Forthcoming in *Philosophy & Technology*

New Possibilities for Fair Algorithms

Michael Nielsen[1] and Rush T. Stewart[2]

[1]The University of Sydney
[2]University of Rochester

September 20, 2024

Abstract

We introduce a fairness criterion that we call Spanning. Spanning i) is implied by Calibration, ii) retains interesting properties of Calibration that some other ways of relaxing that criterion do not, and iii) unlike Calibration and other prominent ways of weakening it, is consistent with Equalized Odds outside of trivial cases.

Keywords. Algorithmic fairness; base rate tracking; calibration; equalized odds

1 Introduction

The genre of much work on algorithmic fairness is tragedy. After the identification of some set of seemingly desirable fairness criteria comes the statement of an impossibility theorem establishing that those criteria are inconsistent or consistent in only utterly unrealistic or trivial cases (Kleinberg et al., 2017; Pleiss et al., 2017; Chouldechova, 2017; Stewart and Nielsen, 2020; Beigang, 2023b). A central example is the result due to Kleinberg and coauthors establishing that two constraints called Calibration and Equalized Odds are inconsistent outside of certain trivial cases (2017). One natural response is to weaken Equalized Odds. Pleiss et al. show that, for a particular way of relaxing Equalized Odds, new possibilities emerge (2017). Ways of weakening Calibration have also been investigated but have led to more impossibility results (Stewart and Nielsen, 2020; Stewart et al., 2024).

We find the relative merits of Calibration and Equalized Odds difficult to evaluate. Consequently, we regard explorations of relaxing each criterion in attempts to skirt the impossibility result as worthwhile. For the present investigation, we will suppose that Equalized Odds is a necessary condition of algorithmic fairness. Given this assumption, we ask what interesting content of Calibration can be retained without running into triviality. Our genre is not tragedy. We identify a way of weakening Calibration that retains some of its interesting properties, but is non-trivially consistent with Equalized Odds. We call this criterion *Spanning*. It is important to emphasize that we are not proposing Spanning as a *sufficient* condition for algorithmic fairness. On its own, it is a weak criterion. In some ways, this means that the case for its status as a necessary condition is easier to make. Conjoined with Equalized Odds, it's stronger, but there may be yet further necessary criteria. After introducing the framework and the criteria in Section 2, we provide three motivations for Spanning in Section 3. Chief among these motivations is Spanning's consistency with Equalized Odds, and, in Section 4, we show that this observation can be strengthened to consistency with an even stronger criterion. We conclude in Section 5 by responding to some potential objections.

2 Criteria of Algorithmic Fairness

We adopt the standard framework for theorizing about algorithmic fairness. We start with a population $N = \{1, ..., n\}$ of individuals. We are trying to predict whether or not each individual has a certain property y. The random variable $Y : N \to \{0, 1\}$ represents that individual i has property y if $Y(i) = 1$, and it represents that i does not have y if $Y(i) = 0$. Typically, there is a vector of features associated with each individual, and it is on the basis of these features predictions are made. Depending on the prediction task, the relevant features might include high school GPA, type of crime committed, employment status, and so on. For simplicity, we will not represent these vectors. Next, we suppose that our population N is divided into groups. The groups are represented as the cells of a partition π of N. Finally, an *assessor* is a function $h : N \to [0, 1]$ that represents predictions about whether individuals have property y. For instance, we can interpret the number $h(i)$ as the probability that an algorithm assigns to the proposition that individual i has property y. Below we will say that an assessor is *binary* if it takes only the values 0 and 1; otherwise it is *continuous*. These four components—the population N, random variable Y, group partition π, and assessor h—comprise what we will call an *assessment scenario*. Examples of assessment scenarios include predicting the risk of recidivism in the criminal justice system, forecasting loan repayment, and predicting which college applicants will eventually meet a specific criterion of success.

To formally represent talk of ratios and proportions in groups, we introduce a uniform distribution P on N. This distribution can be relativized to a group G in a given partition π by conditioning on G, $P_G(\cdot) = P(\cdot|G)$, yielding a uniform distribution on G. Define the *base rate* for a group G in given partition π of population N as $\mu_G = P_G(Y = 1)$. This is the proportion of people in G that have the property y. When defined with respect to P, the expectation $E_G(h|Y = 0)$ is the average score h assigns to individuals in G that do *not* have property y. This quantity, also denoted $f^+_G(h)$, is called the *generalized false positive rate* for group G. Similarly, the *generalized false negative rate* for group G is defined as $f^-_G(h) = E_G(1 - h|Y = 1)$ and is the average of the quantity $1 - h$ for individuals in G that *do* have property y. Together, the generalized false positive and false negative rates allow us to define the Equalized Odds fairness criterion that figures centrally in the result of Kleinberg and coauthors.

> **Equalized Odds**. For an assessor h of N and any groups G, G' in the relevant partition π of N, $f^+_G(h) = f^+_{G'}(h)$ and $f^-_G(h) = f^-_{G'}(h)$ (whenever those terms are defined).

Equalized Odds for a given partition π requires that generalized error rates are the same for all groups in π. For example, in a partition into race groups in the context of predicting recidivism, it rules out assigning a much higher average probability of recidivism to black individuals who do not reoffend than to white individuals who do not reoffend. Equalized

Odds diagnoses such inequalities in errors as unfair bias.[1] Violations of (a form of) this criterion in the criminal justice system in the US have sparked much press, reflection, and research (Angwin et al., 2016).

The other fairness criterion that figures in the theorem of Kleinberg and coauthors is Calibration.

> **Calibration**. For an assessor h of N and any group G in the relevant partition π of N, $P_G(Y = 1|h = p) = p$ for all $p \in [0, 1]$ such that $P_G(h = p) > 0$.

