Forthcoming in *AI & Society*

The Ideals Program in Algorithmic Fairness

Rush T. Stewart

November 8, 2024

Abstract

I consider statistical criteria of algorithmic fairness from the perspective of the *ideals* of fairness to which these criteria are committed. I distinguish and describe three theoretical roles such ideals might play. The usefulness of this program is illustrated by taking Base Rate Tracking and its ratio variant as a case study. I identify and compare the ideals of these two criteria, then consider them in each of the aforementioned three roles for ideals. This ideals program may present a way forward in the normative evaluation of candidate statistical criteria of algorithmic fairness.

Keywords. AI ethics; algorithmic fairness; base rate tracking; bias; calibration; equalized odds

1 Introduction

Predictive algorithms are employed in a number of important social settings including decisions about loans in the financial sector and sentencing in the criminal justice system. Disparate impact across sub-groups of assessed populations has raised urgent ethical concerns (Angwin et al., 2016). While it is widely recognized that algorithmic fairness is an important goal, what fairness amounts to in such contexts is controversial. The dominant approach is to articulate necessary conditions for fairness in the form of statistical criteria that an algorithm's assessments must satisfy for the algorithm to qualify as fair. Hedden (2021) argues that none of the standard statistical criteria—except, perhaps, Calibration—are genuine necessary conditions of algorithmic fairness. Grant (2023) defends (a version of) Equalized Odds, another of the most prominent statistical criteria of algorithmic fairness. In the background of this debate are impossibility theorems, theorems establishing that certain criteria are impossible to jointly satisfy in non-trivial contexts. For example, Calibration and Equalized Odds are jointly unsatisfiable outside of conditions that can be called trivial (Kleinberg et al., 2017). Given these sorts of impossibility results, a crucial line of research scrutinizes conceptions of fairness as formalized in these statistical criteria in an effort to discern which are most compelling to retain.[1] What's more, the rapid proliferation

[1]Others have argued for certain modifications of all criteria featuring in an impossibility result with the express aim of relieving the relevant inconsistency (Beigang, 2023).

of alternative criteria for evaluating *forecasts* only exacerbates the need for general means of evaluating or ways of bringing additional considerations to bear on such *criteria* themselves (Narayanan, 2018; Verma and Rubin, 2018; Eva, 2022; Edenberg and Wood, 2023; Loi et al., 2023).

My aim here is to emphasize a neglected dimension along which fairness criteria can be compared: the *ideals* of fair assessment that candidate criteria encode. A criterion's fairness ideal is what maximal fairness looks like according to that criterion. For the criteria that I discuss here, the ideals take very simple forms that recognizably concern fair treatment. Because of this, ideals may facilitate certain comparisons between candidate statistical criteria. I distinguish different theoretical uses to which such ideals could be put in the normative evaluation of putative statistical criteria of algorithmic fairness. The articulation of this program is the primary conceptual contribution of this essay. I do not propose a satisfactory account of algorithmic fairness; I present a tool to help in formulating such an account. I then illustrate the usefulness of ideals with respect to the recently-proposed criterion of Base Rate Tracking (Eva, 2022), identifying the fairness ideal to which that criterion is committed. The ideal is compelling, but it is not one to which many other prominent criteria—including an alternative form of Base Rate Tracking that Eva suggests—are committed. While certain other considerations so far adduced in the literature fail to distinguish Base Rate Tracking from its alternative form, ideals may contribute to a case for the original form of the criterion. Characterizing maximal fairness for Base Rate Tracking (Theorem 1) and Ratio Base Rate Tracking (Theorem 2) is the primary technical contribution of this study.

2 Statistical Criteria

In assessment problems of the sort in which we are interested, there is some population N of individuals. The population might be candidates for loans, applicants to colleges, inmates coming up for parole, or any number of other things. I focus on the simple case of assessing whether the individuals have some property y or not. The property y might be whether the applicant with graduate within six years or whether the inmate will reoffend. For individual i in N, we write $Y(i) = 1$ if i has y, and $Y(i) = 0$ if i does not have y. In most interesting cases, we do not know whether a given individual has property y. We have to predict on the basis of information that we do have. Often this information is represented as a vector of features associated with the individual. In the case of predicting recidivism, these features might include type of crime, criminal history, employment status, age, etc. For the sake of simplicity, we can abstract from making these lists of features explicit and deal only with the individuals. An *assessor* is a function $h : N \to [0, 1]$. We can interpret the number $h(i)$ as a prediction about whether individual i has property y; in particular, we will construe $h(i)$ as the *probability* that i has y. In the case of forecasting recidivism, $h(i)$ is often called a *risk score* and is produced by a particular forecasting algorithm.

Within a general population N, it is customary to distinguish certain groups. Standard divisions of a population concern race, gender, disability status, nationality, age, and so on. Each such way of looking at the population generates a *partition* π of N, a way of dividing N into groups such that each individual belongs to exactly one group. Defining prominent criteria of fairness requires reference to proportions in N and subpopulations (or groups in a

given partition). To this end, we can introduce a uniform distribution P on N. The quantity $P(Y = 1) = \mu$ is called the *base rate*, the proportion of individuals in N that have property y. For any partition $\pi = \{G_1, \ldots, G_m\}$ of N, $P_k = P(\cdot|G_k)$ is the uniform distribution on G_k so that $P_k(Y = 1) = \mu_k$ is the prevalence of y in group G_k where $k = 1, \ldots, m$.

Standardly, the focus is on fair assessment across a given way of carving the population into groups. The central ethical question is, What properties must an assessor have to count as fair?[2] The standard approach is to propose *necessary* conditions of algorithmic fairness. Perhaps the most important statistical criterion in fair algorithms research is the familiar probabilistic concept of Calibration.

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

In words, Calibration for a partition requires that p% of the individuals in group G given the score p actually have property y. So, of those assigned a score 0.8 in G, for instance, 80% of them have y. Calibration mandates that the assessor's forecast probabilities match the empirical frequencies in this way. In the context of risk scores for recidivism, Calibration rules out forms of overconfidence in one race, say, being more likely to reoffend than another. For Calibrated assessors, it cannot be the case that 80% of white defendants who are assigned a risk score of 2/3 go on to reoffend while only 40% of black defendants assigned a risk score of 2/3 go on to reoffend. Not only do calibrated risk assessors not display this sort of simultaneous overconfidence in black recidivism and underconfidence in white recidivism, they do not display over- or underconfidence at any score for any group. Sometimes Calibration is motivated by the observation that it guarantees that scores for different groups will "mean the same thing," unlike the risk score 2/3 in the example above.

