minimal desiderata to clarify the moral-methodological challenges underlying this problem; we argue that some existing fairness metrics fail these desiderata, but illustrate that the desiderata can be met. We propose fairness metrics that rely on hypothesis testing (instead of performance point estimates) and that understand fairness as sufficiency (instead of equality). On these proposed metrics, fairness is understood as a certain minimum level of expected model performance that is, for all we know, likely enjoyed by all groups. We empirically evaluate the metrics against the proposed desiderata, including on 18 datasets that are widely used for fairness benchmarks.

In light of their technical and normative-theoretical limitations, the metrics we propose should be seen as illustrations. Technically, the simple hypothesis testing needs to be extended to multiple hypothesis testing to allow for interdependent subgroup memberships (see Appendix B). Normative-theoretically, the desiderata that we develop are not exhaustive and they do not uniquely characterize the metrics we propose.

Nevertheless, overall, our findings extend the list of problems that statistical uncertainty raises for algorithmic fairness. Previous work observed that fairness metrics are biased: They "fail to account for statistical uncertainty . . . exaggerating the extent of performance disparities" between groups where such disparities exist and indicating disparities "in cases where model performance is . . . identical across groups" (Lum et al., 2022). Our present findings add that with increasing statistical uncertainty fairness metrics risk becoming either nonsensical (if they aggregate point estimates) or morally inadequate (if they "lower the fairness bar" to enable statistical estimation).

However, we also offer ways of advancing the literature on algorithmic fairness: with desiderata that clarify the challenges at hand and guide the search for solutions, and with fairness metrics that suggest novel avenues for defining such metrics based on hypothesis testing and fairness as sufficiency.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. DMS-1928930 and by the Alfred P. Sloan Foundation under grant G-2021-16778,

while Ellen Veomett was in residence at the Simons Laufer Mathematical Sciences Institute (formerly MSRI) in Berkeley, California, during the Fall 2023 semester.

References

Ashrya Agrawal, Florian Pfisterer, Bernd Bischl, Francois Buet-Golfouse, Srijan Sood, Jiahao Chen, Sameena Shah, and Sebastian Vollmer. Debiasing classifiers: is reality at variance with expectation?, 2021. URL https://arxiv.org/abs/2011.02407.

Michell Bao, Angela Zhou, Samantha Zottola, Brian Brubach, Sarah Desmarais, Aaron Horowitz, Kristian Lum, and Suresh Venkatasubramanian. It's compaslicated: The messy relationship between rai datasets and algorithmic fairness benchmarks. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/92cc227532d17e56e07902b254dfad10-Paper-round1.pdf.

G. Baudart, M. Hirzel, K. Kate, P. Ram, and A. Shinnar. Lale: Consistent automated machine learning. In *AutoML Workshop at KDD*, 2020.

Liam Kofi Bright, Daniel Malinsky, and Morgan Thompson. Causally Interpreting Intersectionality Theory. *Philosophy of Science*, 83(1):60–81, January 2016. ISSN 0031-8248, 1539-767X. doi: 10.1086/684173. URL https://www.cambridge.org/core/journals/philosophy-of-science/article/causally-interpreting-intersectionality-theory/E78BB6C33D0D7DF4316FCD3687912258. Publisher: Cambridge University Press.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

Center for Intersectional Justice. Intersectionality at a glance in europe. https://www.intersectionaljustice.org/img/2020.4.14_cij-factsheet-intersectionality-at-a-glance-in-europe_du2r4w.pdf, 2020.

John Cherian and Emmanuel Candés. Statistical inference for fairness auditing. *arXiv*, https://arxiv.org/pdf/2305.03712.pdf, 2023.

Sam Corbett-Davies, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. The measure and mismeasure of fairness. *Journal of Machine Learning Research*, 24(1), 2024. ISSN 1532-4435.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=bYi_2708mKK.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, January 2012. Association for Computing Machinery. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255. URL https://dl.acm.org/doi/10.1145/2090236.2090255.

European Union. Regulation (eu) 2021/0106 of the european parliament and of the council: Laying down harmonised rules on artificial intelligence (artificial intelligence act) and ammending certain union legislative acts. *Official Journal of the European Union*, 2021a.

European Union. Recital 49 of the eu ai act. https://artificialintelligenceact.eu/recital/49/, 2021b.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921, 2020a. doi: 10.1109/ICDE48307.2020.00203.

James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. *Bayesian Modeling of Intersectional Fairness: The Variance of Bias*, pages 424–432. Society for Industrial and Applied Mathematics, 2020b. doi: 10.1137/1.9781611976236.48. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611976236.48.