Calibration for a given partition π requires that, for each group $G \in \pi$, among the individuals in G assigned a score p, the proportion of individuals that actually has property y is p. Consider the context of forecasting recidivism and a partition into race groups again. If 60% of the white individuals assigned a score of 0.5 were recidivists, then h would be underconfident in its predictions of white recidivism for these individuals. If 40% of Asians assigned a score of 0.5 were recidivists, then h would be overconfident in its predictions of Asian recidivism. Calbiration diagnoses both cases as unfair bias, requiring that forecast probabilities and empirical frequencies match: 50% of the individuals assigned a score of 0.5 in *any* group in the partition are recidivists.

We will call an assessment scenario *trivial* if either h is perfect or all groups have the same base rate. What Kleinberg and coauthors show is that Calibration and Equalized Odds are consistent only in trivial assessment scenarios. As we mentioned in the introduction, several ways of weakening of Calibration have been studied in the literature, inspired in large part by trying to skirt this impossibility result. For our purposes the following two are the most relevant.[2]

> **Predictive Equity**. For an assessor h of N and any groups G, G' in a partition π of N, $P_G(Y = 1|h = p) = P_{G'}(Y = 1|h = p)$ for all $p \in [0, 1]$ such that $P_G(h = p), P_{G'}(h = p) > 0$.

Instead of requiring that the percentage of recidivists among white and Asian individuals assigned a score of 0.5 is 50%, Predictive Equity requires only that those percentages are equal across the partition. So while $P_{White}(Y = 1|h = 0.5) = P_{Asian}(Y = 1|h = 0.5) = 0.6$ is a violation of Calibration, it is not a violation of Predictive Equity. In this sense, Predictive Equity preserves a form of equal treatment that Calibration implies. On the one hand, it is possible to conceive of Calibration as adding to this equal treatment requirement conditions that are not fundamentally about fairness. For instance, the additional requirement that $P_{White}(Y = 1|h = p)$ and $P_{Asian}(Y = 1|h = p)$ are not only equal to each other but equal to precisely p can be interpreted as a sort of epistemic requirement of accuracy (Stewart and Nielsen, 2020). On the other hand, there are ways of thinking of this within-group

[1] For the sake of brevity and because less brief motivations have been given a number of times elsewhere, we restrict ourselves to brief motivations/explanations of each fairness criterion other than Spanning. For further elaboration, see, for example, (Eva, 2022; Stewart, 2022).

[2] We are not treating all candidate ways of relaxing Calibration that have been proposed. For example, a variant of Base Rate Tracking called Ratio Base Rate Tracking (Eva, 2022) has also been shown to be consistent with Equalized Odds only in trivial scenarios (Stewart et al., 2024, Proposition 2). Likewise, a more abstract and perhaps more difficult to interpret weakening of Calibration labeled $*$ has been shown to be consistent with Equalized Odds only in trivial scenarios (Stewart and Nielsen, 2020, Theorem 2).

constraint as a sort of fairness requirement. The idea is that even if $P_{White}(Y = 1|h = p)$ and $P_{Asian}(Y = 1|h = p)$ are equal, it is still unfair to be under- or overconfident in recidivism—even if the assessor is *uniformly* over- or underconfident across groups. We will return to this issue when it comes to interpreting Spanning.

Eva (2022) advances a distinct way of relaxing Calibration.

> **Base Rate Tracking**. For an assessor h of N and any groups G, G' in the relevant partition π of N, $E_G(h) - \mu_G = E_{G'}(h) - \mu_{G'}$.

Continuing with the recidivism context, according to Base Rate Tracking, if the assessor treats one group as more risky than another, it must be because that group *is* riskier. If the base rates for Hispanic and Asian groups were roughly equal, but the Hispanic group were assigned significantly higher average scores, Base Rate Tracking would diagnose that as unfair bias.

The criterion that we would like to introduce for consideration is *Spanning*.

> **Spanning**. For an assessor h of N and any group G in the relevant partition π of N, the base rate μ_G lies in the interval $[\min_{i \in G} h(i), \max_{i \in G} h(i)]$.

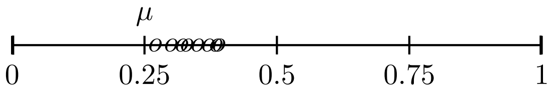
Using the same recidivism context to explain the condition, Spanning requires that the base rate for a given race group lies within the interval spanned by the predictions h makes on that group. So if the base rate for white individuals were 0.4, but the lowest score h assigns to a white individual were 0.5, h would violate Spanning. Like Calibration but unlike the other criteria we have introduced, Spanning is a within-group condition. Instead of requiring certain equalities hold across groups, it requires a particular condition to be satisfied within each group. There are at least two ways of understanding Spanning, which correspond to the two ways of understanding Calibration that we mentioned earlier. Recall that Calibration can be understood either as adding an additional accuracy requirement to Predictive Equity or as diagnosing failures of Calibration as a form of unfair bias in confidence. On the first way, Spanning may be a requirement of good predictions without being a *fairness* requirement. As far as the context of discovery goes, our inspiration is from a literature that takes an analogue of Spanning to be an *epistemic* requirement.[3] But as in the case of Calibration, it is also possible to interpret Spanning as a criterion of fairness. Consider Figure 1. The base rate for the group is 0.25, but *all* of the assessor's predictions are greater than 0.25 in clear violation of Spanning. It is natural to think that we do not need to consider the forecasts in other groups to diagnose the predictions for this group as systematically biased and unfair. If the assessor is predicting recidivism for the group, it is clear that the assessor is overconfident in the riskiness of the group. We turn now to some additional motivations for Spanning.

3 Three Motivations for Spanning

There are at least three motivations for introducing Spanning. First, like Base Rate tracking and Predictive Equity, it weakens Calibration in a natural way. In fact, these three ways of

[3] Calibration is formally similar to the reflection principle (van Fraassen, 1984). Spanning is formally similar to certain weakenings of the reflection principle (van Fraassen, 1999; Huttegger and Nielsen, 2020; Nielsen, 2021; Dorst et al., 2021; Dorst, 2023). We've repurposed "Spanning" from van Fraassen and Halpern (2016). A similar property has also been investigated in the pooling literature under the heading "bracketing the truth" (Larrick and Soll, 2006).

