Algorithmic Fairness and Base Rate Tracking

Abstract

In the last few years, machine learning researchers have proposed a plethora of prospective 'statistical criteria of algorithmic fairness', i.e. purely statistical necessary conditions that a predictive algorithm's predictions must satisfy in order for the algorithm to count as fair. However, a mixture of formal no-go theorems and devastating counterexamples have served to undermine the philosophical credibility of almost all of these conditions. Only one statistical criterion retains anything like universal support, namely *calibration within groups*. In this paper, I (i) argue that calibration within groups is neither a necessary nor a sufficient condition for algorithmic fairness, (ii) propose, motivate and defend a novel criterion, called 'base rate tracking', which evades the theorems and counterexamples that undermined existing criteria and allows us to accurately diagnose and quantify many paradigmatic instances of algorithmic unfairness, and (iii) reevaluate the proper role of statistical criteria of algorithmic fairness in the project of ensuring the fair and equitable application of predictive algorithms in society.

1 Introduction

Suppose you want to buy a home. In order to achieve your goal, you first need to be approved for a mortgage, and the success of your application will be determined by a predictive algorithm whose internal workings you know nothing about. You feel good about your chances, since your credit history is strong and your income is reasonably high in relation to the price of the property you hope to purchase. Sadly, your application is denied, for reasons known only to the algorithm (and maybe its designers). As a member of a disadvantaged group, you suspect that the decision was not entirely fair. You subsequently learn that many other people from the same

disadvantaged group have had similar experiences with the same algorithm, and your suspicion grows stronger. You decide to investigate the algorithm further in order to determine whether it really is operating unfairly. But since the algorithm is proprietary, you are not able to examine its internal operations. Furthermore, you do not have any information about the character or motivations of the people who designed the algorithm. The one thing that you do have access to is data describing the details of all the cases to which the algorithm has been applied so far, and the verdicts that the algorithm gave for those cases. You aim to evaluate the fairness of the algorithm on the basis of this data alone. In order to do so, you need to make use of *statistical criteria of algorithmic fairness*, i.e. purely statistical criteria which specify necessary conditions that must be satisfied by an algorithm's predictions in order for the algorithm to count as fair. For instance, you might employ the popular criterion that, in order to count as fair, the algorithm should not yield more false positives for one group than it does for another (see e.g. Angwin et al (2016), Hardt et al (2016)).[1] As it turns out though, the existence and character of statistical criteria of algorithmic fairness is the topic of sustained and ongoing disagreement, and the criterion I mentioned above (amongst others), has recently been the target of some powerful objections and impossibility results (see e.g. Chouldechova (2017), Corbett-Davies and Goel (2018), Kleinberg et al (2016), Long(manuscript), Miconi (2017)). In fact, Hedden (2021) has recently presented a counterexample which seems to simultaneously refute 10 of the 11 most influential criteria from the literature on algorithmic fairness. As a result, it's far from clear exactly which criteria you should employ when evaluating the fairness of the suspect lending algorithm. In this article, I will present, motivate and defend a novel statistical criterion of algorithmic fairness that is both resistant to Hedden's counterexample, and well equipped to accurately diagnose unfairness in cases of the kind described above.

More precisely, the plan is as follows. In Section 2 I begin by briefly reviewing 11 of the most influential statistical criteria of algorithmic fairness, before recalling

[1] In the present example, a false positive would be a case where the algorithm approves an application where it shouldn't have (where the applicant subsequently defaults on their mortgage).

Hedden's (2021) counterexample, which simultaneously undermines 10 of these 11 criteria. In Section 3, I turn to evaluating the proper formulation and inherent limitations of the 1 surviving criterion, *calibration within groups*, and argue that, even in its most plausible formulation, it is neither a necessary nor a sufficient condition for algorithmic fairness. Section 4 is devoted to introducing and motivating a novel criterion, *base rate tracking*, which evades Hedden's counterexample and allows us to diagnose many instances of algorithmic unfairness to which the calibration within groups criterion is blind. Finally, Section 5 considers whether base rate tracking should be buttressed by any further statistical criteria and draws some general morals regarding the role of statistical criteria in ensuring the fair application of predictive algorithms in society.

