# Identity and the Limits of Fair Assessment

Rush T. Stewart*

King's College London

May 9, 2022

**Abstract**

In many assessment problems—aptitude testing, hiring decisions, appraisals of the risk of recidivism, evaluation of the credibility of testimonial sources, and so on—the fair treatment of different groups of individuals is an important goal. But individuals can be legitimately grouped in many different ways. Using a framework and fairness constraints explored in research on algorithmic fairness, I show that eliminating certain forms of bias across groups for one way of classifying individuals can make it impossible to eliminate such bias across groups for another way of dividing people up. And this point generalizes if we require merely that assessments be approximately bias-free. Moreover, even if the fairness constraints are satisfied for some given partitions of the population, the constraints can fail for the coarsest common refinement, that is, the partition generated by taking intersections of the elements of these coarser partitions. This shows that these prominent fairness constraints admit the possibility of forms of intersectional bias.

**Keywords.** Algorithmic fairness; bias; calibration; equalized odds; intersectionality

## 1 Introduction

Individual identity is multifaceted. Hannah, for instance, is a woman, an American, from New York City (specifically the Upper West Side), but a resident of the South, a person who spent several formative years in England, a Cambridge graduate, a philosophy DPhil, an academic, from an upper middle class background, an advocate of risk literacy, a runner, a violist, Jewish, a flexitarian, heterosexual, an effective altruism enthusiast, a mother, a sister, a wife, and a fan of Andrei Tarkovsky and Townes van Zandt. Any one of these properties applies to a large number of other people, defining a subgroup of a general population of individuals. For a given individual, different social contexts may make membership in different groups more or less salient. The relative importance attached to membership in such groups is also a matter of individual discretion, at least to some degree. But one and the same individual

*Thanks to Marshall Bierson, Mike Bishop, Yang Liu, Michael Nielsen, Ignacio Ojea Quintana, Shanna Slank, Tom Sterkenburg, Reuben Stern, Borut Trpin, audiences at the Center for Advanced Studies (CAS) at LMU Munich and the Faculty of Philosophy at the University of Groningen, three anonymous referees at *Social Choice and Welfare*, and two anonymous referees at the *Journal of Theoretical Politics* for helpful conversations and feedback. I am grateful to CAS for providing research leave, and to the Cambridge-LMU Strategic Partnership for funding the Decision Theory and the Future of Artificial Intelligence group.

can be a member of all of these groups without contradiction. Since similar remarks apply to any individual, there are many legitimate ways to group individuals in a population, from marital status or nationality to religion or taste in music.

Fair treatment of different groups is an objective common to many domains of assessment including aptitude testing in psychometrics (Borsboom et al., 2008), hiring decisions in the labor market (Fang and Moro, 2011), risk assessment in the criminal justice system (Kleinberg et al., 2017; Pleiss et al., 2017), and evaluation of the credibility of testimonial sources in epistemology (Stewart and Nielsen, 2020).[1] Consider the case of risk assessment. Using the same actuarial techniques that are used to calculate insurance premiums, statistical software is employed in the U.S. criminal justice system to assess an individual's risk of re-offending. Given the type of crime committed, age, sex, employment status at the time of arrest, criminal history, etc., an individual is assigned (what can be thought of as) a probability of recidivism. Such scores are used in sentencing and parole decisions among other things. A 2016 *ProPublica* analysis of the risk scores of the COMPAS statistical tool for Broward County, Florida found a form of bias in the data on the tool's predictions (Angwin et al., 2016b). The rate of false positives—the percentage of non-recidivists given a high risk score—was roughly twice as great among black defendants as among white. And the rate of false negatives—the percentage of recidivists being given a low risk—among whites was roughly twice as great as among blacks.[2] The bias is that these types of errors were asymmetrically distributed across black and white sub-populations, affecting the lives of black and white people in very different ways.

Research on algorithmic fairness studies the prospects of unbiased assessment. Bias in error rates is one form of bias, but not the only form and often considered not the most important form. Can bias in error rates and other important forms of bias be simultaneously eliminated? One lesson that emerges from some of these studies is that eliminating one form of bias can mean that it is impossible to eliminate another. Sometimes, then, we face a conflict between eliminating different forms of bias. Here, I argue that, not only do we face a conflict in eliminating different forms of bias, we also face a conflict in eliminating one form of bias across different groupings. Eliminating a certain form of bias across groups for one way of categorizing people in a population can mean that it is impossible to eliminate that form of bias across groups for another way of classifying them. This conflict is significant to the extent that multiple classifications are relevant. And they often are: consider the various classes mentioned in standard non-discrimination clauses, for example.[3] Moreover, even if our assessments are unbiased for certain ways of classifying people—say for both a race classification that includes black and white categories and a gender classification that includes categories for women and men—bias can persist for the coarsest common refinement of these classifications—in this case, the single classification that includes the groups of black women, black men, white women, and white men. In other words, forms of intersectional bias are

[1] A number of different formal frameworks have been brought to bear on issues of group discrimination and disparate treatment. For a recent study of connections between some of these frameworks, see (Patty and Penn, MS).

[2] Here, rearrests must be used as a proxy for recidivism.

[3] Facebook's non-discrimination policy for advertisement: "Ads must not discriminate or encourage discrimination against people based on personal attributes such as race, ethnicity, color, national origin, religion, age, sex, sexual orientation, gender identity, family status, disability, medical or genetic condition" (Facebook, 2022).

possible for the prominent fairness constraints in the fair algorithms literature. Given the conceptions of fairness encoded in these constraints, and confronted with the sorts of limitations in achieving fairness across various classifications discussed below, we must reconcile ourselves with lingering bias against some groups.

## 2 Identity and Population Partitions

In any assessment problem, there is a particular population of individuals that is relevant. For instance, in parole decisions in Broward County in 2015, there are the people coming before the parole boards in the county that year. In SAT testing in the U.S. for the last decade, there is the population of people who took the exam in that time frame. In Faceboook's assessment of the trustworthiness of its users in an effort to combat "fake news," there is the set consisting of nearly 2.5 billion active monthly users (Dwoskin, 2018). The first element of our model, then, is a finite population of individuals, $N = \{1, 2, \ldots, n\}$.[4]

Any population can be divided into groups according to various individual properties or identities. For the question of fair treatment in assessment, certain groups are more customary to consider than others. Often, history and social context make salient particular categories. In the *ProPublica* story mentioned above, the focus is on the disparate treatment of different races. In particular, black and white defendants were treated differently. My interest here is in the possibility of fair treatment across various partitions of a population. A *partition* of a (finite) set $N$ is a collection $\pi = \{G_1, \ldots, G_m\}$ of non-empty subsets of $N$ such that the elements of $\pi$ are mutually disjoint and collectively exhaustive.[5] In other words, each individual in $N$ belongs to exactly one group $G_k$. Whether the black and white racial categories partition a population is a contingent matter that depends on the composition of the population. If the population contains Latino and Asian people, the groups of black and white people might fail to exhaust the population. Similar points apply to a gender partition $\{M, F\}$ consisting of males and females, or a sexuality partition $\{G, S\}$ consisting of groups of homosexual and heterosexual individuals, and so on.[6]

Why should we be concerned about the multiple group memberships of any given individual and the multiple possible ways a population can be partitioned? In his book *Identity and Violence*, Sen argues that there is moral urgency to considering the various aspects of individual identity. Reckoning with "the power of *competing* identities," says Sen, "leads to other ways of classifying people, which can restrain the exploitation of a specifically aggressive use of one particular categorization" (Sen, 2007, p. 4). Sen has in mind the way in which considering one's humanitarian or religious affiliations might weaken the pull of a

[4] Often, it is useful to represent an individual as a vector of features including age, sex, employment status, etc. For simplicity, my representation of individuals abstracts from the list of features.

[5] That is, $G_j \cap G_k = \emptyset$ if $j \neq k$ and $\bigcup_{k=1}^{m} G_k = N$. I will sometimes refer to elements of a partition as *cells* of the partition.

[6] That identities introduce certain partitions of $N$ is to some extent a simplifying assumption. We could consider a collection of groups that aren't mutually exclusive, for example. But the assumption should not be worrisome for a couple of reasons. First, the partition assumption is not so conceptually restrictive. We can always carve out additional cells to secure mutual exclusiveness. So the observations discussed below would still be of significant interest even if the partition assumption were necessary for all of them. Second, for many of the observations made below, the partition assumption is not necessary. For instance, we could easily state a version of Observation 1 for all groups of $N$, not necessarily ones forming partitions.

violent nationalistic or racist movement, for example. Throughout the book, Sen criticizes the idea that there is a uniquely appropriate or privileged partition.

> The insistence, if only implicitly, on a choiceless singularity of human identity not only diminishes us all, it also makes the world much more flammable. The alternative to the divisiveness of one preeminent categorization is not any unreal claim that we are all much the same. That we are not. Rather, the main hope of harmony in our troubled world lies in the plurality of our identities, which cut across each other and work against sharp divisions around one single hardened line of vehement division that allegedly cannot be resisted. (Sen, 2007, p. 16)

My concern is a bit different from Sen's, but there is a relevant lesson in the passage quoted just above. To wit, a single, fixed partition is overly constraining, and may frustrate our goals and lead to sub-optimal outcomes. The goal in assessment that is our focus is the fair treatment of different groups. But since there are various legitimate ways to partition a population into groups, restricting our attention to a single partition potentially commits us to ignoring important forms of group bias.[7]

Consider once again the bias found against black people in the COMPAS data. In that same Broward County data set, there is a similar amount of bias in error rates against women compared to men, as a companion piece in *ProPublica* makes clear (Angwin et al., 2016a). Bias against either group is ethically relevant. Satisfying certain central fairness constraints (described in the following section) for a race partition does not imply that those constraints are satisfied for a gender partition. Still other partitions could be pertinent. The relevant social identities cannot be decided a priori, without appeal to contingent social context and values. Sen points out that even intuitively unimportant aspects of personal identity can become important. Consider, for example, those who wear a size 8 shoe, or those born between nine and ten in the morning, local time. If size 8 shoes were to become extremely difficult to find—think "high noon" of Soviet civilization or broken supply chains due to a novel coronavirus pandemic—then being someone who wears that shoe size may become an important part of one's identity and grounds for solidarity with those similarly unshod. Likewise, if an authoritarian ruler were to elect to severely curtail the freedoms of people born between nine and ten in the morning due to some supernatural belief or other, then the hour of one's birth and the persecution it entails for some is, again, likely to become an important aspect of one's identity and grounds for solidarity (Sen, 2007, pp. 26-27). In either case, new forms of bias become pressing considerations. The priority of particular partitions in eliminating bias might reasonably depend not just on past history of discrimination, but also on current deprivation. What groups suffer discrimination and deprivation is a matter to which we may frequently need to reattend. They may well depend in part on the particular assessment problem confronting us, or on the particular population, as I explain at the end of Section 4. In general, and more to the point for this essay, there can be bias against many different groups.