Another prominent criterion requires that error rates are the same across groups. We need to introduce a bit more machinery to define it. The expectation E_G is defined in terms of P_G. Define the *generalized false positive rate* in G as $f^+_G(h) = E_G(h|Y = 0)$. Since P_G is uniform on G, $E_G(h|Y = 0)$ is just the *average* score in G for individuals that do *not* have property y. Define the generalized *false negative rate* in G analogously as $f^-_G(h) = E_G(1 - h|Y = 1)$, the average of the quantity $1 - h$ (in G) for individuals that *possess* y.

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

[2]There are different ways of understanding the relevant notion(s) of fairness for predictive algorithms. While some have recently stressed the distinction between fairness in forecasts and fairness in decision-making, arguing that different criteria might be appropriate for different tasks (Beigang, 2022), the more standard view seems to be that bias in forecasts will infect decisions made on the basis of such forecasts and that it is sufficient to focus on forecasts. About criteria like those considered here, Di Bello and Gong, for example, write, "They are plausible measures of algorithmic fairness on the assumption that differences in algorithmic performance will eventuate in differences in the allocation of benefits and burdens across groups since algorithmic predictions guide these allocations" (2023, p. 4). Still others argue for special notions of predictive justice that are supposed to apply to algorithmic forecasts (Lazar and Stone, 2023). It seems to me that the relevance of the ideals program outlined below could be worked out for any of these various views.

Table 1: Calibration without Equalized Odds

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

Equalized Odds for a partition requires that the mistakes an assessor makes are not distributed in a skewed way among groups in that partition. It rules out a much higher false positive rate for black people than for white people, for instance. In the contexts of forecasting recidivism for sentencing or parole decisions, an asymmetric distribution of errors would impact the lives of black and white people in very different ways. Table 1 displays assessment for a population of 6 individuals divided into two groups. An asterisk indicates that the individual has property y. While the assessor is calibrated for both Black and White groups, it does not satisfy Equalized Odds. While $f^+_{Black}(h) = 2/3$, for example, $f^+_{White}(h) = 1/3$. This could be the basis for a complaint by the non-recidivist (individual 2) in the Black group.

Violations of Equalized Odds are to be expected of Calibrated assessors in light of Kleinberg et al.'s theorem. To the extent one recognizes something compelling in both criteria, there is motivation for seeking ways of relaxing at least one criterion in compelling fashion. The next property is a way of relaxing Calibration.

> **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 $P_G(h = p), P_{G'}(h = p) > 0$.

Predictive Equity for a partition requires that the percentage of people in any group in the partition who are assigned a given score p is the same. Calibration is one way to secure this since all those percentages are equal to p. But, obviously, those terms all being equal to any other value will secure that Predictive Equity is satisfied just as well.

3 Ideal Fairness

Satisfying such a statistical criterion for *all* partitions is typically impractical in interesting assessment exercises. Insofar as this is so, we have to reconcile ourselves with lingering bias against some groups.[3] This point should perhaps serve to encourage some additional caution in drawing inferences about the nature and source of discovered bias. The infeasibility of fairness across all partitions notwithstanding, I have suggested that such an ideal case may still be theoretically informative.

> 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

[3]The (hopefully) more attainable goal, then, is to minimize biased assessment or classification for the most relevant or protected groups.

> 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. (2022, p. 425)

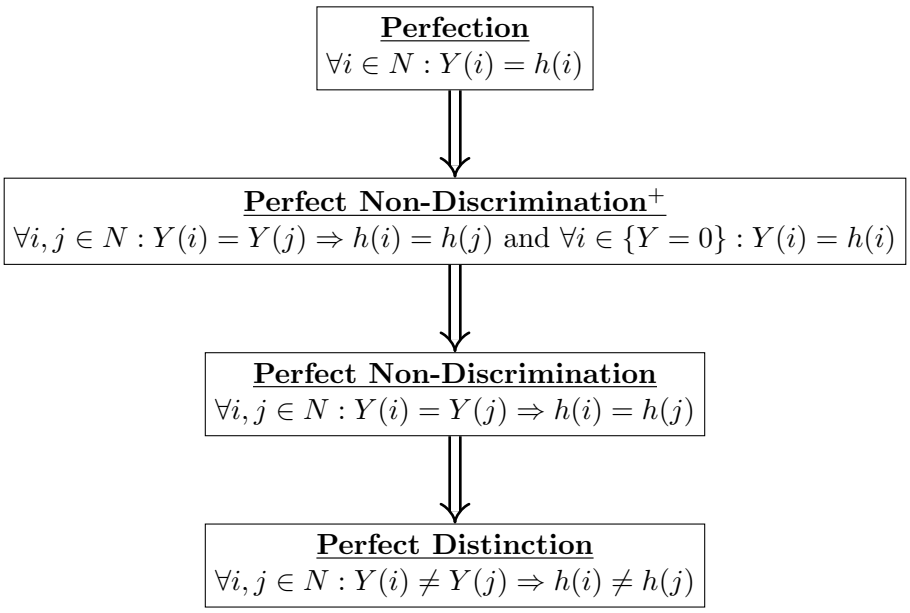
Fairness to salient racial groups is a laudable goal; fairness across races, sexes, nationalities, etc. is even better. In fact, many standard non-discrimination clauses prohibit unfairness across several ways of partitioning a population. Maximal fairness according to a given criterion would be satisfying that criterion across all ways of partitioning the population. This ensures fairness to all groups since any group is a cell of some partition. The idea then is to investigate what ideal assessment, i.e., fairness to all groups, looks like from the perspective of a given statistical criterion. It may be that some ideals are more attractive than others.