2 Extant Criteria

Before introducing 11 extant candidate statistical criteria of algorithmic fairness, note that predictive algorithms can be partitioned into (at least) two kinds, depending on the type of verdict that they yield. Firstly, binary classification algorithms output binary verdicts such as 'approve/deny loan', 'grant/deny parole', 'classify as high/low risk driver' etc. Typically, one of the two possible outcomes of a binary classification will have positive valence (e.g. 'approve loan', 'classify as low risk', 'grant parole') and one will have negative valence (e.g. 'deny loan', 'classify as high risk', 'deny parole'). Secondly, continuous risk scoring algorithms output numerical risk scores that are intended to estimate a subject's risk of exhibiting a certain kind of behaviour. For instance, such an algorithm might assign a subject a risk score that estimates the risk of that subject defaulting on their mortgage payments. A decision on whether to approve a mortgage for that agent will then be made on the basis of their assigned risk score. So whereas binary classification algorithms generally yield direct recommendations (e.g. 'approve/deny mortgage'), continuous risk scoring algorithms yield risk scores that do not directly entail explicit recommendations, although they can easily be made to do so if one imposes some

kind of risk threshold rule of the form 'approve mortgages for all and only those subjects with risk score less than *x*', for some fixed *x* (see e.g. (Corbett-Davies and Goel (2018), Long (manuscript)) for defences of such threshold rules). For the purposes of this essay, we will be focusing specifically on the fairness of the judgements (e.g. classifications, risk scores and predictions) produced by predictive algorithms, rather than the fairness of the decisions that are taken on the basis of those judgements.[2]

When we're considering possible statistical criteria of algorithmic fairness, it is important to distinguish between crtieria used to evaluate binary classification algorithms on the one hand, and criteria used to evaluate numerical risk scoring algorithms on the other. We begin (following Hedden (2021)) by recalling three influential criteria that have been proposed as necessary conditions for the fairness of numerical risk scoring algorithms.

2.1 Criteria for Numerical Risk Scoring Algorithms

(1) Calibration Within Groups: For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.[3]

(2) Balance for the Positive Class: The (expected) average risk score assigned to those individuals who are actually positive is the same for each relevant group.

(3) Balance for the Negative Class: The (expected) average risk score assigned to those individuals who are actually negative is the same for each relevant group.

[2] I take these two issues to be conceptually distinct, although they are clearly intimately related.

[3] Hedden (2021) formulates the candidate criteria in terms of expectation values rather than actual frequencies, 'since an algorithm can satisfy the expectational version but violate its actual relative frequency-based analogue simply due to the vagaries of chance.' (Hedden, 2021: 6), and I am happy to follow that convention for the same reason. Generally, I think of these expectation values as being computed relative to objective probability functions that encode either physical chances, long run frequencies, or the subjective probabilities of some suitably idealised observer.

The motivations for these three criteria are straightforward and intuitively compelling. (1) requires that any given risk score should mean the same thing (have the same evidential significance) for all groups. By way of illustration, consider an insurance pricing algorithm that assigns subjects risk scores encoding their estimated risk of being involved in a car accident. Now imagine that 25% of all the white drivers with a risk score of 25% were involved in an accident, while only 10% of black drivers with a risk score of 25% were involved in an accident. Then it seems that the risk score 25% has different meanings for white and black drivers, since a white driver with that score is more likely to be involved in an accident than a black driver with the same score. This is exactly the kind of situation that is deemed unfair by (1). Sticking with the insurance example, (2) simply requires that in order to count as fair, the algorithms should assign the same average risk score to the set of all white drivers who actually ended up being involved in accidents as it does to the set of all black drivers who actually ended up being involved in accidents. To see the motivation for this, imagine that, on average, white drivers who were actually involved in accidents had lower risk scores than black drivers who were actually involved in accidents. That would seem to imply that the algorithm was more likely to give an 'unsafe driver' (a driver that ended up being involved in an accident) a low risk score if they were white, and that seems unfair.