[7] As I was informed after drafting this paper, this point has recently been acknowledged in the machine learning literature: "There are exponentially many ways of carving up a population into subgroups, and we cannot necessarily identify a small number of these a priori as the only ones we need to be concerned about" (Kearns et al., 2018, p. 2565). Also see Hébert-Johnson et al. (2018).

## 3 Fair Assessment

In order to foreground the issue of different population partitions, let's assume that there is just a single property $y$ of interest (this is a standard assumption in the literature anyhow). Individuals in $N$ either have property $y$ or lack it. We can represent this with a random variable $Y : N \to \{0, 1\}$ that assigns 1 to individual $i$ if $i$ has the property, and assigns 0 to $i$ if $i$ lacks it. In the case of risk assessment, $Y(i) = 1$ would indicate that individual $i$ is a recidivist. In credibility assessment, $Y(i) = 1$ might represent that $i$ is credible or is above some credibility threshold. Call a function $h : N \to [0, 1]$ an *assessor*. In risk assessment, we can think of $h$ as assigning each individual a probability of re-offending. In credibility assessment, $h(i)$ could be interpreted as the probability that $i$ is credible (or above the credibility threshold). But $h$ need not be thought of as assigning probabilities. We could just as well introduce basic risk, aptitude, credibility, etc. scores. For concreteness, I will interpret $h(i)$ as the assessor's probability that $i$ has property $y$.

In order to talk about population proportions or frequencies, let's introduce a uniform probability distribution $P$ on $N$. The quantity $P(Y = 1) = \mu$, for example, is the proportion of people in $N$ that have property $y$, the prevalence of $y$ in the population.[8] Call $\mu$ the *base rate* for $y$ in $N$. Given a partition $\pi = \{G_1, \ldots, G_m\}$ of $N$, let $P_k = P(\cdot|G_k)$ be the uniform probability distribution on $G_k$ for $k = 1, \ldots, m$. The quantity $P_1(Y = 1) = \mu_1$, for example, is the base rate for $y$ in group 1; $P_2(h = 0.5)$ is the proportion of people to which $h$ assigns 0.5 in $G_2$, and so on.

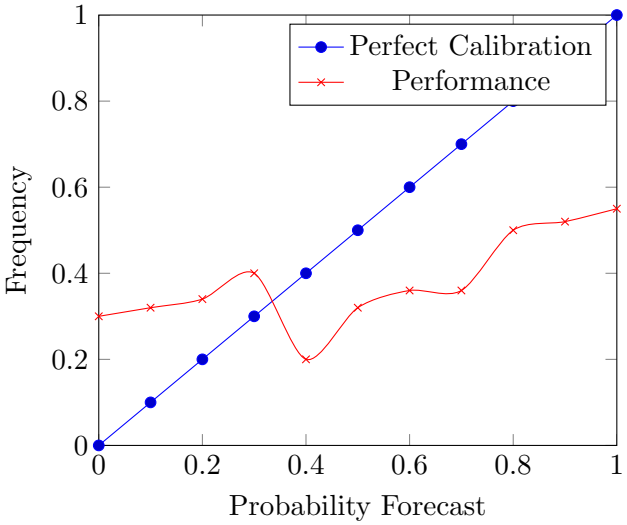
The interesting issue concerns what properties $h$ should have in order to qualify as fair or unbiased. The "dominant fairness criterion" in the literature on risk assessment is calibration (Corbett-Davies et al., 2017, p. 799). An assessor is *calibrated* if $P_k(Y = 1|h = p) = p$ for all $p \in [0, 1]$ and $k = 1, 2, \ldots, m$ such that $P_k(h = p) > 0$. This property is familiar from work on the foundations of probability. According to proponents of calibration, good forecasts should track observed frequencies (van Fraassen, 1983; Shimony, 1988). Consider weather forecasting. Suppose that each day, a forecaster announces a probability of rain for that day. The forecaster is calibrated if it rains on 10% of the days she announces that it will rain with probability 0.1, and it rains on 85% of the days she predicts rain with probability 0.85, etc. Similarly, a risk assessor is calibrated if, among those it assigns a 0.1 probability of re-offending, 10% re-offend, and, among those it assigns a 0.85 probability of re-offending, 85% re-offend, etc.

Why does it make sense to think of calibration as a fairness constraint? One reason is that it guards against a form of bias in confidence. If it rains on 100% of the days a forecaster predicted rain with probability 0.8, then the forecaster is *underconfident* in rain on those days. Likewise, a forecaster would be *overconfident* if it rains on only 50% of the days rain was predicted with probability 0.8. It can be helpful to visualize these concepts with a calibration curve like in Figure 1.[9] Here, 30% of the events the forecaster assigned probability 0 occurred, while only 55% of the events assigned probability 1 occurred. The forecaster's assessments were underconfident in the first case, and overconfident in the latter. Instead of weather forecasting, consider the problem of risk assessment again. If only 50% of black people assigned a score of 0.8 go on to re-offend, then the assessor is overconfident

[8] Here, $Y = 1$ is shorthand for the event $\{i \in N : Y(i) = 1\}$.

[9] Figure 1 is loosely based on a calibration curve for actual forecasts from Tetlock's work on expert political judgment (Tetlock, 2005, p. 55).

Figure 1: Calibration Curve




that blacks assigned that score will be recidivists. If 100% of white people assigned a score of 0.8 go on to re-offend, then the assessor is underconfident that whites assigned that score will be recidivists. If the assessor were calibrated, not only would it not be overconfident in one group and underconfident in another, it would not be over- or underconfident in any of its assessments. Another way calibration is sometimes motivated is by pointing out that it implies that scores "mean" the same for individuals in different groups. About calibrated assessors Kleinberg et al., for example, write, "we are justified in treating people with the same score comparably with respect to the outcome, rather than treating people with the same score differently based on the group they belong to" (Kleinberg et al., 2017, pp. 4-5). If 50% of black people assigned a score of 0.8 go on to re-offend while 100% of white people assigned a score of 0.8 go on to re-offend, there is a sense in which the assessment score of 0.8 means something different for individuals in the two groups.

I think there are some reasonable concerns one might have about construing calibration as a fairness property. It seems to me that this latter line of motivation in terms of meaning—and to some extent even the previous one in terms of under- and overconfidence—fails to fully motivate calibration as a fairness constraint. Scores for individuals in different groups can mean the same without the assessor satisfying the full calibration constraint. Calibration implies that $P_k(Y = 1|h = p) = P_j(Y = 1|h = p)$ for all $G_k, G_j \in \pi$ when those conditional probabilities are defined. Clearly, if $P_k(Y = 1|h = p)$ and $P_j(Y = 1|h = p)$ are both equal to $p$ (calibration), then those terms are equal to each other, but they could both be equal to some other value. Say that an assessor $h$ satisfies *predictive equity* for a partition $\pi$ if $P_k(Y = 1|h = p) = P_j(Y = 1|h = p)$ for all $G_k, G_j \in \pi$ and all $p$ for which the conditional probabilities are defined.[10] Put another way, among people assigned the same assessment

[10]Thanks to Shanna Slank for suggesting that Michael Nielsen and I investigate essentially this property. Predictive equity appears to have first been called *sufficiency* in the literature on algorithmic fairness (e.g., Mitchell et al., 2018; Barocas et al., 2019). Michael and I were unaware of this when we wrote (Stewart and

score, the proportion of people who have property $y$ is the same across all groups in the partition. Predictive equity simply retains the explicit equal treatment aspect of calibration.

There are at least two cautions about relaxing calibration to predictive equity worth considering. The first is that fairness is not the only goal in assessment. We also care about the property being assessed after all. We care about maintaining public safety, admitting a talented class of freshmen, trusting credible testimonial sources, making prudent loan decisions, etc. That is, there is typically a purpose for which an assessment is conducted, with fairness acting as a sort of constraint. So, it may be reasonable to retain the form of accuracy that calibration adds to predictive equity. Not only should it be the case that $P_j(Y = 1|h = p) = P_k(Y = 1|h = p)$, but it should also be the case that those terms track the actual frequencies in the population. According to this way of understanding the property, calibration captures both a fairness concern and an epistemic concern about accuracy (Stewart and Nielsen, 2020). The second concern about relaxing calibration to predictive equity is that accuracy may represent a type of fair treatment itself. While predictive equity does not permit being simultaneously overconfident in individuals in one group at a given assessment score and underconfident in individuals in another group at that same assessment score, it does permit being uniformly under- or overconfident in individuals of a given assessment score. If only 50% of people assigned a risk score of 0.8 are recidivist in each race group, then those individuals might still be considered the victims of unfair assessment. Even if those in one group haven't been treated more harshly relative to those given the same risk score in the other group, they have been treated too harshly in the sense that the assessor is overconfident that they will reoffend. Calibration prevents this.

Rather than considering ways to relax calibration, we might consider alternative fairness constraints, ones that might potentially supplement calibration. Even if calibration is necessary for unbiased assessment, it may not be sufficient. An assessor that simply predicts the group base rate for everyone in the group will be calibrated. Yet, an innocent person in a group with a high recidivism base rate, for example, might have grounds for complaint when he receives a higher risk score than his counterpart in a group with a lower base rate. Similarly, it is consistent with calibration for a recidivist in a low base rate group to receive a lower risk score than a non-recidivist in a high-base-rate group. One reading of these points is that there are other forms of bias to consider besides the one calibration attempts to eliminate. This reading seems supported by *ProPublica*'s analysis of the COMPAS data. The sort of bias that they charge the statistical tool with is not a failure of calibration, but a disparity in error rates across groups. I turn now to a constraint meant to eliminate for exactly this type of bias.