Here is another way to motivate the idea that satisfying a given statistical criterion of fairness across all partitions represents maximal fairness according to that criterion. For a given partition of the population, when the assessor fails to satisfy a criterion like Calibration for the partition, some group has a complaint about being treated unfairly. If the assessor satisfies Calibration for that partition, no group in the partition has such a complaint—at least with respect to the notion of fairness Calibration encodes. But even if the assessor is calibrated for one partition, it may be, and generally is, uncalibrated for another partition. So some group in this other partition has an unfairness complaint. If the assessor were calibrated for all partitions, there would exist no group at all that has such a complaint.[4]

I previously characterized ideal assessment for Calibration, Equalized Odds, and Predictive Equity in terms of very simple properties that recognizably concern fair treatment (2022, Observations 1–3). First, an assessor satisfies Calibration for all partitions if and only if it is

[4]One concern about this conception of ideals or maximal fairness—raised by an anonymous referee—is worth addressing directly. Here as in Stewart (2022), an ideal is equivalent to a criterion's satisfaction for *all* partitions. But perhaps this is overly ambitious. Perhaps we should content ourselves with a notion of maximal fairness that concerns only some set of *relevant* partitions. In practice, these often include categories like race, gender, and so on. I think that there are a number of plausible things one might say in response to this concern. Here are three. First, the conception of ideals presented here is clearly *one* sensible notion. Even if there are others worth investigating, that does not imply that this one is not. Second, it is difficult to see that other notions will be as well-behaved and informative as the notion explored here. Treated generally, any set of partitions could be designated as relevant, and this includes sets consisting of a single partition. In such cases, there may be no nice analytic characterizing condition of when a criterion holds for the relevant partitions other than the criterion itself. By contrast, not only are the ideals as conceived here simple and clearly about fairness, they are linked by a striking logical order. Third, the restriction to "relevant" groups comes with seemingly implausible commitments. For instance, suppose that the set of relevant partition does not include a rural/urban partition, as salient examples of non-discrimination clauses often do not. Still, the discovery of significant bias against people from rural regions does not seem ethically irrelevant. Suppose further that we have an assessor that satisfies some fairness criterion for all the relevant partitions, but not for the rural/urban partition of the population. Another assessor satisfies the criterion, not just for the relevant partitions, but for the rural/urban partition as well. Is the second assessor really not more fair in an ethically relevant and theoretically interesting sense? More generally, I find it plausible that if h satisfies some criterion for a set of partitions, but h' satisfies the criterion for strictly more partitions, there is an important sense in which h' is more fair than h. Clearly, if h satisfies a criterion for all partitions, there can be no "more fair" h' as far as the relevant criterion goes, and so it makes sense to consider h maximally fair in this sense. Despite finding these considerations quite persuasive, I think we should welcome the investigation of alternative notions of maximal fairness/fairness ideals.

Figure 1: Relations among Fairness Ideals for Non-Constant Assessors



perfect. Here, Perfection is defined as h predicting 1 or 0 for all individuals in the population without making any mistakes: h assigns 1 to those individuals with property y, 0 to those without it. Second, an assessor satisfies Equalized Odds for all partitions if and only if it is *perfectly non-discriminatory*. Here, Perfect Non-Discrimination means treating likes alike: all individuals alike in the sense of having or lacking property y are assigned the same score. Finally, an assessor satisfies predictive equity for all partitions if and only if it makes *perfect distinctions*. By Perfect Distinction, I mean that the assessor treats individuals who differ in the relevant way differently, namely, by not assigning individuals who differ with respect to the salient property y the same score. For non-constant assessors, that is, assessors that assign more than one score, these simple characterizing conditions stand in decreasing order of logical strength. I summarize these clean logical relationships (including those of an additional characterizing condition discussed below) in Figure 1. One potentially interesting fact about the logical relationships between these characterizing conditions is that, while Equalized Odds is logically independent of Calibration in general, it is committed to a strictly weaker fairness ideal. Moreover, and perhaps surprisingly, this fairness ideal is intermediate between that of Calibration and that of Predictive Equity.

Regardless of feasibility, it is difficult to object to the conception of ideal assessment as Perfection (cf. Hardt et al., 2016). It is the standard one might expect on a biblical judgment day. Perfection recognizes the moral importance of getting each individual's case right. It also represents ideal *accuracy* (which goes "hand in hand" with access to more individual-level data (Di Bello and Gong, 2023, p. 6)). In other words, Perfection, it seems to me, is *sufficient* for fairness in assessment. This is less true of some other ideals. The idea that fairness involves treating like cases alike and unalike cases differently is the formal equality principle and has an ancient pedigree (Aristotle, 2019, Book 5). The requirement to

treat like cases alike has also been emphasized, in different ways than here, in the literature on algorithmic fairness (Dwork et al., 2012; Zimmermann and Lee-Stronach, 2022). While there's obviously something compelling in the ideal that likes should be treated alike as it is formalized in Perfect Non-Discrimination, it seems equally obvious that something is lacking. A constant assessor or an assessor that assigns everyone with property y a lower score than everyone lacking it would count as maximally fair as far as non-discrimination goes. In other words, Perfect Non-Discrimination does *not* seem sufficient for fairness in assessment. On its own, then, Perfect Non-Discrimination is not a compelling conception of fairness. Similarly, while treating unalike cases unalike is implied by perfect assessment and so plausible, there are a lot of ways to satisfy Perfect Distinction that are intuitively unappealing. Simply assigning everyone different scores, for instance, would work even if the scores are not accurate or are systematically skewed in some way like all people from one racial group receiving very high scores while all other individuals receive very low scores. Again, Perfect Distinction by itself is not a fully compelling conception of fair treatment.

4 Roles for Ideals

The heart of the ideals program is the contention that ideals convey normatively relevant information about candidate statistical criteria of algorithmic fairness.[5] An account of the appropriate use of such information, however, could be developed in different ways. I will briefly canvas three possibilities here, essentially in order of increasing plausibility.

But first, to motivate the evaluative use of ideals, consider a loose analogy to a topic in Bayesian epistemology. That Bayesian agents expect to (eventually) converge to the truth and to (eventually) achieve consensus through inquiry with other sufficiently open-minded Bayesians are supposed to be good-making features of Bayesian theory. But one line of resistance objects to the asymptotic nature of these guarantees, often citing Keynes's famous dictum about the long run (e.g., Earman, 1992; Sprenger, 2018). Such guarantees offer little comfort about real-life, short-run cases that might be our focus. Others see the function of these guarantees differently, as minimal adequacy conditions or sanity checks for an inductive methodology. As Autzen writes, "Under the ideal scenario of an infinitely large data set, an inference procedure should show certain desirable features" (2017, p. 255). Ideals of statistical fairness criteria might function in an analogous fashion as minimal sanity checks. Inductive methodology should yield the right result—namely consensus or convergence—when agents follow it and have access to an infinite stream of evidence, which might be regarded as unrealistic in certain senses. Similarly, while ideals may correspond to remote or unrealistic cases, cases in which fairness is achieved for more groups than is realistically feasible, a sound account of fair assessment should likewise yield the right results—namely, attractive ideals—in at least those extreme cases.