Figure 1: A group with base rate $\mu = 0.25$. Eight predictions, each marked by an o.



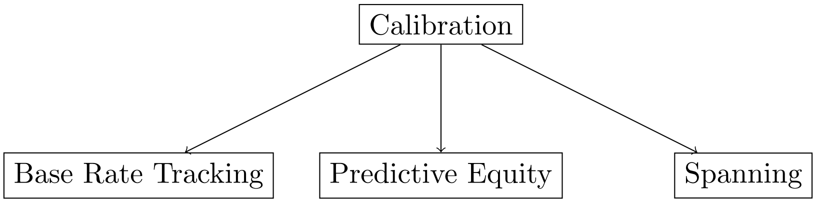


Figure 2: Three independent weakenings of Calibration

relaxing Calibration are logically independent of one another. The relationships summarized in the next proposition are depicted in Figure 2.

Proposition 1. *Calibration implies Base Rate Tracking, Predictive Equity, and Spanning, but those three conditions are independent of each other.*

The proof of this proposition and the proofs of the two propositions below are in the Appendix.

As we have mentioned, the impossibility of satisfying Calibration and Equalized Odds provides strong motivation for the program of studying ways to relax those criteria. Spanning is part of that program. One important property of Calibration that Spanning retains is ruling out the sort of systematic bias displayed in Figure 1. But Spanning retains other content as well as we explain next.

The second motivation for Spanning is that, like Base Rate Tracking but unlike Predictive Equity, Spanning retains a "commitment" to the ideal of perfect assessment (Stewart, 2024). Fairness criteria are supposed to capture some aspect of fair treatment in the context of assessment. If h satisfies some criterion like Calibration for all groups in a partition, no group has a complaint about being treated unfairly in the sense ruled out by Calibration, namely, complaints about the assessor's over- or underconfidence. But other groups, not in that partition, may. An assessor could be calibrated for a race partition but fail to be calibrated for a gender partition, for example. Maximal fairness in the sense captured by a specific criterion, while perhaps unrealistic in interesting cases, would be achieved if h were to satisfy that criterion for *all* partitions, or, equivalently, all groups.[4] In the case

[4] One might think that this is overkill. An anonymous referee expressed concern that the proof of Proposition 2 in the Appendix, for example, appeals to the partition of singletons. Instead of investigating a criterion's satisfaction for *all* partitions, perhaps we should content ourselves with a notion of maximal fairness that requires a criterion's satisfaction for all *relevant* partitions like races, sexes, and so on. We agree that bias against some groups may be a more pressing issue than bias against just any gerrymandered group, but there are several things one might say in response; here are four. First, we cannot rule out singleton groups. Even for core protected classes like races or sexes, there may be applications and populations in which some group is a singleton. So *a priori* constraints on the size of groups have to be abandoned if we want to treat arbitrary populations. Second, there may well be nothing interesting to say about the set of "relevant partitions" in general. Such a set could consist of a single partition with no characterizations of the sort in (Stewart, 2022) and in Proposition 2 available. Third, and modestly, fairness for all partitions or all groups is clearly *one*

of Calibration, no group would have a complaint about the assessor's bias in confidence. (Stewart, 2024) suggests that by considering what maximal fairness looks like according to different algorithmic fairness criteria, we get information relevant to the normative evaluation of those criteria, information about what "ideals" the various criteria are committed to. For central criteria of algorithmic fairness, maximal fairness can be characterized in terms of simple properties that concern fair treatment. For example, h is calibrated for all partitions if and only if h is perfect: $h(i) = Y(i)$. While h satisfies Predictive Equity for all partitions if and only if h satisfies the weaker property of making perfect distinctions: $Y(i) \neq Y(j)$ implies $h(i) \neq h(j)$. Arguably, perfection excludes any unfair treatment. Perfect distinctions does not. It is consistent with assigning every recidivist a risk score 0 and every non-recidivist a risk score of 1, or assigning all individuals different scores in any arbitrary fashion. For example, as long as all individuals are assigned distinct scores, Predictive Equity's ideal is consistent with assigning very high scores to all black individuals and very low scores to all white individuals. The thought is that Calibration encodes a more attractive ideal. What about Spanning?

Proposition 2. *Let h be an assessor for a (non-homogeneous) population N. The following are equivalent.*

1. *h satisfies Calibration for all partitions of N.*
2. *h satisfies Base Rate Tracking for all partitions of N.*
3. *h satisfies Spanning for all partitions of N.*
4. *h is perfect.*

But satisfying Predictive Equity for all partitions of N does not imply that h is perfect.

So while Spanning and Base Rate Tracking retain perfection as maximal fairness, Predictive Equity does not. Those that find the ideals program—and the ideal of perfection in particular—compelling might well take this as a consideration in support of Spanning. One consequence of Proposition 2 is that any putatively necessary criteria of algorithmic fairness that we add to Spanning will at least entail that maximal fairness implies the perfection ideal.

The third and perhaps weightiest motivation for Spanning is that, unlike both Base Rate Tracking and Predictive Equity, it is non-trivially consistent with Equalized Odds, both generally and in the special case of binary assessment. Stewart et al. (2024, Theorem) prove that relaxing Calibration to Base Rate Tracking opens up no new possibilities at all: Base Rate Tracking and Equalized Odds are consistent only in trivial assessment scenarios. And Stewart and Nielsen (2020, Theorem 3) show that relaxing Calibration to Predictive Equity opens up no new possibilities for binary assessment unless the assessor is "maximally error-prone": Predictive Equity and Equalized Odds are consistent for binary assessors only if the assessment scenario is trivial or the assessor's false positive and false negative rates are

sensible conception of maximal fairness. Others may well be worth investigating, but this conception has some precedent in the literature and others may be less interesting analytically. Fourth, even if we agree on some set of partitions as the relevant ones, it does not follow that bias against groups that are not cells in those partitions is ethically unimportant. While the set of people from rural areas may not have been listed as a relevant group, the discovery of significant bias against the group may be diagnostic, ethically concerning, and worth reducing. To us, it seems wrong in such a case to regard an assessor as *maximally* fair when another assessor might reduce bias against rural folks without introducing any new bias. We take considerations like these to suggest that the conception of maximal fairness that we adopt is interesting, legitimate, and worth exploring.

maximal (i.e. equal to 1) for all groups. Crucially, Spanning avoids both of these impossibility results.