To introduce the constraint, we need a few auxiliary definitions. The *false positive rate*

Nielsen, 2020). Sometimes the property is *also* called calibration and formulated as follows: for all $G_k \in \pi$, $P(Y = 1|h = p) = P_k(Y = 1|h = p)$. But, as Mitchell et al. point out, this is unfortunate terminology since this property is distinct from the property we call—and indeed is standardly called—calibration (Mitchell et al., 2018, p. 5). To see that it is the same as predictive equity, assume that $P_k(Y = 1|h = p) = P_j(Y = 1|h = p) = r$ for all $G_k, G_j \in \pi$ and some $r \in [0, 1]$ as predictive equity demands, and use the law of total probability to compute

$$P(Y = 1|h = p) = \sum_{k=1}^{m} P(Y = 1|h = p, G_k)P(G_k|h = p) = \sum_{k=1}^{m} rP(G_k|h = p) = r.$$

of an assessor $h$ for group $G_k$ is given by $f_k^+(h) = E_k(h|Y = 0)$.[11] In words, the false positive rate for group $G_k$ is the average assessment score of individuals *lacking* property $y$ in group $G_k$. The *false negative rate* is $f_k^-(h) = E_k(1 - h|Y = 1)$. In words, the false negative rate for group $G_k$ is equal to 1 minus the average assessment score of individuals *possessing* property $y$ in group $G_k$. An assessor satisfies *equalized odds* if $f_j^+(h) = f_k^+(h)$ and $f_j^-(h) = f_k^-(h)$ for all $G_j, G_k \in \pi$. Equalized odds guarantees that errors are not asymmetrically distributed across groups of the partition.[12] In general, calibration does not imply equalized odds nor does equalized odds imply calibration. For example, a calibrated assessor can assign much greater average risk scores to non-recidivists in one group than to non-recidivists in another (and higher even than the average risk scores assigned to *recidivists* in the other group). Moreover, the COMPAS data gives us at least prima facie reasons to be concerned about the form of bias motivating the introduction of equalized odds.

In this essay, I will grant that calibration and equalized odds are prima facie compelling fairness constraints, though I consider it legitimate to subject them to further scrutiny in general and plan do to so in future work. On the one hand, the reader might agree that the properties are important formalizations of unbiased assessment. Many have found them to have considerable intuitive plausibility. So, the consequences of the properties would seem ethically important for such readers. On the other hand, because of the prominence of these sorts of statistical properties in theories of algorithmic fairness, it is crucial to scrutinize them, to explore their consequences and their limitations. So, even if the reader is unconvinced of the normative status of the properties, the consequences of these properties are relevant to a sober evaluation of them.[13]

## 4 Some Limitations

Unfortunately, there are limits to the extent assessments can be unbiased. Let's look at one central limitative result. An assessor is *perfect* if $h(i) = Y(i)$ for all $i \in N$. In practice, perfect assessment is rarely achievable in interesting or non-trivial cases. So, if some standard of fairness is achievable only by perfect assessment, it is reasonable to think the standard is too high for interesting assessment problems of practical concern. If we think of the assessor as a probability judgment that some individual has property $y$, then perfection requires only assigning probabilities 0 and 1 and not making any mistakes. But much of the power and applicability of probability theory comes with non-extreme judgments. The same is true of assessments. We can now state one of the central limitative results for fair assessment.

**Theorem 1.** *(Kleinberg et al., 2017) Let $h$ be an assessor for $N$. The following are equivalent:*

1. *$h$ is calibrated and satisfies equalized odds for a partition $\pi$.*

[11] Recall that the conditional expectation $E_k(h|Y = 0)$ is the expected value of $h$ as computed with respect to the conditional probability $P_k(\cdot|Y = 0)$. What I am calling the false positive and false negative rates, Pleiss et al. call the *generalized* false positive and *generalized* false negative rates (2017).

[12] In no population are both the false positive and false negative rates defined for all cells of all partitions (consider singleton cells). We will say that $h$ satisfies equalized odds for a partition if, whatever false positive rates are defined are equal, and whatever false negative rates are defined are equal.

[13] A worry about the general approach undertaken here and in computer science and machine learning is that it runs the risk of conflating "statistical parities with more complex concepts" (Mitchell et al., 2018, p. 13). While I will put this issue aside in this paper, I will just note that one might interpret the observations made below as support for explications of fairness that are not purely statistical.

2. *Either i) the base rates in all groups are exactly the same or ii) $h$ is perfect.*

Theorem 1 is widely regarded as an impossibility or triviality result for fair assessment. Corbett-Davies et al. report that Kleinberg et al. "prove that except in degenerate cases, no algorithm can simultaneously satisfy" calibration and equalized odds (Corbett-Davies et al., 2017, p. 799); on the basis of this result, journalists at *ProPublica* published a followup article entitled "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say" (Angwin and Larson, 2016). The idea is that perfection, as discussed, is very rarely achievable in real-life, interesting assessment problems. Similarly, that the base rates for the relevant groups are *exactly* the same is only very rarely the case. As a result, outside of very rare circumstances, it is impossible to achieve both fairness properties. Results like Theorem 1 give us reason to explore ways to relax or modify the fairness constraints.

Each of calibration and equalized odds is meant to eliminate a certain form of bias. What Theorem 1 establishes is that, for a fixed way of carving the population into groups, eliminating one form of bias makes it impossible to eliminate another. Next, I consider requiring the individual fairness constraints on assessment hold for all partitions of the population. Clearly, we cannot expect these constraints to be *jointly* satisfied in assessment for multiple partitions since, by Theorem 1, they cannot be simultaneously satisfied for a single partition. Instead, I consider each property on its own. Under the assumption that each fairness constraint eliminates a form of bias that is desirable to eliminate, I study the possibility of eliminating one form of bias across multiple ways of dividing the population into groups.

Let's consider each constraint in turn, starting with calibration. An alternative way to strengthen calibration for a single partition is to require it for multiple partitions rather than imposing a different sort of fairness constraint like equalized odds on the same partition. Again, Theorem 1 gives us reason to seek such alternatives. When confronted with the limitation expressed in Theorem 1, a number of people have suggested to me in person that calibration is clearly the more compelling condition. Perhaps bias of types that calibration fails to exclude—for example, some version of bias in error rates—could be reduced by requiring calibration for multiple partitions. The best case would be for calibration to hold for all partitions, since that would exclude bias in confidence against *any* group and maybe other forms of bias to boot. Observation 1 states a limitation on this strategy.

**Observation 1.** *Let $h$ be an assessor for $N$. The following are equivalent:*

1. *$h$ is calibrated for all binary partitions.*[14]

2. *$h$ is calibrated for all partitions.*

3. *$h$ is perfect.*

Calibration for all partitions, then, is only achievable in the typically unrealistic case of perfect assessment (cf. Hébert-Johnson et al., 2018, p. 1940).[15] In other words, outside of the unrealistic case of perfect assessment, there will be bias in confidence against some group.

[14]The first clause of Observation 1—just like the first clauses of Observations 2 and 3—could be replaced with the proposition that $h$ is calibrated for all partitions of any cardinality $m$ with $2 \leq m \leq n$.

[15]Throughout the paper, proofs are confined to the Appendix. Proofs for Observations 1, 2, and 3, however, are omitted since they are fairly straightforward and these observations are generalized below by Observations 1′, 2′, and 3′, respectively.

Observation 1 complicates any automatic inference from failure of calibration for some group to *intentional* bias on behalf of the assessor (for further discussion of this point, see Stewart and Nielsen, 2020, Sec. 5).

Next, let's consider the weaker predictive equity property. Say that an assessor $h$ makes *perfect distinctions* if, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. For any score $p$, if $h(i) = p$ and $Y(i) = 1$, then for no individual $j$ such that $Y(j) = 0$ is it the case that $h(j) = p$. Compare calibration for a fixed partition. Calibration allows for individuals that differ with respect to property $y$ to receive the same score $p$ so long as the proportion of individuals who possess the property among those who receive the score $p$ is $p$. In fact, calibration for a fixed partition generally *requires* individuals who differ with respect to property $y$ to receive the same assessment; otherwise, for values in $(0, 1)$, the assessor would be under- or overconfident in the group.

**Observation 2.** *Let $h$ be an assessor for $N$. The following are equivalent:*

1. *$h$ satisfies predictive equity for all binary partitions.*
2. *$h$ satisfies predictive equity for all partitions.*
3. *$h$ makes perfect distinctions.*

Aside from assessors that make perfect distinctions, scores will not "mean" the same thing for all groups; there will be bias against some group. In large populations, perfect distinctions is very difficult to achieve—not as difficult as perfect assessment, but difficult nonetheless. Observations 4 and 5 below relate perfect distinctions to perfection.

Say that $h$ satisfies *perfect non-discrimination* when, for all $i, j \in N$, $Y(i) = Y(j)$ implies that $h(i) = h(j)$. Notice that a degenerate case of perfect non-discrimination is a constant assessor: for some $p \in [0, 1]$, $h(i) = p$ for all $i \in N$. Perfect non-discrimination—the converse of perfect distinction—captures the ideal that likes are treated alike, whereas perfect distinction captures the ideal that unalike individuals are *not* treated alike. We can now make the following observation regarding equalized odds.

**Observation 3.** *Let $h$ be an assessor for $N$. The following are equivalent:*

1. *$h$ satisfies equalized odds for all binary partitions.*
2. *$h$ satisfies equalized odds for all partitions.*
3. *$h$ is perfectly non-discriminatory.*

Here is another way to think about perfect non-discrimination. A higher bar than making perfect distinctions would be making perfect distinctions while limiting assessments to just two scores. Then, the binary assessor perfectly sorts the population into two groups: those having property $y$ are assigned one score $p$, while all of those lacking the property are assigned another score $p' \neq p$. A very low bar, on the other hand, would be a constant assessor, that is, an assessor that assigns the same score to every individual in the population. Such assessments may fail to carry any information at all. If $h$ is perfectly non-discriminatory, then $h$ is either constant and uninformative or, if non-constant, binary, sorting the population into

two groups perfectly (more on this implication after Observation 5 below). So unless $h$ takes one of these rather restrictive forms, it will violate equalized odds for some way of partitioning the population. Since equalized odds rules out bias in error rates, we know that such bias in error rates is unavoidable outside of the two restrictive cases just mentioned.