[5] Stewart (2022) does not propose using ideals in this way to distinguish the normative plausibility of different candidate fairness criteria or of sets of them. Instead, I took the fact that a statistical criterion has an ideal that is difficult to achieve in practice as a strike against the ideal, and did not discuss significant distinctions in plausibility between the ideals. Here, I concede from the start that satisfying some criterion for all partitions is unrealistic, but the contention is that ideals provide information about the contours of the conceptions of fairness that the various criteria encode, and important distinctions between these criteria can be made. As far as I know, this perspective has not been suggested previously.

At least some readers of earlier versions of this essay have thought that the ideals program presupposes that statistical criteria of fairness are *both* necessary and sufficient conditions of fairness. But this is incorrect. Consider the analogy to long-run learning behavior in Bayesian theory again. Even if a sound inductive method should secure consensus in the long run—that is, under certain ideal conditions—it does not follow that merging of opinions is also *sufficient* to secure the rational status of some method. The method may be very unreasonable over shorter time horizons or when agents do not have access to the (perhaps tremendous) amount of information required for the merging theorems, for example. And similarly, for the three uses of ideals surveyed below. One might hold that any candidate necessary condition of fairness should have a sufficiently compelling ideal, or that a candidate set of sufficient conditions should, or that ideals can figure into cumulative cases for different criteria, without construing any of the particular criteria of the sort studied in the literature as sufficient by themselves.

4.1 Tests of Necessity for Individual Criteria

Most ambitiously, one could use ideals as adequacy conditions for candidate necessary criteria of algorithmic fairness. The standard view of the sorts of criteria under consideration in this essay is as necessary criteria for fairness. On this first proposal, then, in order for a candidate criterion to be a genuine necessary condition for algorithmic fairness, it must pass the test of having a sufficiently compelling ideal.[6] Put differently, it must "aim at" a plausible conception of fairness *in the ideal case*. In some ways, this is a natural thought. In response to the question, "why should we think of a statistical condition like Calibration as a *fairness* criterion?," one could point out that, in the ideal or limiting case, it reduces to a property that obviously concerns fairness and, arguably, suffices for fairness.

There are reasonable reservations about this role for ideals, reservations that, to me, seem decisive even. If we were to regard Calibration, for instance, as a necessary condition of fairness, it would follow that *all* logical consequences of Calibration are necessary conditions. But some of these consequences will be very weak—like making forecasts at all—and will not imply compelling ideals on their own. That is, making numerical forecasts does not imply an ideal that is sufficient for fair assessment. So it looks like using ideals to test for necessity is a dead end. Similar reservations arise outside of the case of logical implications of a given necessary condition. It could be, for example, that the ideal implied by each condition in a *set* of necessary conditions is not maximally fair in some pre-theoretic sense while the ideal implied by the set of criteria taken together is. (Since the ideals of prominent criteria that I discuss here stand in the logical relations displayed in Figure 1, it is very simple to identify the ideal to which any subset of these criteria is committed: it is the strongest ideal associated with a property in the set, the ideal highest up in Figure 1.) But this concern

[6]For the sake of clarity and simplicity, I'm describing things in very general terms here and eliding some nuance. There are different ways of construing the tests of necessity interpretation that may be more or less plausible. The one I'm considering requires that candidate criteria have sufficiently compelling ideals. An alternative test of necessity might be that a criterion's ideal is simply clearly concerned with some form of fair or equal treatment. It is plausible that all conditions considered in this essay pass this alternative, weaker test of necessity. This test, however, is still susceptible to the objections to the use of ideals as tests for necessity.

suggests the next possible role for ideals.

4.2 Tests of Insufficiency for Sets of Criteria

While statistical criteria of algorithmic fairness are standardly interpreted as individually necessary conditions for fairness, the ultimate aim is *to have fair algorithms*. In other words, the search is *ultimately* for jointly sufficient conditions for fairness. Can ideals help? As I mentioned above, it's hard to think of perfect assessment as leaving any room for unfairness. It is much less difficult to see how an assessor that is calibrated across a salient partition could be considered unfair (consider Table 1 again). So even if Calibration for all partitions is equivalent to perfect assessment, it is implausible that it is sufficient for fair assessment for a single partition in general. So having an ideal that is sufficient for fair assessment does not imply that the criterion itself is sufficient for fair assessment for a given partition. And since Perfection is quite a strong ideal—what would be a compelling way of strengthening it?—it is implausible that ideals can be used to identify sufficient conditions just as Perfection does not identify Calibration as sufficient for fairness. But maybe ideals can be used to diagnose insufficiency.

To use ideals to test for insufficiency of a set of criteria is to use them as necessary conditions, not for individual criteria exactly, but for candidate overall accounts of algorithmic fairness. The idea is that any account of algorithmic fairness, any set of conditions that are jointly sufficient, must imply a sufficiently compelling ideal—a condition, perhaps like Perfection, that is sufficient for fairness. Put another way, the sufficiency of a set of alleged sufficient conditions should be apparent in the extreme case of satisfaction for all partitions. Crucially, this role does not claim that if a set implies a sufficiently compelling ideal, then the set is sufficient for fairness. Again, one could find Perfection a compelling ideal but reject Calibration as sufficient for fairness. Instead, the proposal is that ideals help to eliminate criteria, or conjunctions of them, as candidate *sufficient* conditions for fairness. Maximal fairness—that is, fairness across *all* partitions—according to an alleged set of sufficient conditions for fairness for a single partition should be clearly and unambiguously fair. Consider once again the leading analogy to Bayesian theory. If an inductive method defined by some set of rules is put forward as a rational inductive method, but those rules fail to secure consensus in the long run (for sufficiently open-minded initial opinions) even on the assumption of an infinite, increasing stream of evidence, some would regard this as disqualifying for the proposed method's claim to rational status. Similarly, if a set of conditions that are supposed to be sufficient for fair assessment in some partition fail to imply a compelling ideal when imposed across all partitions, some might regard this as disqualifying for the account of fairness.