Conversely, (3) requires that the algorithm should assign the same average risk score to the set of all white drivers who didn't actually end up being involved in accidents as it does to the set of all black drivers who didn't actually end up being involved in accidents. Again, if the algorithm violated (3), that would mean that the algorithm was more likely to give a 'safe driver' (a driver that did not end up being involved in an accident) a high risk score if they were black, and that seems unfair.

2.2 Criteria for Binary Classification Algorithms

The remaining seven criteria all impose necessary conditions for the fairness of binary classification algorithms.

(4) Equal False Positive Rates: The (expected) percentage of actually negative individuals who are falsely predicted to be positive is the same for each relevant group.

(5) Equal False Negative Rates: The (expected) percentage of actually positive individuals who are falsely predicted to be negative is the same for each relevant group.

(6) Equal Positive Predictive Value: The (expected) percentage of individuals predicted to be positive who are actually positive is the same for each relevant group.

(7) Equal Negative Predictive Value: The (expected) percentage of individuals predicted to be negative who are actually negative is the same for each relevant group.

(8) Equal Ratios of False Positive Rate to False Negative Rate: The (expected) ratio of the false positive rate to the false negative rate is the same for each relevant group.

(9) Equal Overall Error Rates: The (expectation of) the number of false positives and false negatives, divided by the number of individuals, is the same for each relevant group.

(10) Statistical Parity: The (expected) percentage of individuals predicted to be positive is the same for each relevant group.

(11) Equal Ratios of Predicted Positives to Actual Positives: The (expectation of) the number of individuals predicted to be positive, divided by the number of individuals who are actually positive, is the same for each relevant group.

Again, the basic philosophical motivations behind most of these criteria seem quite robust at first glance. Sticking with the insurance pricing example, we can imagine now that, as well as a numerical risk scoring algorithm that assigns subjects quantitative risk scores, we also have a binary classification algorithm that simply attempts to determine whether or not drivers are 'high risk' (as opposed to 'low

risk'). Criteria (4) and (5) are directly analogous to (2) and (3),[4]respectively, although they are formulated in terms of binary predictions rather than numerical risk scores. While (2) required that safe black drivers should not be more likely to be assigned a high risk score than safe white drivers, (4) requires that safe black drivers should not be more likely to be designated as 'high risk' than safe white drivers. Similarly, while (3) required that unsafe white drivers should not be more likely to be assigned a low risk score than unsafe black drivers, (5) requires that unsafe white drivers should not be more likely to be designated as 'low risk' than unsafe black drivers.

Just as (4) and (5) are directly analogous to (2) and (3), both (6) and (7) are directly analogous to (1). While (1) required that each possible risk score have the same evidential import for all groups, (6) and (7) likewise require that both possible binary predictions have the same evidential import for all groups. For instance, if 30% of white subjects that are classed as 'high risk' by our binary algorithm are actually involved in accidents while only 20% of black subjects that are classed as 'high risk' are actually involved in accidents, that would suggest that the prediction 'high risk' means different things (has different evidential import) for white drivers and black drivers, and that seems unfair.

(8) is motivated by the idea that predictive errors should lean in the same direction for all groups. For instance, if white drivers had more false negatives than false positives, but black drivers had more false positives than false negatives, that would mean that the algorithm was erring on the side of caution (in the sense of classifying many safe drivers as unsafe) for black drivers, but erring on the side of risk (in the sense of classifying many unsafe drivers as safe) for white drivers. That would seem unfair. (9) simply requires that the algorithm should be equally accurate for all groups. Although violations of this condition do not obviously imply that one group is being systematically mistreated in comparison to another, it obviously seems desirable that our algorithms should be equally accurate for all groups.