### 4.1 Connections

Certain mathematical relationships between the limitations in Observations 1, 2, and 3 are easy to state. For instance, a simple numerical transformation can convert an assessor that makes perfect distinctions into a perfect assessor.

**Observation 4.** *Let $h$ be an assessor for $N$.*

1. *$h$ makes perfect distinctions if and only if there exists a function $g : [0, 1] \to \{0, 1\}$ such that $g \circ h$ is perfect.*

2. *$h$ is non-constant and satisfies perfect non-discrimination if and only if the population is not homogeneous with respect to $y$ and there exists an injective function $g : [0, 1] \to \{0, 1\}$ such that $g \circ h$ is perfect.*

Barocas et al. point out that, for any assessor that satisfies predictive equity, there is a transformation of it that is calibrated (Barocas et al., 2019, Proposition 1, p. 52). On the basis of this observation, they conclude that predictive equity and calibration are "essentially equivalent notions" (Barocas et al., 2019, p. 52). In my view, such an interpretation is unwarranted,[16] but these sorts of transformations at least succinctly state a type of connection between concepts. The existence of a transformation to a perfect assessor characterizes those assessors that make perfect distinctions and so characterizes assessors that satisfy predictive equity for all partitions. For non-homogeneous populations, the existence of an injective transformation to a perfect assessor characterizes those assessors whose scores sort the population into a $y$ group and its complement, thereby characterizing the class of non-constant assessors that satisfy equalized odds for all partitions.

Using the limitations, we can also easily mark some logical relationships between the fairness constraints when they hold for all partitions for non-constant assessors.

**Observation 5.** *Let $h$ be a non-constant assessor for $N$.*

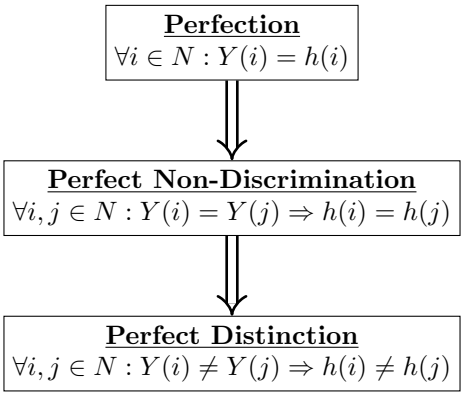
1. *If $h$ is calibrated for all partitions, then $h$ satisfies equalized odds for all partitions.*

2. *If $h$ satisfies equalized odds for all partitions, then $h$ satisfies predictive equity for all partitions.*

Since the observation is a fairly immediate consequence of preceding ones, I will just sketch a quick supporting argument here. If $h$ is calibrated for all partitions, then, by Observation 1, $h$ is perfect. Perfect assessors clearly give the same score to $i$ and $j$ when $Y(i) = Y(j)$, namely, 1 or 0 depending on whether $Y(i)$ is 1 or 0. So clause 1 follows from Observations 3. For clause 2, suppose that $h$ satisfies equalized odds for all partitions. By Observation 3, $h$ is perfectly non-discriminatory. Since $Y$ is a binary random variable, with the assumption that

[16] Consider, for example, that, unlike the case with calibration, predictive equity and equalized odds *are* jointly satisfiable outside of the cases of equal base rates and perfect prediction (Stewart and Nielsen, 2020).

Figure 2: Relations among Fairness Ideals for Non-Constant Assessors




$h$ is non-constant, it follows that $h$ is binary, assigning each individual one of just two scores. Suppose that, for some $i, j \in N$, $Y(i) \neq Y(j)$. For reductio, suppose that $h(i) = h(j)$. Since $h$ is non-constant, there is some $k \in N$ such that $h(k) \neq h(i) = h(j)$. Since $Y$ is binary, either $Y(k) = Y(i)$ or $Y(k) = Y(j)$. In either case, perfect non-discrimination implies that $h(k) = h(i) = h(j)$, which is a contradiction. Hence, $h(i) \neq h(j)$. It follows that $h$ makes perfect distinctions. By Observation 2, it follows that $h$ satisfies predictive equity for all partitions.

We could think of what happens when a constraint is satisfied for *all* partitions as revealing what ideal of fairness the constraint is committed to. As satisfying one of the constraints is supposed to represent a form of fair assessment for the groups in a partition, satisfying a constraint for *all* partitions represents fair assessment for *all* groups. This is, plausibly, the ideal case. For the three criteria under consideration here, the ideals are very simple and so are the relationships between them (Figure 2).

## 4.2 Objections

I want to consider two objections to the significance of the foregoing limitative results. Both concern the potentially overly exacting nature of what is being asked for in avoiding bias completely against all groups. First, we might consider satisfying certain fairness constraints approximately rather than exactly. That is, we could confine the amount of bias to which any group is subject to a certain margin of tolerance. Second, we might consider avoiding bias for a certain collection of partitions, even if that collection is not the set of *all* partitions. I discuss these objections in turn.

One potential source of stringency that could be driving the limitative results is the requirement that a constraint has to be satisfied *exactly*. Instead, we could consider requiring that an assessor satisfies a fairness constraint *approximately*. Kleinberg et al. consider such a possibility for satisfying multiple fairness criteria approximately in light of Theorem 1. On this approach, an assessor is approximately fair for some margin of tolerance if, for each group, the assessments are within that margin. The guiding idea is that the fairness standards are relaxed to requiring only that assessors are unbiased "enough" for each group.

Only sufficiently small amounts of bias, in other words, are tolerated. Let's look at each constraint in turn.

Say that $h$ is *$\varepsilon$-approximately calibrated* for some partition $\pi$ if, for some $\varepsilon \geq 0$, all $G_k \in \pi$, and all $p$ such that the conditional probability $P_k(Y = 1|h = p)$ is defined, $|P_k(Y = 1|h = p) - p| \leq \varepsilon$. Rather than requiring that $P_k(Y = 1|h = p) = p$ exactly, in other words, approximate calibration requires only that, for each group $G_k \in \pi$, $P_k(Y = 1|h = p)$ is within $\varepsilon$ of $p$. Say that $h$ is *$\delta$-approximately perfect* if, for all $i \in N$, $|Y(i) - h(i)| \leq \delta$. The next observation establishes that approximate calibration for all partitions is equivalent to approximate perfection with $\delta = \varepsilon$.

**Observation 1′.** Let $h$ be an assessor for $N$. The following are equivalent:

1. $h$ is $\varepsilon$-approximately calibrated for all binary partitions.
2. $h$ is $\varepsilon$-approximately calibrated for all partitions.
3. $h$ is $\varepsilon$-approximately perfect.

Put another way, relaxing calibration in a continuous fashion is equivalent to relaxing perfection in a continuous way. Small deviations from calibration allow only (equally) small deviations from perfect assessment. Observation 1′ expresses the same qualitative limitation as Observatoin 1 for avoiding bias across all partitions.

Say that $h$ satisfies *$\varepsilon$-approximate predictive equity* for a partition $\pi$ if, for all $G_j, G_k \in \pi$, $|P_j(Y = 1|h = p) - P_k(Y = 1|h = p)| \leq \varepsilon$ for all $p$ for which the conditional probabilities are defined. Again, the idea is that we might require the relevant conditional probabilities to be "close enough" rather than exactly equal. (Notice that, when $\varepsilon = 1$, $\varepsilon$-approximate predictive equity is completely vacuous, placing no constraints on the respective probabilities. Observation 2′ excludes this case, assuming $0 \leq \varepsilon < 1$ in generalizing Observation 2.) We might consider saying that assessor $h$ makes *$\delta$-approximately perfect distinctions* if there is some $\delta > 0$ such that, for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $|h(i) - h(j)| > \delta$. But this *strengthens* the property of making perfect distinctions. Since $Y$ is binary, alternative formulations of approximate versions of making perfect distinctions are limited. The following observation, however, establishes a connection between $h$'s satisfying approximate predictive equity for any $\varepsilon \in [0, 1)$ and making perfect distinctions.

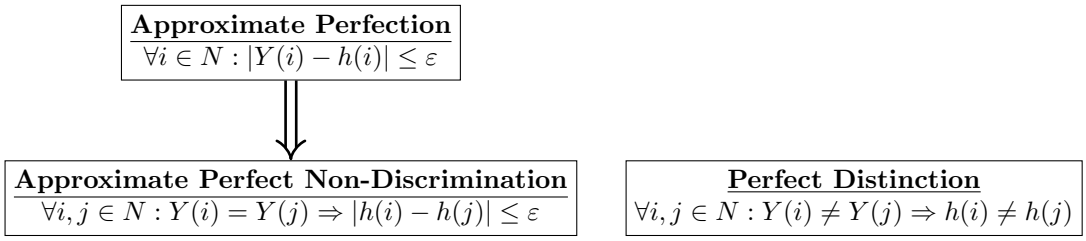
**Observation 2′.** Let $h$ be an assessor for $N$. The following are equivalent:

1. $h$ satisfies $\varepsilon$-approximate predictive equity for all binary partitions with $0 \leq \varepsilon < 1$.
2. $h$ satisfies $\varepsilon$-approximate predictive equity for all partitions with $0 \leq \varepsilon < 1$.
3. $h$ makes perfect distinctions.

Observations 2 and 2′ imply that (non-vacuous) approximate predictive equity for all partitions is, somewhat surprisingly, equivalent to predictive equity for all partitions. Moving to the approximate version of the constraint creates no new possibilities.