To illustrate this role of ideals, consider that while Calibration is inconsistent with Equalized Odds, the weaker Predictive Equity criterion is not. What is maximally fair assessment according to the conjunction of Equalized Odds and Predictive Equity? It is the formal equality principle as formalized by the conjunction of Perfect Non-Discrimination and Perfect Distinction; for non-constant assessors, this is equivalent to Perfect Non-Discrimination (Figure 1). While this ideal requires a non-constant assessor to assign the same score to everyone with property y and a different score to everyone without the property, the two scores are otherwise unconstrained. This leaves ample room for assessments that are not

intuitively fair. If that is right, the proposal under consideration would reject the sufficiency of the conjunction of Predictive Equity and Equalized Odds for fair assessment.

There are ways, however, of resisting this use of ideals as well. It could be that while Perfection is the correct standard in ideal cases, the best we can do in non-ideal cases requires more of a concession to practicalities, requires using criteria that move decidedly away from Perfection as their characterizing condition for when they are satisfied for all partitions. Consider the distinction between ideal and non-ideal theory that Sen, for example, has stressed. According to Sen, the identification of an ideal political state is neither necessary nor sufficient to make at least a great many social and political comparisons and choices (2010, pp. 15–16). It is not necessary since many such choices may be readily made without appeal to an ideal state. And the identification of an ideal is not sufficient because it may fail to inform us on how to compare non-ideal states that are feasible options, or how to move from our present state to a non-ideal but better alternative state. On Sen's view, ideal and non-ideal theory address different questions, with ideal theory being "of no direct relevance" to non-ideal theory (2010, pp. 16–17). Might one say the same about non-ideal algorithmic assessment, assessment in cases in which Perfection is not remotely feasible?[7] The situation with ideals of algorithmic fairness is not directly analogous. We are considering certain statistical criteria for fair assessment that are supposed to apply in non-ideal cases, namely, in a given, salient partition of the population. The ideals I identify are not disconnected from these criteria, are not addressing different questions, as Sen claims ideal and non-ideal political theory are. Instead, ideals are simply what maximally fair assessment looks like according to each criterion we have considered. They are defined in terms of the very criteria that are supposed to be applicable in non-ideal cases. We need make no assertion that ideals are the end of theorizing about fairness or that maximal fairness is realistically achievable. Rather, they can be interpreted as providing information about the normative content of the relevant criteria, as informing about how appropriate or compelling the various criteria might be. Still, to insist that the concern about concessions to practicalities can be dismissed outstrips the arguments I have at hand.

[7]The idea that different standards might be appropriate under ideal and non-ideal conditions finds sympathetic expression in the algorithmic fairness literature. For instance, in response to Hedden's alleged chance-based counterexample to most statistical criteria, Di Bello and Gong say this:

> One might argue that if a certain criterion of fairness is shown to be inapplicable under idealized conditions, then a fortiori the same criterion would be inapplicable under more realistic conditions. But this argument is too quick. Even if [...] predictive algorithms necessarily violate several performance criteria under idealized conditions and this violation is not intuitively unfair, the same violation under more realistic conditions may still count as unfair. For example, a predictive algorithm whose risk scores perfectly track the objective risks may count as intuitively fair even if it violates predictive parity. And yet, when the algorithm's risk score no longer track the objective risks, the algorithm need not be regarded as intuitively fair. Under more realistic conditions, the violation of classification parity may become morally problematic. (2023, p. 10)

4.3 Elements of Cumulative Cases

More minimally still, ideals can be taken as simply important considerations in formulating accounts of algorithmic fairness. A compelling ideal may be a consideration in favor of a putative criterion or set of criteria without by itself being dispositive of the status of criteria as necessary or sufficient conditions for fairness. While it is difficult to conceive of reasons for the wholesale rejection of this use of ideals, it is also difficult to specify details of the nature of cumulative cases in the abstract. It would be helpful to consider an example. In the next section, I consider a case study. Among other things, ideals provide a means of making certain subtle conceptual distinctions that may otherwise go unnoticed. My aim now is to substantiate this claim, extending the ideals program to an interesting, recently-proposed criterion as a case study. I identify the ideals of both Base Rate Tracking and Ratio Base Rate Tracking below. They are distinct. Moreover, as I point out, there are grounds for finding the ideal of Ratio Base Rate Tacking less compelling. This point could be used, depending on the role for ideals found most compelling, as a test of necessity, insufficiency, or just a consideration in a more encompassing case for preferring Base Rate Tracking to its ratio form.

5 Case Study: (Ratio) Base Rate Tracking and Ideal Fairness

Eva (2022) introduces a criterion he dubs *Base Rate Tracking*. The idea behind this condition is that the average score for a given group should not deviate from that group's base rate more or less than the average score deviates from the base rate in another group. Put differently, a group's average score has to be related to the prevalence of the property in the group in a uniform way.

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

Assigning groups with very similar base rates very different average scores clearly violates Base Rate Tracking. Stewart et al. establish that Base Rate Tracking, like Predictive Equity, is a way of weakening Calibration (2024, Proposition 1). Since Base Rate Tracking is a way of relaxing Calibration while maintaining some of its appeal as a fairness criterion, it is especially interesting in light of the impossibility theorems. Such results motivate the search for weaker criteria. Given that Base Rate Tracking weakens Calibration, it may be somewhat surprising that Base Rate Tracking, unlike Predictive Equity, retains a commitment to Perfection. Call a population *non-homogeneous* if there are $i, j \in N$ such that $Y(i) \neq Y(j)$. Non-homogeneity is itself a kind of non-triviality assumption about the relevant assessment exercise.

Theorem 1. *An assessor h for a non-homogeneous population N satisfies Base Rate Tracking for all partitions iff h is perfect.*

Table 2: Base Rate Tracking without Ratio Base Rate Tracking

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

(Rather than proving Theorem 1, I prove a generalization, Theorem 3, in the Appendix. The generalization extends Theorem 1 to cases in which Base Rate Tracking is *approximately* satisfied for all partitions.)