Theorem 1. *There exist non-trivial assessment scenarios in which continuous assessors satisfy both Spanning and Equalized Odds. And there exist non-trivial assessment scenarios in which binary assessors satisfy both Spanning and Equalized Odds without being maximally error-prone.*

The theorem is proved by considering examples. We begin with a particularly simple example that establishes Theorem 1's first claim. In the examples below, an asterisk indicates that the individual has property y.

Example 1: A non-trivial assessment scenario with a continuous assessor satisfying Spanning and Equalized Odds

Group 1	$h(1) = 1/3$	$h(2^*) = 2/3$	$h(3) = 1/3$
Group 2	$h(4^*) = 2/3$	$h(5^*) = 2/3$	$h(6) = 1/3$

Spanning is easy to verify after observing that the base rate in Group 1 is 1/3 and the base rate in Group 2 is 2/3. The generalized false positive rate in Group 1 is $f^{+}_{G_1}(h) = E_G(h|Y = 0) = 1/3$, which is the same as in Group 2. Likewise, the generalized false negative rate in Groups 1 and 2 is 1/3. So the assessor in Example 1 satisfies both Spanning and Equalized Odds. Note that this assessor is by no means unique. Given the same population there are many other assessors that also non-trivially satisfy Spanning and Equalized Odds. For instance, with $c \in (0, 1/3)$:

Example 2: A continuum of non-trivial assessment scenarios with continuous assessors satisfying Spanning and Equalized Odds

Group 1	$h(1) = c$	$h(2^*) = 1 - c$	$h(3) = c$
Group 2	$h(4^*) = 1 - c$	$h(5^*) = 1 - c$	$h(6) = c$

At this point, it becomes a straightforward exercise to show that larger populations give rise to an even greater variety of possibilities. Example 3 shows that h need not be two-valued.

Example 3: A non-trivial assessment scenario with a continuous, non-two-valued assessor satisfying Spanning and Equalized Odds

Group 1	$h(1^*) = 3/5$	$h(2^*) = 9/10$	$h(3) = 1/10$
	$h(4) = 1/10$	$h(5) = 1/5$	$h(6) = 1/5$
Group 2	$h(7^*) = 3/5$	$h(8^*) = 3/5$	$h(9^*) = 9/10$
	$h(10^*) = 9/10$	$h(11) = 1/10$	$h(12) = 1/5$

The next example establishes Theorem 1's second claim.

Example 4: A non-trivial assessment scenario with a binary, not maximally error-prone assessor satisfying Spanning and Equalized Odds

Group 1	$h(1^*) = 0$	$h(2^*) = 1$	$h(3) = 0$	
Group 2	$h(4^*) = 1$	$h(5^*) = 0$	$h(6) = 0$	h(7)=0

Again, the assessor in Example 4 is not unique for the given population (though there cannot be a *continuum* of binary fair assessors). For instance, interchanging the values of $h(1^*)$ and $h(2^*)$ preserves the relevant features of the example.

4 Strengthening the Consistency Result: Strong Equalized Odds

The foregoing examples actually establish something stronger than the non-trivial consistency of Spanning with Equalized Odds. Part of explaining this point involves clarifying an issue about which some confusion exists in the literature. Unfortunately, a number of formal conditions are called *Calibration* in the literature on algorithmic fairness. Similarly, and more to our point in this section, a number of formal conditions go by the name *Equalized Odds*. One consequence of this unfortunate state of affairs is that impossibility theorems are sometimes misreported and incorrectly described. For instance, it's been claimed that the result due to Kleinberg and coauthors establishes the non-trivial inconsistency of the criterion that we call Predictive Equity and the following criterion, sometimes itself called Equalized Odds, but which, for reasons that will become obvious, we refer to as *Strong Equalized Odds* (Beigang, 2023a, p. 169).

> **Strong Equalized Odds**. For an assessor h of N and any groups G, G' in the relevant partition π of N, $P_G(h = p|Y = 1) = P_{G'}(h = p|Y = 1)$ and $P_G(h = p|Y = 0) = P_{G'}(h = p|Y = 0)$ for any p (whenever those probabilities are defined).

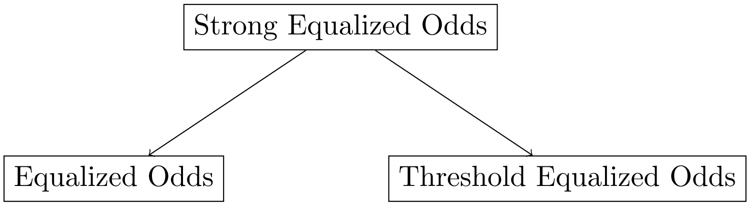


Figure 3: Two independent weakenings of Strong Equalized Odds

Strong Equalized Odds requires that, across groups, equal *proportions* of individuals with (respectively, without) property y are given score p. So, if 50% of black recidivists are assigned a score of 0.8, for example, then 50% of white recidivists should be assigned a score of 0.8. But we have two quick points to make about the report that this property is inconsistent with Predictive Equity outside of trivial scenarios. First, contrary to that report, Predictive Equity and Strong Equalized Odds are not the criteria that figure in the Kleinberg et al. impossibility theorem. Second, and more importantly, these criteria are not inconsistent outside of trivial scenarios. Example 6 in the Appendix witnesses this fact.