Finally, say that an assessor satisfies *$\varepsilon$-approximately equalized odds* for a partition $\pi$ if there exists some $\varepsilon \geq 0$ such that, for all $G_j, G_k \in \pi$, $|f_j^+(h) - f_k^+(h)| \leq \varepsilon$ and $|f_j^-(h) - f_k^-(h)| \leq \varepsilon$. In words, the false positive rates for any two cells of the partition are within $\varepsilon$ of each

Figure 3: Relations among Approximate Fairness Ideals for Non-Constant Assessors




other, as are the false negative rates. Say that an assessor $h$ satisfies $\delta$-*approximately perfect non-discrimination* when likes are treated approximately the same: for some $\delta \geq 0$ and for all $i, j \in N$, $Y(i) = Y(j)$ implies that $|h(i) - h(j)| \leq \delta$. The next observation establishes that when $h$ satisfies approximately equalized odds with respect to $\varepsilon$ for all partitions, $h$ satisfies approximately perfect non-discrimination with respect to $\delta = \varepsilon$. However, the equivalence between satisfying a constraint for all partitions and satisfying it for all *binary* partitions breaks down here. Now, we have that if $h$ satisfies approximately equalized odds with respect to $\varepsilon$ for all binary partitions, $h$ satisfies approximately perfect non-discrimination with respect to $\delta = 2\varepsilon$.

**Observation 3′.** Let $h$ be an assessor for $N$. Then:

1. $h$ satisfies $\varepsilon$-approximately equalized odds for all partitions if and only if $h$ is $\varepsilon$-approximately perfectly non-discriminatory.

2. If $h$ satisfies $\varepsilon$-approximately equalized odds for all binary partitions, then $h$ is $2\varepsilon$-approximately perfectly non-discriminatory.

On the basis of Observations 1′, 2′, and 3′ it seems fair to say that, for all three fairness criteria, the spirit of the limitations indicated by Observations 1, 2, and 3 survives the transition to the approximate versions of those criteria. Tolerating a small amount of bias relaxes the restrictiveness with which fairness can be achieved only by a correspondingly small amount—or, in the case of predictive equity, not at all.

When we require, not that the criteria hold exactly for all partitions, but only that they hold approximately for all partitions, the logical connections between the characterizing conditions are a bit different. These connections are represented in Figure 3. We can take any $\delta \geq \varepsilon$ in the statement of approximate perfect non-discrimination for the implication in the figure to go through. Excluding the trivial case of $\varepsilon = 1$, however, perfect distinctions—which is not stated in an approximate form for reasons previously given—is independent of approximate perfection and approximate perfect non-discrimination.

The second objection councils restraining our ambitions in a different way. One might be inclined to think that, while (a particular type of) unbiased assessment for multiple partitions is often desirable, we have overshot the mark by requiring it for *all* partitions. Consider calibration. There are simple examples of populations that allow for a imperfect assessor that is simultaneously calibrated for, say, two different non-trivial ways of partitioning the population.

**Example 1.** Let $N = \{1, 2, 3, 4, 5, 6\}$, and let $Y(i) = 1$ for $i = 1, 5, 6$. Having property $y$ is represented by an asterisk in Table 1. We consider a race partition $\{B, W\} = \{\{1, 2, 4\}, \{3, 5, 6\}\}$ and a sex partition $\{M, F\} = \{\{1, 2, 3\}, \{4, 5, 6\}\}$. To see that $h$ is cal-

Table 1: Calibration for Two Binary Partitions

| | B | W |
|---|---|---|
| M | $h(1^*) = 1/2,\ h(2) = 0$ | $h(3) = 1/2$ |
| F | $h(4) = 1/2$ | $h(5^*) = 1/2,\ h(6^*) = 1$ |

ibrated for the $\{M, F\}$ partition, observe that, in both groups, half of those assigned $1/2$ have property $y$. Similarly, none of those assigned 0 have property $y$ and all of those assigned 1 have property $y$. The same sort of inspection confirms that $h$ is also calibrated for the $\{B, W\}$ partition. ◇

So one idea might be that there is a small set of relevant partitions $\Pi$, and we should look for calibrated assessments only for members of $\Pi$. However, there are also simple examples of populations—involving a different distribution of property $y$—such that no imperfect assessor is calibrated for these same two pre-specified categories that partition the population.

**Example 2.** Let $N = \{1, 2, 3\}$, and let $Y(i) = 1$ for $i = 1, 3$. Having property $y$ is represented by an asterisk in Table 2. Supposing that $h$ is imperfect and calibrated for the

Table 2: No Calibration for Two Binary Partitions

| | B | W |
|---|---|---|
| M | $1^*$ | 2 |
| F | $3^*$ | |

$\{M, F\}$ partition of $N$ implies that individual 1 must receive a score in $(0, 1)$. The only such assessment consistent with calibration is $h(1) = h(2) = 1/2$. But then $h$ cannot calibrated for $B$ since, by calibration for $F$, $h(3) = 1$. Similarly, $h$ cannot be calibrated for $W$ since $P_W(Y = 1|h = 1/2) \neq 1/2$. ◇

One feature of Example 2 that may incline some to regard it as a corner case is that the base rate for property $y$ is extreme (in $\{0, 1\}$) in the $\{B, W\}$ partition. But that is only for simplicity of illustration. The Appendix includes a simple example (Example 4) of a population partitioned three ways. The base rate of $y$ in each cell is in $(0, 1)$. No imperfect, calibrated assessor exists. So impossibilities emerge even in the simplest cases.

Still, it makes sense to investigate conditions that are both necessary and sufficient for the existence of an imperfect but calibrated assessor, for example, for all partitions in a set $\Pi$. (Again, in general, $\Pi$ may not be the set of *all* partitions of $N$ or all partitions of $N$ of a given cardinality.) We could investigate analogous characterizing conditions for predictive

equity, equalized odds, or other fairness constraints of concern. I will have to leave the task of discovering conditions that are both intuitive and interesting to future work. However, we can already mark a few limitations of this approach. First, Observations 1, 2, and 3 still imply that, for any assessment problem and any non-trivial assessor, bias against *some* groups will persist—even if that group is not considered highest priority. Second, by Examples 2 and 4, unbiased assessment for even two or three given partitions is sometimes impossible outside of trivial cases. Even if these partitions represent the categories typically considered of most urgent concern in some context—like race and gender in certain settings, say—fairness as gauged by the metrics under discussion may be impossible to achieve in a non-trivial way for these partitions.[17] Moreover, even if they are achievable for those partitions of one population (like for race and gender partitions in Example 1), they might fail to be achievable in another population for those same categories (like for race and gender partitions in Example 2). For example, it could be the case that a non-perfect, calibrated risk assessor exists for races and genders for defendants one year, but not the next. And since we do not typically know the values of either $Y$ or $P_k$ at the time assessments must be made, we typically do not know whether non-trivial unbiased assessment is possible for all partitions in the set. Specifying a set of relevant partitions $\Pi$ for a given assessment problem, then, confers no guarantee that a given fairness constraint can be satisfied for all members of $\Pi$ in a non-trivial fashion.

To summarize, on the one hand, requiring the satisfaction of a fairness constraint for some *single* partition is generally unsatisfactory since we may care about the fair treatment of groups from different partitions. On the other hand, requiring any of the fairness constraints considered here be satisfied for *all* partitions of the population or all partitions of some cardinality places unrealistically high demands on assessment as Observations 1, 2, and 3 establish. Kearns et al. make a similar point: "we cannot insist on any notion of statistical fairness for *every* subgroup of the population: for example, any imperfect classifier could be accused of being unfair to the subgroup of individuals defined ex-post as the set of individuals it misclassified. This simply corresponds to 'overfitting' a fairness constraint" (2018, p. 2565). The foregoing observations refine this point, providing, for each fairness constraint, an explicit characterization of when the constraint holds or holds approximately for all partitions. What about imposing the fairness constraint on only some set of partitions? There are at least three problems with resisting the limitative nature of Observations 1, 2, or 3 by relaxing the assumption that the relevant fairness constraint holds for all partitions (or all partitions of a given cardinality) to the assumption that it holds for just multiple partitions. First, by the observations above, bias against *some* group is a foregone conclusion. Which and how many partitions are ethically relevant is not invariant across assessment problems and cannot be decided a priori, so the implied bias may be more or less ethically relevant. As Examples 2 and 4 show, for some assessment problems, we can run into impossibilities even for small sets of partitions. These examples can be adapted to show that this sort of limitation emerges even for approximate versions of the fairness constraints. Second, for some given set of partitioning

[17] One might hold that the limitative results discussed above are grounds to seek other formalizations of unbiased assessments, other fairness constraints. One thing that underwrites such a view, I guess, is optimism that fair treatment of all groups should be possible. There are also the reservations, mentioned earlier, about defining fairness in terms of certain types of statistical parities (see, e.g., (Bright et al., 2016) for *causal* interpretations of intersectionality, a topic I turn to next). For readers sympathetic to that sort of view, this essay is best understood as an investigation of the consequences and limitations of current proposals for capturing unbiased assessment.

categories, certain populations may admit non-trivial fair assessments while others do not. For example, it could be the case that the population of Broward County defendants in 2015 admits non-trivial fair assessment for races and genders, while the 2016 population does not. This raises yet further concerns about fair treatment: some populations can be treated fairly, while others cannot unless the assessment meets the highest bar of perfection. The third problem that arises for resisting the limitative nature of the foregoing observations by focusing only on a set of pre-determined partitions is the prospect of intersectional bias, which I turn to next.

## 5 Intersectionality

We have seen that membership in multiple social groups is universal and is, in a sense, a truism. Intersectionality theory is concerned with membership in multiple *socially disadvantaged* groups, and with how membership in multiple socially disadvantaged groups can compound disadvantage in a nonlinear way, so to speak. Calibration, predictive equity, and equalized odds each presents a particular conception of unbiased assessment. A natural question to ask is whether any of these conceptions of unbiased assessments admits the possibility of intersectional bias. Do the intersectionality theorist's concerns emerge here?

Kimberlé Crenshaw, who introduced the term "intersectionality," makes use of a court case to explain how bias against black women, for example, is consistent with the lack of that form of bias against black people or against women (Crenshaw, 1989). In *DeGraffenreid v. General Motors*, five black women alleging discrimination by General Motor's seniority-based system sued the company. Prior to 1964, General Motors did not hire black women. All of the black women hired after 1970 lost their jobs through a seniority-based layoff during a later recession. The district court rejected the plaintiffs' attempt to bring a suit on behalf of black women in particular rather than on behalf of black people or women. According to the court, the suit must present "a cause of action for race discrimination, sex discrimination, or alternatively either, but not a combination of both" (qtd. in Crenshaw, 1989, p. 141). The court noted that, while General Motors did not hire black women prior to 1964, they did hire female employees for a number of years prior to 1964. So there was no sex discrimination. And what if General Motors had hired black people—specifically black men—for a number of years prior to 1964? Crenshaw's point is that that would not really absolve General Motors of the charge of discrimination against black women. It certainly does not follow that there could be no discrimination against black women.