Harry Frankfurt. Equality as a Moral Ideal. *Ethics*, 98(1):21–43, October 1987. ISSN 00141704. doi: 10.1086/292913. URL http://www.jstor.org/stable/2381290.

Sarah Giest and Annemarie Samuels. 'For good measure': data gaps in a big data world. *Policy Sciences*, April 2020. ISSN 1573-0891. doi: 10.1007/s11077-020-09384-1. URL https://doi.org/10.1007/s11077-020-09384-1.

Christine Herlihy, Kimberly Truong, Alexandra Chouldechova, and Miroslav Dudík. A structured regression approach for evaluating model performance across intersectional subgroups. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 313–325. ACM, 2024. ISBN 979-8-4007-0450-5. doi: 10.1145/3630106.3658908. URL https://dl.acm.org/doi/10.1145/3630106.3658908.

M. Hirzel and M. Feffer. A suite of fairness datasets for tabular classification. https://arxiv.org/pdf/2308.00133.pdf, 2023.

IBM/lale. Lale fairness dataset sample notebook. https://github.com/IBM/lale/blob/master/examples/demo_fairness_datasets.ipynb, 2023.

Abigail Z. Jacobs and Hanna Wallach. Measurement and Fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, March 2021. doi: 10.1145/3442188.3445901. URL http://arxiv.org/abs/1912.05511. arXiv: 1912.05511.

M. Kearns, S. Neel, A. Roth, and Z.S. Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *The 35 th International Conference on Machine Learning*, 2018.

Falaah Arif Khan, Denys Herasymuk, and Julia Stoyanovich. On fairness and stability: Is estimator variance a friend or a foe?, 2023. URL https://arxiv.org/abs/2302.04525.

Goda Klumbytė, Claude Draude, and Alex S. Taylor. Critical tools for machine learning: Working with intersectional critical concepts in machine learning systems design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1528–1541. Association for Computing Machinery, 2022. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533207. URL https://dl.acm.org/doi/10.1145/3531146.3533207.

Youjin Kong. Are "intersectionally fair" AI algorithms really fair to women of color? a philosophical analysis. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 485–494. Association for Computing Machinery, 2022. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533114. URL https://dl.acm.org/doi/10.1145/3531146.3533114.

Seth Lazar and Jake Stone. On the Site of Predictive Justice. *Noûs*, 58(3), September 2024. ISSN 1468-0068. doi: 10.1111/nous.12477.

Jonas Lerman. Big Data and Its Exclusions. *Stanford Law Review*, September 2013. URL https://www.stanfordlawreview.org/online/privacy-and-big-data-big-data-and-its-exclusions/.

Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 379–389, New York, NY, USA, June 2022. Association for Computing Machinery. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533105. URL https://dl.acm.org/doi/10.1145/3531146.3533105.

Mathieu Molina and Patrick Loiseau. Bounding and approximating intersectional fairness through marginal fairness. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/6ae7df1f40f5faeda474b36b61197822-Abstract-Conference.html.

Giulio Morina, Viktoriia Oliinyk, Julian Waton, Ines Marusic, and Konstantinos Georgatzis. Auditing and achieving intersectional fairness in classification problems. *CoRR*, abs/1911.01468, 2020. URL http://arxiv.org/abs/1911.01468.

Derek Parfit. Equality and Priority. *Ratio*, 10(3):202–221, December 1997. ISSN 0034-0006. doi: 10.1111/1467-9329.00041. URL http://www.blackwell-synergy.com/links/doi/10.1111%2F1467-9329.00041.

Amartya Sen. *On economic inequality*. Clarendon Press, Oxford, enl. ed., edition, 1997. ISBN 978-0-19-829297-5.

Michael A. Slote. *Beyond Optimizing: A Study of Rational Choice*. Harvard University Press, Cambridge Mass., 1989. ISBN 978-0-674-06918-3.

Rush T. Stewart. Identity and the limits of fair assessment. *Journal of Theoretical Politics*, 34(3):415–442, 2022. ISSN 0951-6298. doi: 10.1177/09516298221102972. URL https://doi.org/10.1177/09516298221102972.

Harini Suresh, Rajiv Movva, Amelia Lee Dogan, Rahul Bhargava, Isadora Cruxen, Angeles Martinez Cuba, Guilia Taurino, Wonyoung So, and Catherine D'Ignazio. Towards intersectional feminist and participatory ML: A case study in supporting feminicide counterdata collection. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 667–678. Association for Computing Machinery, 2022. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533132. URL https://dl.acm.org/doi/10.1145/3531146.3533132.

Angelina Wang, Vikram V Ramaswamy, and Olga Russakovsky. Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 336–349. Association for Computing Machinery, 2022. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3533101. URL https://dl.acm.org/doi/10.1145/3531146.3533101.