Another criterion often referred to as Equalized Odds is the following condition that we will call *Threshold Equalized Odds* (e.g., Crespo et al., 2024).

> **Threshold Equalized Odds**. For an assessor h of N, a specified threshold $t \in [0, 1]$, and any groups G, G' in the relevant partition π of N, $P_G(h > t|Y = 0) = P_{G'}(h > t|Y = 0)$ and $P_G(h < t|Y = 1) = P_{G'}(h < t|Y = 1)$ for any p (whenever those probabilities are defined).

The quantities $P_G(h > t|Y = 0)$ and $P_G(h < t|Y = 1)$ are sometimes referred to as the false positive and false negative rates, respectively, for group G. Suppose we specify a threshold of 0.25. If 60% of Asian non-recidivists are assigned a score above 0.25, Threshold Equalized Odds requires that 60% of non-recidivists in every other racial group be assigned a score above 0.25. And if 0% of Asian recidivists are assigned a score below 0.25, the false negative rate for all groups must be 0.

We now clarify the relationships between these three criteria that have all been called Equalized Odds. Again, the logical relationships summarized in Proposition 3 are also depicted in Figure 3.

Proposition 3. *Strong Equalized Odds implies Equalized Odds and Threshold Equalized Odds, but Equalized Odds and Threshold Equalized Odds are independent of one another.*

Proposition 3 allows us to state our main point: Spanning is non-trivially consistent with Strong Equalized Odds, as Examples 1 and 3 establish. It follows from Proposition 3 that it is also consistent with Threshold Equalized Odds. Of course, neither Calibration nor Base Rate Tracking nor Predictive Equity is consistent with Strong Equalized Odds outside of trivial scenarios.

5 Objections and Responses

We have been arguing that weakening Calibration to Spanning opens up new possibilities for fair assessment. Throughout the previous section, we assumed that Equalized Odds is necessary for fair assessment. This assumption has received plenty of support in the literature,[5] but Equalized Odds also has its detractors. Hedden (2021), in particular, argues that Equalized Odds is vulnerable to counterexamples. Hedden's arguments have already garnered a lot of attention, and although this debate extends beyond the scope of our paper, it is worth making two brief remarks.

First, Hedden's argument is compatible with our main point. If fair assessors need not satisfy Equalized Odds, then there are even more possibilities for fair assessment than we imagined in the previous section. However, one might think that Hedden's counterexamples do threaten our argument with otiosity. By Hedden's lights, his counterexamples refute every prominent statistical criterion of fairness except Calibration, and if Calibration is the only fairness criterion, then it's already clear that the impossibility results can be avoided. So, arguably, Hedden's counterexamples obviate the motivation to explore weakenings of Calibration such as Spanning. Our second remark about Hedden's argument, however, is that not everyone agrees that his counterexamples to Equalized Odds succeed (Viganò et al., 2022; Søgaard et al., 2024). We are sympathetic to a line of objection on which Hedden's alleged counterexamples have features that genuine assessment problems lack. For instance, Hedden's examples involve assessors that have access to the "objective chances" that individuals have property y. But this never occurs in the cases that motivate research on algorithmic fairness. We never know the objective chance that a given individual will recidivate, for example—even supposing it makes sense at all to speak of chances in this case. In genuine prediction problems, assessors must make predictions without being able to take objective chances into account. To be fair, Hedden concedes, "I am not claiming that the case of people, coins, and rooms is realistic or completely analogous to cases like COMPAS. Of course it is not [. . .] human behavior is not normally chancy, at least not in the same way that coin flips are" (2021, p. 223). However, the objection we mean to express sympathy with here is not merely that Hedden's examples are unrealistic, but that their unrealistic features make them different in kind from the cases that Equalized Odds is supposed to constrain.[6] If that's right, then Hedden's examples don't show that Equalized Odds isn't required for fairness in genuine assessment problems. But providing detailed support for this objection will have to be left for another time. In any case, having noted some of the controversies surrounding Hedden's argument, our assumption that Equalized Odds is necessary for fair assessment is not implausible.

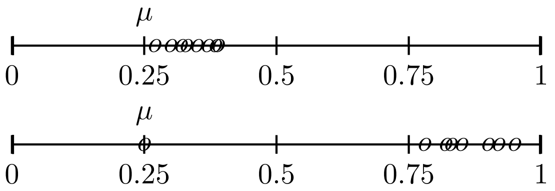
A second potential worry about our argument is that Spanning yields the wrong verdicts in specific cases. Consider Figure 4. Interpret the two images as two different assessors both issuing predictions about a single group of eight individuals. Given that the base rate for the group is 0.25, the top assessor violates Spanning whereas the bottom one satisfies it. But the top assessor's predictions are, on average, much closer to the base rate. Is that reason to think the top assessor is *more fair*? We would like to make two points in response to this concern.

First, the alternatives to Spanning that we have considered—Calibration, Base Rate Tracking, Predictive Equity—also fail to yield the verdict that the top assessor is "more fair" than the bottom. Both the top and bottom assessors violate Calibration. And since we are

[5] See Grant (2023) for a recent defense.

[6] Søgaard et al. (2024) develop a similar point in detail, arguing that Hedden's counterexample is not a standard machine learning problem. Their focus is the fact that the performance of Hedden's assessors cannot be evaluated on "held-out data (a new random sample) from the underlying distribution" (p. 9).

Figure 4: Base rate $\mu = 0.25$ *with eight predictions, each marked by an* o



considering a single group, Base Rate Tracking and Predictive Equity are silent since they are inter-group criteria. We might seek to supplement Calibration or one of its other weakenings with an account of how "close" an assessor is to satisfying the relevant constraint. So for some ways of specifying closeness, the top assessor is closer to Calibration than the bottom assessor is. Even with an appropriate account of closeness in hand, however, the question of the relative fairness of the two assessors is not settled, which leads us to the next point.