To address the spectre of intersectional bias in the fair assessment setting considered here, we need to introduce the notion of the coarsest common refinement of a set of partitions. As technical work that deals with partitions has shown, the coarsest common refinement of a set of partitions is a very handy concept (e.g., Aumann, 1976).[18] A partition $\pi$ *refines* (is a refinement of) another partition $\pi'$ if every cell of $\pi$ is a subset of a cell of $\pi'$. In turn, $\pi'$ *coarsens* or is a coarsening of $\pi$. The partition $\pi$ is a *common refinement* of partitions $\pi'$ and $\pi''$ if $\pi$ refines both $\pi'$ and $\pi''$. The *coarsest common refinement* of partitions $\pi'$ and $\pi''$ is the partition $\pi$ that is a common refinement of $\pi'$ and $\pi''$ such that any other common refinement

[18]Thanks to Ignacio Ojea Quintana for urging that I look into this issue. Since drafting this paper, I learned of a more recent and intentional spin on this sort of intersectional bias that goes by the label "fairness gerrymandering" in the machine learning literature (e.g., Kearns et al., 2018).

of $\pi'$ and $\pi''$ is also a refinement of $\pi$. The coarsest common refinement of a set of partitions is given by the partition consisting of the nonempty *intersections* of the cells of those partitions. In the aforementioned court case, for example, from a race partition containing the categories of black people and white people and a gender partition containing the categories of women and men, we form a new partition with the categories of black women, black men, white women, white men.

How does the fairness framework under consideration in this essay bear on the issue of intersectionality? It is not true that if $h$ is calibrated for each group in some set of partitions, then $h$ is calibrated for the coarsest common refinement of those partitions. For example, even if an assessor exhibits no bias against black job candidates nor against women candidates in the sense that it is calibrated for these groups, it can still fail to be calibrated for black women. That bias not present against certain groups can be present against intersections of those groups is easily established in the assessment framework under consideration.

**Example 3.** Let $N = \{1, 2, 3, 4, 5, 6, 7\}$, and let $Y(i) = 1$ for $i = 2, 4, 5, 7$. Again, having property $y$ is represented by an asterisk in Table 3. Consider two binary partitions, one for gender, $\{M, F\}$, and one for race, $\{B, W\}$. The partitions and assessments scores are also displayed in Table 3.

Table 3: Intersectional Bias

| | B | W |
|---|---|---|
| M | $h(1) = 2/3,\ h(2^*) = 2/3$ | $h(3) = 0,\ h(4^*) = 2/3$ |
| F | $h(5^*) = 2/3$ | $h(6) = 2/3,\ h(7^*) = 2/3$ |

The assessor $h$ is calibrated for both the $\{M, F\}$ partition and the $\{B, W\}$ partition. In all of those groups, two thirds of those who receive an assessment of $2/3$ have property $y$. The coarsest common refinment is the four-cell partition $\{B \cap M, B \cap F, W \cap M, W \cap F\}$ composed of the groups of black men, black women, white men, and white women. Since $P_{B\cap F}(Y = 1|h = 2/3) = 1$, $h$ is underconfident in (and so not calibrated for) black women. At the same time, $h$ is overconfident in both black men and white women. $\diamond$

Two further points are worth emphasizing. First, the example also shows that it is conceptually possible for membership in multiple socially disadvantaged groups to lead to favorable assessment bias. This could be the case if $y$ is an undesirable feature like recidivism. Or, if $y$ is a desirable feature, simply swap the $M$ and $F$ labels to generate a case of favorable bias for black women. Second, since calibration implies predictive equity, it follows that $h$ satisfies predictive equity for the race and gender partitions in Example 3. But $h$ also violates predictive equity for the coarsest common refinement: $P_{B\cap F}(Y = 1|h = 2/3) = 1 \neq P_{W\cap F}(Y = 1|h = 2/3) = 1/2$, for example. In the Appendix, Example 5 establishes that equalized odds can also be violated for the coarsest common refinement of two partitions even when the constraint is satisfied for these coarser partitions. So intersectional bias is possible for all of the fairness constraints under considerations. I summarize these points in the following observation.

**Observation 6.** *Let $h$ be an assessor for $N$.*

1. *Even if $h$ is calibrated for each partition in a set $\Pi$ of partitions of $N$, $h$ can fail to be calibrated for the coarsest common refinement of $\Pi$.*

2. *Even if $h$ satisfies predictive equity for each partition in a set $\Pi$ of partitions of $N$, $h$ can fail to satisfy predictive equity for the coarsest common refinement of $\Pi$.*

3. *Even if $h$ satisfies equalized odds for each partition in a set $\Pi$ of partitions of $N$, $h$ can fail to satisfy equalized odds for the coarsest common refinement of $\Pi$.*

I end on at least a slightly more positive note. Suppose that we manage to specify some number of population partitions for which unbiased assessment is most ethically relevant in a particular assessment problem. The next observation provides one consideration in favor of focusing on unbiased assessment for the coarsest common refinement of the relevant partitions.

**Observation 7.** *Let $h$ be an assessor for $N$.*

1. *If $h$ is calibrated for the coarsest common refinement of a set $\Pi$ of partitions of $N$, then $h$ is calibrated for each partition in $\Pi$.*

2. *If $h$ satisfies predictive equity for the coarsest common refinement of a set $\Pi$ of partitions of $N$, then $h$ satisfies predictive equity for each partition in $\Pi$.*

3. *If $h$ satisfies equalized odds for the coarsest common refinement of a set $\Pi$ of partitions of $N$, then $h$ satisfies equalized odds for each partition in $\Pi$.*

Observation 7 assures us that, if we can specify the population partitions that are ethically relevant in an assessment problem, then satisfying a particular fairness constraint for the coarsest common refinement implies that the constraint is satisfied on all of the relevant partitions. The constraint only needs to be satisfiable in the coarsest common refinement in a non-trivial way. This contrasts with the lack of a corresponding guarantee, indicated in Observation 6, when we focus on unbiased assessments for the coarser partitions as the district court did in *DeGraffenreid v. General Motors*. Nevertheless, satisfaction of one of these fairness constraints on the coarsest common refinement of a set of partitions is only a sufficient condition for the constraint's satisfaction for all partitions in the set; it is not necessary, as simple examples illustrate. As a result, focusing only on the coarsest common refinement unduly restricts the set of fair assessors. Furthermore, we generally do not know at the time of assessment whether non-trivial unbiased assessment is possible for the coarsest common refinement.

## 6 Conclusion

There are multiple ways to carve a population, multiple social identities, for which it may be important to avoid biased assessments. Fixing a single partition of identities is overly restrictive, committing us to ignoring both relevant forms of bias against other groups and changing social context. Allowing even a set of partitions to ossify into *the* relevant partitions

may fail to make us sufficiently attentive. In the *DeGriffenreid v. General Motors* decision, the court resisted the idea that relevant classifications are open to reconsideration or refinement: "The prospect of the creation of new classes of protected minorities, governed only by the mathematical principles of permutation and combination, clearly raises the prospect of opening the hackneyed Pandora's box" (qtd. in Crenshaw, 1989, p. 142). Pandora's box or not, the alternative seems to be refusal to confront different possible forms of group bias. If Sen is right, such dogmatism also "makes the world much more flammable."

But we confront limitations in taking the relevant classes in fair assessment to be "governed only by the mathematical principles of permutation and combination." Requiring that certain forms of bias be avoided for *all* possible social groupings is overly constraining, placing unrealistic demands on assessment as Observations 1, 2, and 3 attest. Put another way, bias against some group is inevitable for non-trivial assessment problems. Requiring only that an assessor be "close" to bias-free for all groups does not change this picture much (Observations 1′, 2′, 3′). For example, an assessor is $\varepsilon$-approximately calibrated for all groups if and only if it is $\varepsilon$-approximately perfect. Specifying a set of relevant partitions in advance, non-trivial, unbiased assessment may still be impossible for all partitions in the set as is the case in Examples 2 and 4. Moreover, even if assessments satisfy one of the fairness properties for some set of partitions, it does not follow that the property is satisfied for the coarsest common refinement (Observation 6), establishing the possibility of forms of intersectional bias for these fairness constraints. However, if non-trivial unbiased assessment *is* possible for the coarsest common refinement of the partitions in the set, then, by Observation 7, assessments will be unbiased for the partitions in the set. But even this may be too stringent since unbiased assessment is sometimes possible for each partition in a set even if imperfect assessments *must* be biased for the coarsest common refinement of those partitions.

Where does this leave us? What the foregoing analysis helps us to make clear is that, not only is there a conflict between eliminating different forms of bias, but there are serious limits to the extent to which a given form of bias can be eliminated across different partitions. Often, many partitions are important. In "auditing" an assessor for bias (e.g., Kearns et al., 2018), outside of some rather restrictive cases, we are guaranteed to find bias against *some* group. While that is pertinent data for the ethical *evaluation* of an assessor, what are the *prescriptive* implications for assessment? Some have recently argued for rejecting nearly all of the statistical criteria of fairness proposed in the literature for reasons of a different nature than those I have considered (e.g., Hedden, 2021). Do the observations recorded here add to this case? It seems there are two broad approaches we might pursue. First, we could reconcile ourselves with bias against some groups since it is essentially inevitable on this way of understanding fair assessment, hoping and doing what we can to ensure that implied forms of bias minimally impact what we take to be the most relevant protected classes for the given time and place (cf. Hébert-Johnson et al., 2018). Second, we could seek a different conception of fair assessment.

# Appendix

## Example of Three Partitions for which the Only Calibrated Assessor Is Perfect

**Example 4.** Let $N = \{1, 2, 3, 4, 5, 6\}$ and let $Y(i) = 1$ for $i = 1, 2, 6$. Possession of property $y$ is represented by an asterisk in both tables 4 and 5. Consider three partitions of $N$: one for gender $\{M, F\}$, one for race $\{B, L, W\}$, and one for sexual orientation $\{G, S\}$. The three partitions are represented in tables 4 and 5.