As Eva points out, the motivations he adduces for Base Rate Tracking do not pin down the functional form the criterion should take. Rather than defining the condition in terms of differences of expectations and base rates, we could use ratios (2022, p. 260).

> **Ratio Base Rate Tracking**. For an assessor h of N and any groups G, G' in a partition π of N, $\dfrac{E_G(h)}{\mu_G} = \dfrac{E_{G'}(h)}{\mu_{G'}}$ (whenever both ratios are defined).

On the one hand, it is clear that Base Rate Tracking and Ratio Base Rate Tracking are distinct. For example, Table 2 presents an assessor that satisfies Base Rate Tracking but does not satisfy Ratio Base Rate Tracking in the partition consisting of the two racial groups indicated. On the other hand, not only do Eva's stated motivations for the criteria fail to distinguish Base Rate Tracking from Ratio Base Rate Tracking, non-trivial consistency with Equalized Odds also fails to differentiate the two forms of the criterion. Ratio Base Rate Tracking like Base Rate Tracking is consistent with Equalized Odds only when the assessor is perfect or the base rates for all groups in π are identical (Stewart et al., 2024, Theorem and Proposition 2). Similarly, Hedden's argument against statistical criteria other than Calibration based on considerations having to do with objective chance would likewise fail to mark a normatively relevant distinction between Base Rate Tracking and Ratio Base Rate Tracking since, as consequences of Calibration, the argument does not apply to them. What sorts of reasons can be used to adjudicate between these alternatives?[8]

[8]There is a well-known, relevantly analogous issue in Bayesian epistemology, where a plurality of non-equivalent measures of confirmation have been proposed. Different functional forms of these measures failed to distinguished by relevant normative considerations, at least for some time (Eells and Fitelson, 2000). As Fitelson points out, "a great many of the arguments surrounding quantitative Bayesian confirmation theory are implicitly *sensitive to choice of measure of confirmation*. Such arguments are *enthymematic*, since they tacitly presuppose that certain relevance measures should be used (for various purposes) rather than other relevance measures that have been proposed and defended in the philosophical literature" (1999, p. S362). The "most standard" measure of confirmation is the difference between the posterior and prior—in analogy to Base Rate Tracking—and other "more or less standard" measures include the log of the ratio between posterior and prior—in analogy to Ratio Base Rate Tracking (Eells and Fitelson, 2000, p. 663). For those arguments that presuppose certain measures should be used rather than others, relevant, normative distinctions between measures must be made and considerations brought to bear on a comparative analysis. The same thing is true of statistical criteria for algorithmic fairness. The ideals program is an attempt making some relevant such distinctions.

Table 3: Criteria and Their Associated Fairness Ideals

Perfection	Calibration, Base Rate Tracking
Perfect Non-Discrimination$^+$	Ratio Base Rate Tracking
Perfect Non-Discrimination	Equalized Odds
Perfect Distinction	Predictive Equity

There *is* a distinction between these two forms of Base Rate Tracking in terms of ideal fairness according to each property that may be helpful. From the perspective of fairness ideals, we can see that Ratio Base Rate Tracking loses at least a bit of the content of Calibration that Base Rate Tracking retains. In addition to non-homogeneity, the following characterization result for Ratio Base Rate Tracking makes the further mild assumption that at least two individuals in the population have property y.

Theorem 2. *An assessor h for a non-homogeneous population N such that $|\{Y = 1\}| \geq 2$ satisfies Ratio Base Rate Tracking for all partitions iff h is perfectly non-discriminatory with $h(i) = 0$ whenever $Y(i) = 0$.*

According to Ratio Base Rate Tracking, an assessor that assigns all people with property y the same score and all people who lack property y a score of 0 is maximally fair. Call this ideal *Perfect Non-Discrimination$^+$*. I summarize the fairness ideals to which these prominent criteria of algorithmic fairness are committed in Table 3.

While Base Rate Tracking shares the compelling Perfection ideal with Calibration, Ratio Base Rate Tracking does not. Perfect Non-Discrimination$^+$, the ideal associated with Ratio Base Rate Tracking, is weaker. A uniform, very low score could be intuitively unfair to individuals with property y if loan or admissions decisions, for instance, are made on the basis of a threshold above that score. In fact, a constant assessor could assign all individuals a score of 0 without running afoul of Perfect Non-Discrimination$^+$. While this is not a conclusive consideration, pre-systematic judgment seems to find such an ideal does not suffice for fair assessment. At the very least, we have a potentially interesting normative distinction between the different forms of Base Rate Tracking in terms of their associated ideals.

Let's now consider applications of the three uses of ideals introduced above to the case of Base Rate Tracking and Ratio Base Rate Tracking. First, if a candidate necessary condition for algorithmic fairness must have a fully compelling ideal, this suggests that Ratio Base Rate Tracking, unlike Base Rate Tracking, is *not* a necessary condition of algorithmic fairness. (Notice that could be so even though the ideal associated with Base Rate Tracking implies the ideal associated with Ratio Base Rate Tracking. But as Table 2 shows, in general, Base Rate Tracking does not imply Ratio Base Rate Tracking.) Reservations about this general use of ideals were expressed above. For instance, Calibration shares an ideal with Base Rate Tracking and implies Ratio Base Rate Tracking. If Calibration is deemed a necessary condition of fairness, Ratio Base Rate Tracking would be, too, even though its ideal is less

than fully compelling.

Second, due to the monotonicity of logical consequence, adding further necessary conditions to a set that includes Base Rate Tracking will preserve a commitment to the ideal of Perfection for the set. So if we find Perfection a compelling ideal, there will not be promising ideal-based arguments against candidate sets of sufficient conditions that include Base Rate Tracking. The situation is different for Ratio Base Rate Tracking. If all of the ideals discussed here other than Perfection are deficient in some way, then *no* set of these conditions that excludes both Calibration and Base Rate Tracking is a candidate set of sufficient conditions for algorithmic fairness. This follows again from the relations displayed in Figure 1. The tenability of Ratio Base Rate Tracking, then, depends crucially on supplemental criteria in a way that the case for Base Rate Tracking does not. Admittedly, this is not a *decisive* consideration against Ratio Base Rate Tracking, but, on this use of ideals, there is a burden for an account starting from Ratio Base Rate Tracking that is automatically met by an account starting from Base Rate Tracking.