The second point about this second objection that we would like to make is that we are not proposing Spanning as a sufficient condition for fairness. And neither is Calibration nor either of its alternative weakenings advanced as a sufficient condition. We can even concede that there is something unfair about each assessor in Figure 4. Nothing that we have said commits us to insisting that the bottom assessor is fair. Whether an assessor's forecasts count as fair will depend on the other conditions required for fairness. If we were to interpret the two images as a *single* assessor's predictions for two *different* groups, each with the same base rate, it would be clear that in addition to violating Spanning, these forecasts violate Equalized Odds. Since the base rate is 0.25, six out of the eight individuals do not have property y. So the generalized false positive rate is considerably greater in the bottom group. If we suppose that Spanning and Equalized Odds are necessary conditions for fairness, the defects of the assessor in the single-assessor interpretation of Figure 4 are multiple. If we insist on Calibration or one of its alternative weakenings, on the other hand, since these criteria are generally inconsistent with Equalized Odds, Equalized Odds cannot also be a general necessary criterion of fairness if any of those are. As a result, we are prevented from diagnosing the bias in generalized error rates in the example.

Another way of developing the wrong verdicts concern challenges the notion that failures of Spanning can diagnose unfair treatment of a group independently of the assessor's forecasts for other groups.[7] Consider the top image in Figure 4 again, but suppose that the assessment problem is about successful performance in some college if admitted, for instance. In this case, the worry goes, the assessor's rosy evaluations of the group are not unfair to that group. After all, the assessor's overconfidence in the group is to the group's benefit. We think that there are several potentially mollifying considerations here. First, as we noted above, this objection does not rule out the possibility that certain failures of Spanning are unfair regardless of the assessor's performance in other groups. Consider the same image in Figure 4 under the interpretation of recidivism forecasting, for instance. Second, as we also mentioned above when we introduced Spanning, even in cases like the college admissions example, Spanning could represent a sort of epistemic defect in the forecasts. Overconfidence, even if beneficial, is a form of inaccuracy. Such inaccuracy could (and typically does) frus-

[7] Thanks to an anonymous referee for putting this objection to us.

trate the primary purpose of the forecasts—like admitting a talented and successful class of students. Moreover, such inaccuracy could also have problematic consequences for the very group that apparently benefits. As Borsboom and coauthors note in the context of college admissions, the likely eventual underperformance of the group subject to overly optimistic assessment could reinforce negative group stereotypes (2008). So concluding that the assessor's overconfidence is to the group's benefit in the college admissions setting is hasty. Third, Spanning doesn't imply that the assessor in Figure 4 is unfair *to the top group*. If Spanning is necessary for fairness, and an assessor violates Spanning, then it follows that the assessor is unfair, but it is a further question *which group(s)* the assessor treats unfairly. Answering this question probably is not a purely statistical matter and might require an investigation of the moral/ethical features of the case under consideration. In the college admissions problem, for example, it could be that the bottom group of Figure 4 is treated unfairly. The bottom group may complain that the admissions assessor is overconfident about the top group. Fourth, and most concessively, we could reformulate Spanning in various accommodating ways. For instance, in assessment problems in which only one type of error is harmful or intuitively unfair, Spanning could be relaxed to requiring only that not all the forecasts are above or below the group base rate depending on whether higher or lower probabilities are considered harmful. In the top image in Figure 4, such a modification of Spanning might classify the forecasts as unfair when recidivism is being predicted, but not when college success is the forecasting task.

A third worry, related to the previous ones, is that, due either to its weakness or to its formal structure, Spanning is "gameable" in problematic ways. Some agents, like the proprietors of algorithms used in industry or the criminal justice system, may have incentives to avoid accusations of algorithmic bias regardless of whether the algorithms are in fact fair or not. Perhaps, the worry goes, it is too easy to avoid accusations of bias based on violations of Spanning. Similarly, machine learning algorithms may seek to satisfy Spanning in trivial or unappealing ways. In *The Ethical Algorithm*, Kearns and Roth observe, "One theme running throughout this book is that algorithms generally, and especially machine learning algorithms, are good at optimizing what you ask them to optimize, but they cannot be counted on to do things you'd like them to do but didn't ask for, nor to avoid doing things you didn't want but didn't tell them not to do" (2019, p. 87). In the most extreme case, can't an algorithm "optimize" for fairness and avoid charges of bias simply by assigning two individuals in each group a score of 0 and 1, respectively?

If Spanning were all we cared about, maybe so. But, if Equalized Odds is also necessary for fairness, then accusations of bias are not so easily avoided. If an algorithm haphazardly assigns 0s and 1s in order to satisfy Spanning, it might make error rates differ across groups. In that case, the attempt to "game" Spanning actually creates biases in the form of Equalized Odds violations. But is there a way to game Spanning and Equalized Odds together? For instance, is there an algorithm that assigns some 0s and 1s in order to satisfy Spanning and is guaranteed to satisfy Equalized Odds as well? In general, no. In order to assign 0s and 1s while guaranteeing that Equalized Odds is satisfied, one must know which individuals have property y and which do not. And that is precisely what we do not know in situations that call for predictive algorithms. So, while it's true that it is easy to guarantee that Spanning is satisfied, that's not tantamount to a guarantee that accusations of bias can be avoided.

But perhaps the ease with which Spanning can be satisfied is troubling on its own, even if it does not point to a general method for gaming algorithmic fairness. In our view, this is not a serious problem. Equalized Odds is easy to satisfy too—any constant assessor will do—but that is not taken as a mark against it. The reason why mirrors what we have just said about Spanning: while a constant assessor is guaranteed to satisfy Equalized Odds, in general it will not satisfy Spanning or other weakenings of Calibration, so in general one cannot game Equalized Odds without violating other plausible fairness criteria. The lesson here is that we should be concerned about the "gameability" of proposed fairness criteria only to the extent that there are general methods for guaranteeing that *all* of the criteria are satisfied. That an individual constraint like Spanning (or Equalized Odds) is easy to satisfy tells us little about its plausibility as a criterion for fairness.