Table 4: Gender and Race Partitions

| | B | L | W |
|---|---|---|---|
| M | $1^*$ | $2^*$ | 3 |
| F | 4 | 5 | $6^*$ |

Table 5: Sexual Orientation Partition

| | | | | |
|---|---|---|---|---|
| G | $1^*$ | 5 | | |
| S | $2^*$ | 3 | 4 | $6^*$ |

Suppose that there is some imperfect assessor $h$ that is calibrated simultaneously for both $\{M, W\}$ and $\{B, L, W\}$. Note first that some individual $i \in N$ must receive a score in $(0, 1)$. (Otherwise, fallibility would imply that some individual with property $y$ receives an assessment of 0 or some individual lacking it receives an assessment of 1. Either results in a failure of calibration whichever individual and partition we use.) Consider the gender partition. Suppose that the witness to the fallibility of $h$ is in $M$. (I will check one case, leaving the others to the reader.) Calibration only allows for two possible witnessing assessment scores: $\{1/2, 2/3\}$. If the witnessing score is $2/3$, then $h(i) = 2/3$ for $i = 1, 2, 3$. But then, $h$ is not calibrated for any cell of the $\{B, L, W\}$ partition. If the witnessing score is $1/2$, then calibration implies $h(3) = 1/2$. Since $h$ is calibrated for $W$, $h(6) = 1/2$. Either $h(1) = 1/2$ or $h(2) = 1/2$. If the former, then $h(2) = 1$. Calibration for $\{B, L, W\}$ implies that $h(4) = 1/2$. Now, in group $S$, individuals 3, 4, and 6 all have score $1/2$, but only $1/3$ of them have property $y$. So, $h$ is not calibrated for $S$. The other cases can be similarly verified. $\diamond$

## Proof of Observation 4

*Proof.* (1). Suppose that $h$ makes perfect distinctions: for all $i, j \in N$, $Y(i) \neq Y(j)$ implies $h(i) \neq h(j)$. To define $g$, for all $r \in [0, 1]$, set

$$g(r) = \begin{cases} 1, & \text{if } Y(i) = 1 \text{ for some } i \text{ such that } h(i) = r; \\ 0, & \text{otherwise.} \end{cases}$$

If $g(h(i)) = 1$, then for some $j \in \{j : h(j) = h(i)\}$, $Y(j) = 1$. By perfect distinctions, $Y(i) = Y(j) = 1$. Similarly, if $g(h(i)) = 0$, then $Y(i) = 0$. So, $g \circ h$ is perfect.

Now suppose that $h$ does not make perfect distinctions. Then, for some $i, j \in N$ such that $Y(i) \neq Y(j)$, $h(i) = h(j)$. There cannot exist a function $g : [0,1] \to \{0,1\}$ such that $g \circ h$ is perfect since $g$ must map $h(i)$ and hence $h(j)$ to 0 or 1. This will misclassify either $i$ or $j$. Hence, if $h$ does not make perfect distinctions, then there exists no function $g : [0,1] \to \{0,1\}$ such that $g \circ h$ is perfect.

(2). Suppose that $h$ is non-constant and satisfies perfect non-discrimination. Then, by perfect non-discrimination, $h$ assigns all individuals with $y$ the same score $p$, and all individuals without $y$ the same score $p'$. Since $h$ is non-constant, $p \neq p'$, so the population cannot be homogeneous. Defining $g$ as before, $g \circ h$ is perfect and injective since $g(h(i)) = g(h(j)) = 1$ implies that $i$ and $j$ are both in the set of individuals individuals with $y$ and hence $h(i) = h(j) = p$, and $g(h(i)) = g(h(j)) = 0$ implies that $i$ and $j$ are both in the set individuals without $y$ and hence $h(i) = h(j) = p'$.

Now suppose that the population is not homogeneous and there exists an injective function $g : [0,1] \to \{0,1\}$ such that $g \circ h$ is perfect. Since the population is not homogeneous, $g \circ h$ takes both values 0 and 1. Then, since $g \circ h$ is injective, $h$ takes only two values. Now, since $g \circ h$ is perfect, $h$ must satisfy perfect non-discrimination. □

## Proof of Observation 1′

*Proof.* (2) ⇒ (1). Trivial.

(1) ⇒ (3). Suppose that, for some $\varepsilon \geq 0$, $h$ is approximately calibrated for all binary partitions. Then, in particular, for each $i \in N$, $h$ is approximately calibrated for the partition $\{\{i\}, N \setminus \{i\}\}$. It follows that,

$$|P_{\{i\}}(Y = 1|h = p) - p| \leq \varepsilon. \tag{1}$$

Since $\{i\}$ is a singleton, the conditional probability $P_{\{i\}}(Y = 1|h = p)$ is defined only for the one $p$ such that $h(i) = p$. Now, for the defined conditional probabilities, if $Y(i) = 1$, then $P_{\{i\}}(Y = 1|h = p) = 1$; and if $Y(i) = 0$, then $P_{\{i\}}(Y = 1|h = p) = 0$. So, for all $i \in N$, we have $Y(i) = P_{\{i\}}(Y = 1|h = p)$ whenever the right hand side is defined. Substituting in 1,

$$|Y(i) - h(i)| \leq \varepsilon. \tag{2}$$

Since 2 holds for each $i$, it follows that $h$ is $\varepsilon$-approximately perfect.

(3) ⇒ (2). Suppose that $h$ is approximately perfect for some $\varepsilon \geq 0$: for all $i \in N$, 2 holds. Let $G$ be an arbitrary, nonempty subset of $N$. Let $p$ be any value such that the conditional probability $P_G(Y = 1|h = p)$ is defined. Consider first the case in which $p \leq 0.5$. Let $G_+ = \{i \in N : Y(i) = 1\}$. Either $P_{G_+}(\cdot|h = p)$ is defined or not. If it is defined, then $P_{G_+}(Y = 1|h = p) = 1$ since $Y(i) = 1$ for all $i \in G_+$. So, for any $i \in G$ such that $Y(i) = 1$ and $h(i) = p$, we have

$$|P_G(Y = 1|h = p) - p| \leq |P_{G_+}(Y = 1|h = p) - p| = |Y(i) - h(i)| \leq \varepsilon. \tag{3}$$

If $P_{G_+}(\cdot|h = p)$ is not defined, then $h(i) = p$ for no $i \in N$ such that $Y(i) = 1$. Then, for any $i \in G$ such that $h(i) = p$,

$$|P_G(Y = 1|h = p) - p| = |Y(i) - h(i)| \leq \varepsilon. \tag{4}$$

By 3 and 4, whether $P_{G_+}(\cdot|h=p)$ is defined or not, we have $|P_G(Y=1|h=p)-p| \leq \varepsilon$. When $p > 0.5$, an analogous argument using $G_- = \{i \in N : Y(i) = 0\}$ instead of $G_+$ establishes that $|P_G(Y=1|h=p)-p| \leq \varepsilon$ in that case, too. It follows that $|P_G(Y=1|h=p)-p| \leq \varepsilon$ for any $p \in [0,1]$ such that $P_G(\cdot|h=p)$ is defined. Since any partition $\pi$ of $N$ consists of such groups $G$, we conclude that $h$ is $\varepsilon$-approximately calibrated for any partition $\pi$.

□

## Proof of Observation 2′

*Proof.* (2) ⇒ (1). Trivial.

(1) ⇒ (3). Let $h$ satisfy $\varepsilon$-approximate predictive equity for all binary partitions with $0 \leq \varepsilon < 1$. Suppose that, for $i, j \in N$, $Y(i) \neq Y(j)$ but $h(i) = h(j) = p$. Then, in particular, $h$ satisfies $\varepsilon$-approximate predictive equity for the partition $\pi = \{G_+, G_-\}$, where $G_+ = \{i \in N : Y(i) = 1\}$, and both of the conditional probabilities in 5 are defined:

$$|P_{G_+}(Y=1|h=p) - P_{G_-}(Y=1|h=p)| \leq \varepsilon < 1. \tag{5}$$

But 5 implies that

$$|1-0| \leq \varepsilon < 1,$$

which is a contradiction. Hence, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. It follows that approximate predictive equity with $\varepsilon \in [0,1)$ implies perfect distinction.

(3) ⇒ (2). Suppose that $h$ satisfies perfect distinction: for all $i, j \in N$, $Y(i) \neq Y(j)$ implies that $h(i) \neq h(j)$. Let $\pi = \{G_1, \ldots, G_m\}$ be an arbitrary partition of $N$. Since for any individuals $i, j \in N$ such that $h(i) = h(j) = p$ we have $Y(i) = Y(j)$, it follows that $P_{G_k}(Y=1|h=p) = P_{G_l}(Y=1|h=p)$—which is equal to 1 or 0 according to whether $Y(i) = Y(j) = 1$ or $Y(i) = Y(j) = 0$—for any two groups $G_k, G_l \in \pi$ whenever those two conditional probabilities are defined. That is, $h$ satisfies approximate predictive equity for $\pi$ with $\varepsilon = 0$. Since $\pi$ was arbitrary, $h$ satisfies $\varepsilon$-approximate predictive equity for all partitions with $0 \leq \varepsilon < 1$. □

## Proof of Observation 3′

*Proof.* (1 ⇒). Let $h$ satisfy $\varepsilon$-approximately equalized odds for all partitions. Consider the partition of singleton cells. For any two $i, j \in N$ such that $Y(i) = Y(j) = 0$, $|f_i^+(h) - f_j^+(h)| \leq \varepsilon$. In particular, since $\max_{i \in N} f_i^+(h) = \max_{i \in \{Y=0\}} h(i)$ and $\min_{i \in N} f_{\{i\}}^+(h) = \min_{i \in \{Y=0\}} h(i)$, it follows that $\max_{i \in \{Y=0\}} h(i) - \min_{i \in \{Y=0\}} h(i) \leq \varepsilon$. A similar argument applies if $Y(i) = Y(j) = 1$. Hence, for all $i, j \in N$, if $Y(i) = Y(j)$, then $|h(i) - h(j)| \leq \varepsilon$.