[4] Pleiss et al (2017) refer to the average risk scores referenced in (2) and (3) the 'generalised false negative' and 'generalised false negative' rates, respectively.

(10) stipulates that the percentage of subjects predicted to be positive should be the same for each group, i.e. the percentage of black drivers that are classed as 'high risk' should be equal to the percentage of white drivers that are classed as 'high risk'. Note that (10) will be violated by an optimal predictive algorithm whenever the base rates of the groups are not equal. (11) generalises (10) in the sense that satisfaction of (11) entails satisfaction of (10) when the base rates are equal, but allows for the fairness of optimal predictive algorithms when base rates are not equal.

(1)–(11) represent the most influential and widely discussed statistical criteria of algorithmic fairness from the literature. As we've seen, the majority of them are motivated by prima-facie compelling philosophical intuitions about the nature of fairness. If we were able to accept even a decent subset of these criteria as genuine necessary conditions for algorithmic fairness, then we would have access to powerful diagnostic tools that would allow us to evaluate the fairness of proprietary algorithms whose inner workings and design processes are often opaque and mysterious. As it turns out though, there is good reason to think that the vast majority of these criteria aren't really necessary conditions for algorithmic fairness at all. First of all, there exist a number of famous impossibility results which show that various combinations of these 11 conditions can only be jointly satisfied in unrealistic and trivial special cases (see e.g. Chouldechova (2017), Kleinberg et al (2016), Miconi (2017)). Secondly, and more pertinently for present purposes, Hedden (2021) has provided an example of an obviously fair algorithm that simultaneously violates 10 of the 11 criteria, thereby refuting their claim to constitute necessary conditions for algorithmic fairness. I turn now to briefly reviewing this example.

Suppose that there exist two rooms, A and B, each containing 20 people. Of the room A people, 12 are each assigned coins with a bias of $\frac{3}{4}$ and the remaining 8 are each assigned coins with a bias of $\frac{1}{8}$. Of the room B people, 10 are each assigned coins with a bias of $\frac{3}{5}$ and 10 are each assigned coins with a bias of $\frac{2}{5}$. We want to design an algorithm that predicts, for each person, whether their coin will land heads when tossed. Of course, the best (most predictively accurate) possible numerical risk scoring algorithm is the one that assigns each person a risk score equal to the bias of

their coin. The best possible binary classification algorithm is the one that assigns each person whose coin has a bias greater than $\frac{1}{2}$ a prediction of 'heads' and assigns everyone else a prediction of 'tails'. Apart from being predicatively optimal, these algorithms are perfectly fair. Nobody could plausibly object to the application of algorithms such as these on the ground that they treat one group unfairly in comparison to another. But, as Hedden notes, they violate 10 of the 11 criteria listed above. For instance, the false positive and false negative rates for room A are 3/10 and 1/10, respectively, while the corresponding rates for room B are both 4/10, which entails a violation of criteria (4) and (5). The only prospective criterion that is satisfied by either of these algorithms is (1), calibration within groups, which is satisfied by the numerical risk scoring algorithm. The algorithm assigns each individual a risk score equal to the bias of their assigned coin, regardless of their group. This entails that every risk score has the same evidential import across both groups, which is all that is required by (1). In sum, this example clearly demonstrates that only the first of the eleven prospective statistical criteria of algorithmic fairness is plausibly a necessary condition for an algorithm to count as fair.

3 Calibrating Calibration

At this stage then, calibration within groups is the only candidate statistical criterion of algorithmic fairness that we have good grounds to endorse. This raises a number of questions. Most pertinently, one is compelled to ask whether calibration within groups might actually be both a necessary *and* a sufficient condition for algorithmic fairness. And if the answer to this first question turns out to be 'no' (as it will), one will also be compelled to ask whether we can identify any further statistical criteria of algorithmic fairness that help us to accurately diagnose injustice in the application of predictive algorithms whose design processes and inner workings may be completely unknown. In this section, I will actually argue that calibration within groups is neither a necessary nor a sufficient condition for algorithmic fairness. But before doing this, it will be useful to pause briefly to consider the proper formulation

of the criterion. Towards this end, recall that, as stated in (1), calibration within groups requires

Calibration Within Groups (Strong): For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group and is equal to that risk score.