Acknowledgments. Thanks to Lee Elkin, Jens Kipper, Ignacio Ojea Quintana, Tom Sterkenburg, Reuben Stern, two anonymous referees, and an audience at the Northeastern Ethics & Epistemology Workshop 2024 for helpful conversations and feedback.

Appendix

Proof of Proposition 1

Proof. Proposition 1 in (Stewart et al., 2024) establishes that Calibration for some partition π implies Base Rate Tracking for π. That same proposition also establishes that Calibration for π implies that $E_G(h) = \mu_G$ for all $G \in \pi$. But this fact clearly requires that $\mu_G \in [\min_{i \in G} h(i), \max_{i \in G} h(i)]$ for all $G \in \pi$. So Calibration for π implies Spanning for π. And that Calibration implies Predictive Equity is immediate.

Next, to see that Base Rate Tracking implies neither Predictive Equity nor Spanning, consider the following.

Example 5: Base Rate Tracking without Predictive Equity or Spanning

Group 1	$h(1^*) = 3/4$	$h(2) = 2/3$
Group 2	$h(3^*) = 2/3$	$h(4) = 3/4$

In Example 5, $E_{G_1}(h) = E_{G_2}(h)$ and $\mu_{G_1} = \mu_{G_2}$, so h satisfies Base Rate Tracking. But $P_{G_1}(Y = 1|h = 3/4) = 1 \neq P_{G_2}(Y = 1|h = 3/4) = 0$, so h does not satisfy Predictive Equity. And since the base rate in each group is 1/2 and lies below the least prediction in each group, h does not satisfy Spanning.

Next is an example of an assessor that satisfies Predictive Equity for a partition but neither Base Rate Tracking nor Spanning for that partition.

Example 6: Predictive Equity without Base Rate Tracking or Spanning

Group 1	$h(1^*) = 4/5$	$h(2^*) = 4/5$	$h(3) = 3/4$
Group 2	$h(4^*) = 4/5$	$h(5) = 3/4$	

To verify Predictive Equity, notice that there are just two scores so two cases to check. First, $P_{G_1}(Y = 1|h = 4/5) = P_{G_2}(Y = 1|h = 4/5) = 1$. Second, $P_{G_1}(Y = 1|h = 3/4) = P_{G_2}(Y = 1|h = 3/4) = 0$. So h satisfies Predictive Equity. But h does not satisfy Base Rate Tracking since $E_{G_1}(h) - \mu_{G_1} = 7/60$ and $E_{G_2}(h) - \mu_{G_2} = 11/40$. The base rate for Group 1 is 2/3 which is less than the least score h assigns in that group. (Spanning likewise fails in Group 2.) So h does not satisfy Spanning.

Finally, Spanning implies neither Base Rate Tracking nor Predictive Equity. Example 1 provides an example of an assessor that satisfies Spanning but not Base Rate Tracking.[8] (This is as should be expected in light of the theorem in (Stewart et al., 2024).) And Example 4 is an example of an assessor that satisfies Spanning but not Predictive Equity. (Again, this is as we should expect given Theorem 3 in (Stewart and Nielsen, 2020).) □

Proof of Proposition 2

Proof. That Calibration for all partitions is equivalent to perfection and not equivalent to Predictive Equity for all partitions is established in (Stewart, 2022). The equivalence of Base Rate Tracking for all partitions and perfection, under the assumption that the population is non-homogeneous, is established in Stewart (2024). For the equivalence of Spanning for all partitions and perfection, suppose first that Spanning holds for all partitions. Then it holds for the partition of N into singletons. Since the span of h on a singleton is the degenerate interval $[h(i), h(i)]$, Spanning requires $\mu_{\{i\}} = h(i)$. But for any singleton $\{i\}$, the base rate $\mu_{\{i\}}$ is either 1 if $Y(i) = 1$ or 0 if $Y(i) = 0$. This implies that $h(i) = 1$ if $Y(i) = 1$, and $h(i) = 0$ if $Y(i) = 0$, which is to say that h is perfect. Conversely, suppose that h is perfect. For any group G in some partition π of N, there are three cases to consider: the base rate is either 0, 1, or something in between. If the base rate of G is 0 (respectively, 1), then h assigns 0 (1) to every member of G because it is perfect, and Spanning is satisfied. If the base rate of G is strictly in between 0 and 1, then some member i of G is such that $Y(i) = 1$, which implies $h(i) = 1$ by perfection, and some other member j of G is such that $Y(j) = 0$, which implies $h(j) = 0$. So, in this case too, Spanning is satisfied. □

Remark. Like with Calibration but unlike with Base Rate Tracking, the equivalence of Spanning for all partitions and perfection does not require the restriction to non-homogeneous populations.

[8] Here, it is important to stress that Base Rate Tracking is *not* formulated in terms of the absolute value of $E_G(h) - \mu_G$.