(1 ⇐). Suppose that $h$ is $\varepsilon$-approximately perfectly non-discriminatory: $Y(i) = Y(j)$ implies that $|h(i) - h(j)| \leq \varepsilon$. For any cells $G_k, G_l$ of any partition $\pi$ of $N$ such that the conditional expectations are defined, $f_k^+(h), f_l^+(h) \in [\min_{i \in \{Y=0\}} h(i), \max_{i \in \{Y=0\}} h(i)]$. Hence, $|f_k^+(h) - f_l^+(h)| \leq |\max_{i \in \{Y=0\}} h(i) - \min_{i \in \{Y=0\}} h(i)| \leq \varepsilon$. An analogous argument works for the false negative rates for $\pi$. Since $\pi$ was arbitrary, it follows that $h$ satisfies approximately equalized odds for all partitions.

(2). Let $h$ satisfy $\varepsilon$-approximately equalized odds for all binary partitions with $\varepsilon \geq 0$. Define $G_-^{\min} = \{i \in N : h(i) = min_{i \in \{Y=0\}} h(i)\}$, $G_-^{\max} = \{i \in N : h(i) = \max_{i \in \{Y=0\}} h(i)\}$.

By the assumption,

$$\left|f^+_{G^{\min}_-}(h) - f^+_{N\setminus G^{\min}_-}(h)\right| \leq \varepsilon \tag{6}$$

$$\left|f^+_{G^{\max}_-}(h) - f^+_{N\setminus G^{\max}_-}(h)\right| \leq \varepsilon \tag{7}$$

(If any of the terms in 6 or 7 are undefined, then either all individuals in the population have $y$ or there is a single value $p$ such that, for any $i$ with $Y(i) = 0$, $h(i) = p$. In neither case is it possible to get a violation of $2\varepsilon$-approximate perfect non-discrimination for any binary partition.) Using the triangle inequality, 6, and 7, we have, for any $i, j \in N$ such that $Y(i) = Y(j) = 0$,

$$\begin{aligned}
|h(i) - h(j)| &\leq \left|f^+_{G^{\max}_-}(h) - f^+_{G^{\min}_-}(h)\right| \\
&\leq \left|f^+_{G^{\max}_-}(h) - f^+_{N\setminus(G^{\min}_- \cup\, G^{\max}_-)}(h)\right| + \left|f^+_{N\setminus(G^{\min}_- \cup\, G^{\max}_-)}(h) - f^+_{G^{\min}_-}(h)\right| \\
&\leq \left|f^+_{G^{\max}_-}(h) - f^+_{N\setminus G^{\max}_-}(h)\right| + \left|f^+_{N\setminus G^{\min}_-}(h) - f^+_{G^{\min}_-}(h)\right| \\
&\leq \varepsilon + \varepsilon.
\end{aligned}$$

Hence, for any two $i, j \in N$ such that $Y(i) = Y(j) = 0$, $|h(i) - h(j)| \leq 2\varepsilon$. An analogous argument establishes that for any two $i, j \in N$ such that $Y(i) = Y(j) = 1$, $|h(i) - h(j)| \leq 2\varepsilon$. □

## Example of a Violation of Equalized Odds for the Coarsest Common Refinement of Two Partitions

**Example 5.** Let $N = \{1, 2, 3, 4, 5, 6, 7, 8\}$, and let $Y(i) = 1$ for $i = 1, 3, 5, 7$. Again, having property $y$ is represented by an asterisk in Table 6. Consider two binary partitions, one for gender, $\{M, F\}$, and one for race, $\{B, W\}$. The partitions and assessments scores are also displayed in Table 6.

Table 6: Intersectional Bias

| | B | W |
|---|---|---|
| M | $h(1^*) = 1,\ h(2) = 0$ | $h(3^*) = 1/2,\ h(4) = 1/2$ |
| F | $h(5^*) = 1/2,\ h(6) = 1/2$ | $h(7^*) = 1,\ h(8) = 0$ |

The assessor $h$ satisfies equalized odds for both the $\{M, F\}$ partition and the $\{B, W\}$ partition: $f^+_M(h) = f^+_F(h) = 1/4$, $f^-_M(h) = f^-_F(h) = 1/4$, $f^+_B(h) = f^+_W(h) = 1/4$, and $f^-_B(h) = f^-_W(h) = 1/4$. The coarsest common refinement is the four-cell partition $\{B \cap M, B \cap F, W \cap M, W \cap F\}$ composed of the groups of black men, black women, white men, and white women. To see that $h$ does not satisfy equalized odds for this partition, it suffices to observe that $f^+_{B\cap F}(h) = 1/2$ while $f^+_{W\cap F}(h) = 0$ and $f^-_{B\cap F}(h) = 1/2$ while $f^-_{W\cap F}(h) = 0$. The assessor is perfect, in fact, for black men and white women, and has positive error rates for black women and white men. ◇

## Proof of Observation 7

*Proof.* (1). Let $\Pi$ be a set of partitions of $N$ and $\pi = \{G_1, \ldots, G_m\}$ the coarsest common refinement of the partitions in $\Pi$. Assume that $h$ is calibrated for $\pi$. We want to show that for any $\pi' \in \Pi$, $h$ is calibrated for $\pi'$. Any cell $G'$ of $\pi'$ is a union of some number of cells of $\pi$: $G' = \bigcup_k G_k$ where the $G_k$ are cells of $\pi$. Since, for each such $G_k$, $P_k(Y = 0|h = p) = p$, by the law of total probability

$$\begin{aligned} P_{G'}(Y = 0|h = p) &= \sum_k P_{G'}(Y = 0|h = p, G_k) P_{G'}(G_k|h = p) \\ &= \sum_k p P_{G'}(G_k|h = p) \\ &= p \sum_k P_{G'}(G_k|h = p) \\ &= p. \end{aligned}$$

So, for any cell $G'$ of any partition $\pi' \in \Pi$, $h$ is calibrated.

(2). Analogous to the previous proof.

(3). Let $\Pi$ be a set of partitions of $N$ and $\pi = \{G_1, \ldots, G_m\}$ the coarsest common refinement of the partitions in $\Pi$. Assume that $h$ satisfies equalized odds for $\pi$. We want to show that for any $\pi' \in \Pi$, $h$ satisfies equalized odds for $\pi'$. Any cell $G'$ of $\pi'$ is a union of some number of cells of $\pi$: $G' = \bigcup_k G_k$. Since the $G_k$ partition $G'$ and there is some $r \in [0, 1]$ such that, for each $G_k \in \pi$, $E_k(h|Y = 0) = r$ (when that conditional expectation is defined), by the law of total expectation,

$$\begin{aligned} E_{G'}(h|Y = 0) &= \sum_k E_{G'}(h|Y = 0, G_k) P_{G'}(G_k|Y = 0) \\ &= \sum_k r P_{G'}(G_k|Y = 0) \\ &= r \sum_k P_{G'}(G_k|Y = 0) \\ &= r. \end{aligned}$$

Hence, $f^+_{G'}(h) = f^+_{G''}(h) = r$ for all cells $G', G'' \in \pi'$ for which those error rates are defined. (If defined for no $G_k$, we need only check the false negative rates.) By the same line of reasoning, *mutatis mutandis*, $f^-_{G'}(h) = f^-_{G''}(h)$ for all $G', G'' \in \pi'$. Thus, for any $\pi' \in \Pi$, $h$ satisfies equalized odds. □

## References


Angwin, J. and J. Larson (2016, December). Bias in criminal risk scores is mathematically inevitable, researchers say. https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016a, May). How we analized the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016b). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Aumann, R. J. (1976). Agreeing to disagree. *The Annals of Statistics 4*(6), 1236–1239.

Barocas, S., M. Hardt, and A. Narayanan (2019). *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

Borsboom, D., J.-W. Romeijn, and J. M. Wicherts (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods 13*(2), 75–98.

Bright, L. K., D. Malinsky, and M. Thompson (2016). Causally interpreting intersectionality theory. *Philosophy of Science 83*(1), 60–81.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Huq (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806. ACM.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum 1989*(Article 8), 139–167.

Dwoskin, E. (2018, August). Facebook is rating the trustworthiness of its users on a scale from zero to 1. https://www.washingtonpost.com/technology/2018/08/21/facebook-is-rating-trustworthiness-its-users-scale-zero-one/?noredirect=on&utm_term=.aaa0972fc65f.

Facebook (2022, April). Discriminatory practices. https://www.facebook.com/policies/ads/prohibited_content/discriminatory_practices.

Fang, H. and A. Moro (2011). Theories of statistical discrimination and affirmative action: A survey. In J. Benhabib, A. Bisin, and M. O. Jackson (Eds.), *Handbook of Social Economics*, Volume 1, pp. 133–200. Elsevier.

Hébert-Johnson, U., M. Kim, O. Reingold, and G. Rothblum (2018). Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, pp. 1939–1948. PMLR.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, Forthcoming.

Kearns, M., S. Neel, A. Roth, and Z. S. Wu (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, Stockholm, Sweden, pp. 2564–2572. PMLR.

Kleinberg, J., S. Mullainathan, and M. Raghavan (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, Dagstuhl, Germany, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Mitchell, S., E. Potash, S. Barocas, A. D'Amour, and K. Lum (2018). Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

Patty, J. and E. M. Penn (MS). Algorithmic fairness and statistical discrimination. Unpublished Manuscript.

Pleiss, G., M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pp. 5680–5689.

Sen, A. (2007). *Identity and Violence: The Illusion of Destiny*. London: Penguin Books.

Shimony, A. (1988). An adamite derivation of the principles of the calculus of probability. In *Probability and Causality*, pp. 79–89. Springer.

Stewart, R. T. and M. Nielsen (2020). On the possibility of testimonial justice. *Australasian Journal of Philosophy 98*(4), 732–746.

Tetlock, P. E. (2017, originally published in 2005). *Expert Political Judgment: How Good Is It? How Can We Know? (New Edition)*. Princeton University Press.

van Fraassen, B. C. (1983). Calibration: A frequency justification for personal probability. In R. Cohen and L. Laudan (Eds.), *Physics, Philosophy, and Psychoanalysis*, pp. 295–319. Springer.