This formulation of the condition can be contrasted with the following, logically weaker formulation,

Calibration Within Groups (Weak): For each possible risk score, the (expected) percentage of individuals assigned that risk score who are actually positive is the same for each relevant group.

Like the strong formulation, the weak formulation requires that every possible risk score should have the same evidential import for all relevant groups in order for the algorithm to count as fair. In the insurance example, it requires that if 10% of the white drivers assigned a risk score of 10% actually end up in accidents, then it should also be the case that 10% of the black drivers assigned that risk score end up in accidents. The difference between the formulations is that the strong formulation imposes a stringent constraint on the accuracy of the algorithm: that the proportion of subjects from any group that are assigned a given risk score should actually be equal to that risk score. For instance, if 9% of the white drivers who are assigned a risk score of 10% actually get involved in accidents, then the strong formulation will deem the algorithm to be unfair, even if it's also the case 9% of those drivers from all other relevant groups who are assigned a risk score of 10% get involved in accidents. In essence, the weak formulation requires that the risk score have the same evidential import for all groups, but doesn't impose any further restrictions on how risk scores relate to actual or expected frequencies. In contrast, the strong

formulation poses a non-relational constraint on the accuracy of the risk scores. Here's an example which shows that the strong formulation is too strong.

Suppose again that there are two rooms, A and B, containing 10 people each. All people are assigned 2 coins, the first of which is a fair coin with a known bias of $\frac{1}{2}$. The second coins have unknown biases that are not available to the algorithm. As it turns out, the biases of the second coins are all $\frac{3}{5}$. The algorithm aims to predict whether both of a subject's two coins will land heads when flipped. Since the biases of the second coins are not available to the algorithm, it operates by assuming that all the second coins have a uniform bias of $\frac{1}{2}$ and then assigns each subject a risk score equal to the products of the biases of their two coins, i.e. $\frac{1}{2} * \frac{1}{2} = \frac{1}{4}$. I think this is obviously fair. The algorithm assigns everyone from both groups the same risk score on the basis of the same evidence. And indeed, the algorithm trivially satisfies the weak formulation of calibration within groups. The only risk score assigned by the algorithm is $\frac{1}{4}$ and the proportion of Room A people assigned this score whose coins both land heads is $\frac{3}{10}$, which is equal to the proportion of Room B people assigned the score whose coins both land heads. However, the algorithm also violates the strong formulation of calibration within groups, since the expected proportion of people from either room assigned the risk score $\frac{1}{4}$ who actually tossed two heads ($\frac{3}{10}$) is not equal to that risk score.

I take this example to show that only the weaker formulation of calibration within groups is plausibly a necessary condition for algorithmic fairness. While I agree that the above algorithm is non-ideal in the sense that it systematically underestimates the risk of agents tossing two heads, I also think it's clear that this shortcoming is not helpfully described as 'unfairness'. If one insists on calling this kind of shortcoming 'unfair', then it's clear that we need to distinguish between two conceptions of algorithmic unfairness: one that applies to uniform failings of accuracy that do not track divisions between groups, and one that manifests itself in inequitable differences in the way that different groups are treated by the algorithm. I think it's apparent that the first conception is not the target of extant investigations into the

nature of algorithmic fairness, and hence that it can legitimately be bracketed in subsequent discussion. Overall, the above example makes it obvious that only the weaker formulation of calibration within groups can plausibly be considered as a candidate necessary condition for algorithmic fairness, since the stronger formulation depends on features of an an algorithm's predictions that are not directly relevant to its fairness.

3.1 The Non-Necessity of Calibration

But even under this weaker and more plausible formulation, calibration within groups fails as a necessary condition for algorithmic fairness. To see why, consider the following insurance pricing algorithm, which assigns risk scores to drivers on the basis of their credit scores.