Proof of Proposition 3

Proof. To see that Strong Equalized Odds implies Equalized Odds, suppose that, for all $G, G' \in \pi$, $P_G(h = p|Y = 1) = P_{G'}(h = p|Y = 1)$ and $P_G(h = p|Y = 0) = P_{G'}(h = p|Y = 0)$. Let $h[A]$ be the image of h on $A \subseteq N$ so that

$$h[G \cap \{Y = 0\}] = \{p : h(i) = p \text{ for some } i \in G \text{ such that } Y(i) = 0\}.$$

By Strong Equalized Odds, $h[G \cap \{Y = 0\}] = h[G' \cap \{Y = 0\}]$ for all $G, G' \in \pi$ since, otherwise, there would be some $G, G' \in \pi$ and some $p \in [0, 1]$ such that $P_G(h = p|Y = 0) > 0 = P_{G'}(h = p|Y = 0)$, which contradicts Strong Equalized Odds. Using this fact and Strong Equalized Odds,

$$\begin{aligned} E_G(h|Y = 0) &= \sum_{p \in h[G \cap \{Y=0\}]} pP_G(h = p|Y = 0) \\ &= \sum_{p \in h[G' \cap \{Y=0\}]} pP_{G'}(h = p|Y = 0) \\ &= E_{G'}(h|Y = 0). \end{aligned}$$

Hence, Strong Equalized Odds implies equal generalized false positive rates: $f^+_G(h) = f^+_{G'}(h)$ for all $G, G' \in \pi$. (And note that, when $P_G(h = p|Y = 0)$ is undefined, so is $E_G(h|Y = 0)$.) The argument for equal generalized false negative rates is analogous. So, Strong Equalized Odds implies Equalized Odds.

For the implication from Strong Equalized Odds to Threshold Equalized Odds, assume Strong Equalized Odds again and observe that for any $G, G' \in \pi$ and any $t \in [0, 1]$,

$$\begin{aligned} P_G(h > t|Y = 0) &= \sum_{p>t} P_G(h = p|Y = 0) \\ &= \sum_{p>t} P_{G'}(h = p|Y = 0) \\ &= P_{G'}(h > t|Y = 0), \end{aligned}$$

whenever $P_G(\cdot|Y = 0)$ and $P_{G'}(\cdot|Y = 0)$ are both defined. An analogous argument establishes that $P_G(h < t|Y = 1) = P_{G'}(h < t|Y = 1)$ for any $t \in [0, 1]$ whenever both terms are defined. So, Strong Equalized Odds implies Threshold Equalized Odds.

To see that Equalized Odds and Threshold Equalized Odds are independent, first consider Example 7.

Example 7: Equalized Odds without Threshold Equalized Odds for $t = 0.4$

Group 1	$h(1^*) = 1/3$	$h(2^*) = 2/3$	$h(3) = 1/5$
Group 2	$h(4^*) = 1/2$	$h(5) = 1/5$	

Here, $f^+_{G_1}(h) = f^+_{G_2}(h) = 1/5$ and $f^-_{G_1}(h) = f^-_{G_2}(h) = 1/2$, securing Equalized Odds. But suppose we set $t = 0.4$. Then, $P_{G_1}(h < 0.4|Y = 1) = 1/2 \neq P_{G_2}(h < 0.4|Y = 1) =$

0. So h violates Threshold Equalized Odds. Next, Example 5 presents an assessor that satisfies Threshold Equalized Odds for many choices of threshold but not Equalized Odds. For instance, set $t = 0.4$ again. Then, $P_{G_1}(h > 0.4|Y = 0) = P_{G_2}(h > 0.4|Y = 0) = 1$ and $P_{G_1}(h < 0.4|Y = 1) = P_{G_2}(h < 0.4|Y = 1) = 0$. But $f^+_{G_1}(h) = 2/3 \neq f^+_{G_2}(h) = 3/4$. □

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner (2016). 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.

Beigang, F. (2023a). Reconciling algorithmic fairness criteria. *Philosophy & Public Affairs 51*(2), 166–190.

Beigang, F. (2023b). Yet another impossibility theorem in algorithmic fairnes. *Minds and Machines*, 1–21.

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.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data 5*(2), 153–163.

Crespo, V., B. Eva, and W. Sinnott-Armstrong (2024). Applying base rate tracking and COMPAS.

Dorst, K. (2023). Rational polarization. *The Philosophical Review 132*(3), 355–458.

Dorst, K., B. A. Levinstein, B. Salow, B. E. Husic, and B. Fitelson (2021). Deference done better. *Philosophical Perspectives 35*(1), 99–150.

Eva, B. (2022). Algorithmic fairness and base rate tracking. *Philosophy and Public Affairs 50*(2), 239–266.

Grant, D. G. (2023). Equalized odds is a requirement of algorithmic fairness. *Synthese 201*(3), 101.

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs 49*(2), 209–231.

Huttegger, S. M. and M. Nielsen (2020). Generalized learning and conditional expectation. *Philosophy of Science 87*(5), 868–883.

Kearns, M. and A. Roth (2019). *The Ethical Algorithm: The Science of Socially Aware Algorithm Design.* Oxford, UK: Oxford University Press.

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)*, pp. 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Larrick, R. P. and J. B. Soll (2006). Intuitions about combining opinions: misappreciation of the averaging principle. *Management Science 52*(1), 111–127.

Nielsen, M. (2021). A new argument for Kolmogorov conditionalization. *Review of Symbolic Logic 14*(4), 930–945.

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.

Søgaard, A., K. Kappel, and T. Grünbaum (2024). On Hedden's proof that machine learning fairness metrics are flawed. *Inquiry*, 1–20.

Stewart, R. T. (2022). Identity and the limits of fair assessment. *Journal of Theoretical Politics 34*(3), 415–442.

Stewart, R. T. (2024). The ideals program in algorithmic fairness. Unpublished Manuscript.

Stewart, R. T., B. Eva, S. Slank, and R. Stern (2024). An impossibility theorem for base rate tracking and equalized odds. Forthcoming in *Analysis*.

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

van Fraassen, B. C. (1984). Belief and the will. *The Journal of Philosophy 81*(5), 235–256.

van Fraassen, B. C. (1999). Conditionalization, a new argument for. *Topoi 18*(2), 93–96.

van Fraassen, B. C. and J. Y. Halpern (2016). Updating probability: Tracking statistics as criterion. *The British Journal for the Philosophy of Science 68*(3), 725–743.

Viganò, E., C. Hertweck, C. Heitz, and M. Loi (2022). People are not coins: Morally distinct types of predictions necessitate different fairness constraints. In *FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2293–2301.

Statements and Declarations

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript. The authors have no relevant financial or non-financial interests to disclose.