Third, even if one were to reject these first two uses, ideals provide one philosophically significant distinction between statistical fairness criteria including between Base Rate Tracking and Ratio Base Rate Tracking. This distinction may be just one piece of a complex case to be made for or against one of the properties. For example, if maximal fairness according to Base Rate Tracking is clearly fair, that may be a consideration in its favor. Conversely, to the extent that Ratio Base Rate Tracking's ideal is recognizably defective in some way, that may be a factor in a case against it. On a view that resists seeing ideals as a means of testing for the necessity or sufficiency of some set of statistical fairness criteria, how such considerations are to be weighed up ultimately may be very complex and involve a large number of other considerations. Or ideals could simply function as tie-breaking considerations when the cases for two criteria like Base Rate Tracking and Ratio Base Rate tracking are otherwise on a par. So even if the first two roles for ideals are rejected, ideals may have even critical roles to play in the normative evaluation of putative criteria of algorithmic fairness. Dismissing ideals on the basis of not finding the first two roles compelling would be hasty.

6 Conclusion

The morass of a rapidly expanding set of inconsistent statistical criteria of algorithmic fairness is a problem of both theoretical and practical significance. What's needed are systematic means of assessing the appropriateness of various criteria, of marking normatively relevant distinctions between candidate criteria of fairness. The ideals program, while unlikely to be the whole story, is an important part of such an enterprise. It is also an aspect of algorithmic fairness that, to the best of my knowledge, has received scant attention so far. The relevant ideals are much simpler to evaluate as standards of fair treatment than the associated statistical criteria. The challenge now is in specifying how to make use of these ideals in assessing their associated criteria. I canvassed three potential roles—as tests for necessity of individual criteria, tests for insufficiency of sets of criteria, and elements of cumulative cases—finding grounds to regard the latter two roles as more plausible than the first. How ideals might be brought to bear in comparing particular criteria can be seen in the discussion of Base Rate Tracking. Both Eva (2022) and Stewart et al. (2024) observe that the considerations in their

discussions fail to make normatively relevant distinctions between Base Rate Tracking and Ratio Base Rate Tracking. For the ideals program, such is not the case, as Theorems 1 and 2 show.[9] These initial fruits suggests further harvests for the ideals program.

Acknowledgments. Thanks to Michael Nielsen, Tom Sterkenburg, Reuben Stern, and two anonymous referees for helpful comments on earlier drafts.

Data Availability Statement. No associated data for this manuscript.

Appendix

Proof of Theorem 2

Proof. Let h be an assessor for a non-homogenous population N with $|\{Y = 1\}| \geq 2$.

($\Rightarrow$). I note first that the special assumptions about the population are unnecessary for establishing that $h(i) = h(j)$ for all $i, j \in \{Y = 1\}$. It suffices to consider the partition of singletons. For $i, j \in \{Y = 1\}$, we have $\mu_{\{i\}} = \mu_{\{j\}}$, so Ratio Base Rate Tracking implies that $h(i) = E_{\{i\}}(h) = E_{\{j\}}(h) = h(j)$.

Next, since $|\{Y = 1\}| \geq 2$ and N is non-homogeneous, there is a partition

$$\pi = \{\{i\}, N \setminus \{i\}\},$$

where $\mu_{\{i\}}, \mu_{N\setminus\{i\}} > 0$. Using Ratio Base Rate tracking, the law of total expectation, and the claim established above that $h(i) = h(j)$ for all $i, j \in \{Y = 1\}$,

$$\begin{aligned} h(i) &= \frac{E_{\{i\}}(h)}{\mu_{\{i\}}} \\ &= \frac{E_{N\setminus\{i\}}(h)}{\mu_{N\setminus\{i\}}} \\ &= \frac{E_{N\setminus\{i\}}(h|Y=1))\mu_{N\setminus\{i\}} + E_{N\setminus\{i\}}(h|Y=0))(1-\mu_{N\setminus\{i\}})}{\mu_{N\setminus\{i\}}} \\ &= \frac{h(i)\mu_{N\setminus\{i\}} + E_{N\setminus\{i\}}(h|Y=0))(1-\mu_{N\setminus\{i\}})}{\mu_{N\setminus\{i\}}}. \end{aligned} \tag{1}$$

Equation 1 implies that $E_{N\setminus\{i\}}(h|Y = 0))(1 - \mu_{N\setminus\{i\}}) = 0$. Since $0 < \mu_{N\setminus\{i\}} < 1$ by the assumptions about the population, it follows that $E_{N\setminus\{i\}}(h|Y = 0)) = 0$. This establishes that $h(j) = 0$ for all $j \in \{Y = 0\}$. Hence, if h satisfies Ratio Base Rate Tracking for all partitions of N, then h is perfectly non-discriminatory and assigns score 0 to all individuals without property y.

($\Leftarrow$). Suppose that there is some $r \in [0, 1]$ such that, for all i such that $Y(i) = 1$, we have $h(i) = r$, and for all i such that $Y(i) = 0$, we have $h(i) = 0$. Let G be any group in a

[9]The ideals program has also been recently exploited in (Nielsen and Stewart, 2024).

partition π of N. If $\mu_G = 0$, then $\frac{E_G(h)}{\mu_G}$ is undefined. So suppose that $\mu_G > 0$. Then, using the law of total expectation and the assumption,

$$\begin{aligned} E_G(h) &= E_G(h|Y=1)\mu_G + E_G(h|Y=0)(1-\mu_G) \\ &= E_G(h|Y=1)\mu_G + 0(1-\mu_G) \\ &= E_G(h|Y+1)\mu_G. \end{aligned} \tag{2}$$

By 2 and the assumption,

$$\frac{E_G(h)}{\mu_G} = \frac{E_G(h|Y=1)\mu_G}{\mu_G} = r. \tag{3}$$

Since G is arbitrary, 3 implies that for any $G' \in \pi$ such that $\frac{E_{G'}(h)}{\mu_{G'}}$ is defined, we have $\frac{E_G(h)}{\mu_G} = \frac{E_{G'}(h)}{\mu_{G'}} = r$. Hence, h satisfies Ratio Base Rate Tracking. □

Proof of Theorem 3

Say that h approximately tracks base rates (with respect to a specified margin of error $\varepsilon \geq 0$) in a partition π if, for any cells G, G' of π,

$$|(E_G(h) - \mu_G) - (E_{G'}(h) - \mu_{G'})| \leq \varepsilon.$$

Approximate base rate tracking requires that the magnitude by which the expected score deviates from the base rate is, not necessarily the same for all groups, but bounded by some margin of error ε. An assessor is approximately perfect (for a specified margin of error $\varepsilon \geq 0$) if $|h(i) - Y(i)| \leq \varepsilon$ for all $i \in N$.