Age	**Credit Score**	**Base Rate**	**Risk Score**
Young	Good	$\frac{1}{30}$	$\frac{1}{20}$
Young	Bad	$\frac{1}{30}$	$\frac{1}{10}$
Old	Good	$\frac{1}{40}$	$\frac{1}{20}$
Old	Bad	$\frac{1}{20}$	$\frac{1}{10}$

On average, 3/80 young drivers are involved in accidents, regardless of their credit scores, while $\frac{1}{20}$th of older drivers with bad credit scores are involved in accidents, compared to only $\frac{1}{40}$th of those with good credit scores. The algorithm simply assigns risk scores of $\frac{1}{20}$ to all drivers with good credit scores, and $\frac{1}{10}$ to drivers with bad credit scores. For simplicity, assume that the algorithm is applied to an equal number of drivers from each of the four profiles, which implies that young drivers and old drivers both have an overall base rate of 3/80. Then the algorithm violates calibration within groups, since the base rate for young drivers with a risk score of $\frac{1}{20}$ is $\frac{1}{30}$ while the base rate for old drivers with the same risk score is $\frac{1}{40}$, which means that the risk score $\frac{1}{20}$ has different evidential implications for young drivers than it does for older drivers. However, it seems wrong to say that the algorithm treats older drivers unfairly in comparison to young drivers. For, while older drivers

with a risk score of $\frac{1}{20}$ are actually less risky than their younger counterparts, the converse is true for older drivers with a risk score of $\frac{1}{10}$, who have a higher true risk ($\frac{1}{20}$) than their younger counterparts ($\frac{1}{30}$). The algorithm does not systematically treat younger drivers more favourably than older drivers or vice versa. On balance, it gives them equal treatment, evinced by the fact that the average risk score for both groups is 3/40, equal to half the overall base rates for both groups. Calibration within groups says that the algorithm treats old drivers unfairly in comparison to young drivers, but that is clearly not correct in this case. Neither group is systematically preferred to the other.

One might be tempted to reply that the algorithm is still unfair, even if it does not treat old drivers unfairly in comparison to young drivers, since it treats old drivers with good credit scores unfairly in comparison to young drivers with good credit scores. I argued that young drivers are not treated unfairly in comparison to old drivers because the two failures of calibration evened each other out – old drivers with good credit scores are treated unfairly compared to young drivers with good credit scores, but young drivers with bad credit scores are treated unfairly compared to old drivers with bad credit scores. But this doesn't change the fact that old drivers with good credit scores are treated unfairly compared to young drivers with good credit scores (since they have a lower base rate but the same risk score), which suggests that the algorithm is still unfair, even if that unfairness doesn't stem from an overall age bias.

In response to this objection, it is important to draw a distinction between two distinct possible interpretations of the calibration within groups criterion. Firstly, one can interpret the criterion as a diagnostic tool for identifying whether an algorithm treats some specific groups unfairly in comparison to some others. On this interpretation the criterion can be used to check whether the pricing algorithm above treats young drivers unfairly in comparison to old drivers, for instance.

And as we've just seen, the criterion gives an intuitively incorrect verdict here, since it identifies

age bias where there does not seem to be any.[5] Secondly, one can interpret the criterion as a more coarse grained diagnostic tool that simply helps to identify whether the algorithm is unfair *overall*. On this interpretation, the algorithm is unfair just in case it is possible to identify any groups with respect to which the calibration criterion is violated. The insurance pricing algorithm described above is not necessarily a counterexample to this interpretation of the criterion, since one could argue that the algorithm is unfair overall (on the grounds that it treats old drivers with good credit scores unfairly compared to young drivers with good credit scores), and there are multiple failures of calibration. But adopting the second interpretation doesn't solve all of our problems. Firstly, note that, even if calibration within groups is a necessary condition for an algorithm being fair overall, it is still desirable to have access to more fine grained criteria that allow us to identify not only whether an algorithm is biased in general, but also which groups are being treated unfairly in comparison to which other groups, since this information is clearly crucial to understanding and addressing the unfairness. Secondly, I am skeptical of the idea that we should treat all violations of calibration as conclusive evidence of injustice. For instance, one can imagine an algorithm that is calibrated with respect to age, gender, race, education, income, nationality, zip code, sexual orientation and political and religious beliefs, but that is not calibrated with respect to whether someone lives in an odd or even numbered house. In this case, it might be right to say that the algorithm treats even dwellers unfairly in comparison to odd dwellers, but that doesn't seem like a good reason to simply dismiss the algorithm as 'unfair'.

[5] I'll briefly address another possible criticism here. One could imagine a case in which an algorithm gives preferential treatment to black women and harsh treatment to black men in such a way that the two biases `cancel out' as in the previous example. In this case, there is a strong inclination to charge the algorithm with racial bias, which seems to undermine the alleged counterexample to the necessity of calibration. I think this kind of case can be reasonably explained by the fact that racial bias has, historically, been the norm rather than the exception in western society, and that we therefore have strong reasons to treat these subpopulation biases as prima facie evidence for a more general race bias. But in cases like the age/credit score case described above, the situation is different and it does not seem that the subpopulation biases (which point in opposite directions) generally count as evidence for more coarse grained biases in the same way. That's why we intuitively think that there's no age bias at play in the age/credit score example, whilst also being strongly disposed to suspect race bias in the case I just described.

Clearly, we are more interested in evaluating statistical markers of 'significant' group distinctions (e.g. race, gender, age etc) that track group distinctions with important social, political, economic and historical origins and ramifications.[6] Indeed, it seems unrealistic to expect our algorithms to be even roughly calibrated with respect to every possible group distinction, which suggests that the most we can reasonably demand is that they be calibrated with respect to all 'significant' group distinctions. But then we run straight back into the counterexample outlined above. One could certainly make a case for the claim that the group distinction young/old is a significant one, while the distinction young & good credit/young & bad credit/old & good credit/old & bad credit is not (if one isn't convinced by this case, replace credit score with something more trivial). This then suggests that the algorithm is actually fair after all, since it seems to be fair with respect to age, which is the only significant group distinction in play. But since the algorithm is not calibrated across the age distinction, the criterion will give the incorrect verdict that the algorithm is unfair overall. So the defender of the second interpretation of calibration within groups has two choices. They can either (i) argue that all group distinctions are equally relevant to an algorithm's fairness, in which case they avoid the counterexample (because the more fine grained distinction is treated as relevant to the algorithm's fairness, which implies that the algorithm is unfair), at the cost of placing unrealistic and unreasonable demands on predictive algorithms, or (ii) argue that only 'significant' group distinctions really matter when it comes to an algorithm's fairness, in which case the counterexample still stands (because the more fine grained partition is not treated as relevant to the algorithm's fairness, which means that the algorithm is fair even though calibration is violated).

Overall then, I do not think that calibration within groups can be helpfully employed as a necessary condition for the overall fairness of predictive algorithms. And even if one is not convinced on that point, the observation (established by the

[6] It goes without saying that the question of what counts as a significant group distinction is a deep and difficult one that goes well beyond the scope of the present work, but see e.g. Lippert-Rasmussen (2013), Thomsen (2017) for discussions of `socially salient traits'.

insurance pricing example) that calibration is not a plausible necessary condition for identifying when an algorithm treats one group unfairly in comparison to another still stands. As I stressed above, statistical criteria of algorithmic fairness should be capable of accurately identifying not only when an algorithm is generally unfair, but also when it treats one specific group unfairly in comparison to another. I've shown that being calibrated with respect to the relevant groups is not a necessary condition for treating those groups fairly.