Theorem 3. *Let h be an assessor for a non-homogeneous population N. If h is ε-approximately base rate tracking for all partitions, then h is ε-approximately perfect. And if h is ε-approximately perfect, then h is 2ε-approximately base rate tracking for all partitions.*

Proof. Let h be an assessor for a non-homogeneous population N.

($\Rightarrow$). Suppose that h is approximately base rate tracking for all partitions. That is, for a given $\varepsilon \geq 0$, and for all groups G, G',

$$|(E_G(h) - \mu_G) - (E_{G'}(h) - \mu_{G'})| \leq \varepsilon. \tag{4}$$

In particular, h approximately tracks the base rates in the finest partition $\{\{1\}, \{2\}, \ldots, \{n\}\}$. Substituting in 4, we have, for any $i, j \in N$,

$$|(h(i) - Y(i)) - (h(j) - Y(j))| \leq \varepsilon. \tag{5}$$

Since N is non-homogeneous, there are i, j such that $Y(i) = 1$ and $Y(j) = 0$. For such i, j, we can rewrite 5 as

$$|(h(i) - 1) - (h(j) - 0)| \leq \varepsilon. \tag{6}$$

If $|h(i) - 1| > \varepsilon$, then $h(i) - 1 < -\varepsilon$. Since $h(j) \geq 0$, $h(i) - 1 - h(j) < -\varepsilon$. Hence, $|(h(i) - 1) - (h(j) - 0)| > \varepsilon$, contradicting 6. It follows that $|h(i) - 1| = |h(i) - Y(i)| \leq \varepsilon$.

Similarly, if $h(j) > \varepsilon$, then since $h(i) - 1 \leq 0$, we have $(h(i) - 1) - h(j) < -\varepsilon$. Hence, $|(h(i)-1)-(h(j)-0)| > \varepsilon$, contradicting 6. It follows that $h(j) = |h(j)-0| = |h(j)-Y(j)| \leq \varepsilon$. Given that the population is non-homogeneous, for any $i \in N$ there is a $j \in N$ such that $Y(i) \neq Y(j)$, and the foregoing reasoning can be repeated using j and the finest partition to establish $|h(i) - Y(i)| \leq \varepsilon$.

($\Leftarrow$). Suppose that h is ε-perfect. That is, for any $i \in N$,

$$-\varepsilon \leq h(i) - Y(i) \leq \varepsilon. \tag{7}$$

For any non-empty $G \subseteq N$, it follows that

$$-\varepsilon \leq E_G(h - Y) \leq \varepsilon, \tag{8}$$

that is, the average value of $h - Y$ in G lies between $-\varepsilon$ and ε. But by the linearity of expectation,

$$\begin{aligned} E_G(h - Y) &= E_G(h) - E_G(Y) \\ &= E_G(h) - P_G(Y = 1) \\ &= E_G(h) - \mu_G. \end{aligned} \tag{9}$$

From 8 and 9, we have $|E_G(h) - \mu_G| \leq \varepsilon$. Since G was arbitrary, it follows that, for any partition π and any $G, G' \in \pi$, $|(E_G(h) - \mu_G) - (E_{G'}(h) - \mu_{G'})| \leq 2\varepsilon$. This establishes that h is 2ε-approximately base rate tracking in all partitions. □

Remark 1. Since one can find simple examples of ε-approximately perfect assessors that are exactly 2ε-approximately base rate tracking, this upper bound is tight.

Remark 2. Observe that for $\varepsilon = 0$, Theorem 3 reduces to Theorem 1.

Remark 3. While Calibration and Base Rate Tracking share the same ideal, their relationships to Perfection are not the same in the approximate case. Compare Theorem 3 to Observation 1′ in (Stewart, 2022).

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`.

Aristotle (2019). *Nicomachean Ethics* (Third ed.). Cambridge: Hackett Publishing.

Autzen, B. (2017). Bayesian convergence and the fair-balance paradox. *Erkenntnis 83*(2), 253–263.

Beigang, F. (2022). On the advantages of distinguishing between predictive and allocative fairness in algorithmic decision-making. *Minds and Machines 32*(4), 655–682.

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

Di Bello, M. and R. Gong (2023). Informational richness and its impact on algorithmic fairness. *Philosophical Studies*, 1–29.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226.

Earman, J. (1992). *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.

Edenberg, E. and A. Wood (2023). Disambiguating algorithmic bias: From neutrality to justice. In F. Rossi, S. Das, J. Davis, K. Firth-Butterfield, and A. John (Eds.), *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 691–704.

Eells, E. and B. Fitelson (2000). Measuring confirmation and evidence. *The Journal of Philosophy 97*(12), 663–672.

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

Fitelson, B. (1999). The plurality of bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of science 66*(S3), S362–S378.

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

Hardt, M., E. Price, and N. Srebro (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems 29*.

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

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.

Lazar, S. and J. Stone (2023). On the site of predictive justice. *Noûs*.

Loi, M., A. Herlitz, and H. Heidari (2023). Fair equality of chances for prediction-based decisions. *Economics & Philosophy*, 1–24.

Narayanan, A. (2018). 21 fairness definitions and their politics (tutorial). *Conference on Fairness, Accountability & Transparency*, `https://www.youtube.com/watch?v=jIXIuYdnyyk`.

Nielsen, M. and R. T. Stewart (2024). New possibilities for algorithmic fairness. *Philosophy & Technology 37*(116).

Sen, A. (2010). *The Idea of Justice*. London: Penguin.

Sprenger, J. (2018). The objectivity of subjective bayesianism. *European Journal for Philosophy of Science 8*(3), 539–558.

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

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

Verma, S. and J. Rubin (2018). Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pp. 1–7.

Zimmermann, A. and C. Lee-Stronach (2022). Proceed with caution. *Canadian Journal of Philosophy 52*(1), 6–25.