3.2 The Insufficiency of Calibration

The discussion so far has motivated the view that none of the 11 most influential statistical criteria from the literature are plausible necessary conditions for algorithmic fairness. Before moving on the propose a novel criterion, it is worth pausing to consider why calibration within groups is not a sufficient condition for algorithmic fairness. While many authors (incorrectly, in my view) still consider calibration to be a plausible necessary condition, it is widely acknowledged that it falls short of sufficiency. Hedden, for example, writes, 'there may be other necessary conditions for fairness that concern the algorithm's inner workings and so do not count as statistical criteria...For instance, fairness may require that the algorithm be blinded to protected class membership, and to any proxies for protected class.' (Hedden, 2021: p17). While I agree with Hedden that there are certainly some necessary conditions for algorithmic fairness that cannot be properly articulated purely in terms of the statistical properties of an algorithm's predictions (more on this later), I also think it's clear that (i) it's important to identify the strongest possible statistical criteria, since we often lack any access to the design or inner working of the relevant algorithms and therefore often need to rely on statistical criteria, and (ii) identifying useful and plausible non-statistical criteria is not obviously any easier than identifying statistical ones. For instance, the requirement that an algorithm should be blinded to group membership and should not rely on any proxies for protected groups is far from obvious. In fact, it has been argued (see e.g. Corbett-Davies and Goel (2018)) that there are some cases in which fairness *requires* that the

algorithm explicitly take group membership into account (see below). Furthermore, the task of defining what counts as a 'proxy' for group membership is a non-trivial one that, it seems to me, can likely only be answered in terms of statistical relationships between the group and the prospective proxy variable. Overall then, I think it's important that we don't give up on the goal of identifying further statistical criteria of algorithmic fairness, even if we acknowledge that they will probably never be able to tell the full story. After all, in most real world cases, the statistical properties of the algorithm's predictions are all we have access to. In the next section, I turn to introducing a novel statistical criterion of algorithmic fairness that is able to diagnose many instances of unfairness that would be missed by the calibration within groups criterion, and does so without falling foul of Hedden's counterexample or the counterexample to calibration given §3.1. But before doing so, it will be helpful to consider an example that clearly illustrates why calibration within groups is not a sufficient condition for algorithmic fairness.

Imagine a bank that wants to discriminate against black loan applicants, and suppose that black applicants tend to live in zip codes with higher than average default rates, although, within any given zip code, black applicants actually have the same average default rate as other applicants from the same area. The bank can achieve its discriminatory agenda by assigning risk scores to applicants based purely on their zip code, and ignoring other relevant factors like income, credit history etc. This is an idealised illustration of a real historical phenomena called 'redlining', which lenders used to avoid giving mortgages to minority applicants in the 1930s (see e.g. Hillier (2003)). This kind of case is widely cited as a paradigmatic example of an unfair algorithm, and I think it's clear that any adequate account of algorithmic fairness should yield the verdict that this kind of practice is indeed unfair. But it's easy to see that the calibration within groups criterion is unable to properly diagnose the unfairness inherent in cases of this kind. Specifically, consider the toy example illustrated by the table below.

Redlining 1

Race	Zip	Credit	Number	Default Rate	Risk Score
White	TR10	Good	90	$\frac{1}{10}$	$\frac{1}{4}$
White	TR10	Bad	30	$\frac{1}{5}$	$\frac{1}{4}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{3}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{3}{4}$
Black	TR10	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR10	Bad	20	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{3}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{3}{4}$

Here, we consider two zip codes, TR10 and TR11. Blacks are a minority in TR10 but are a majority in TR11. On average, applicants in TR10 have a lower default rate than those in TR11.

The discriminatory algorithm assigns all applicants in TR10 a risk score of $\frac{1}{4}$ and applicants in TR11 a risk score of $\frac{3}{4}$. It's also true that, for both zip codes, the proportion of black and white applicants with good credit scores is the same ($\frac{3}{4}$ for TR10 and $\frac{1}{2}$ for TR11), as is the default rate ($\frac{1}{8}$ for TR10 and $\frac{3}{20}$ for TR11). Furthermore, an applicant's credit score is a perfect indicator of their true default risk, in the sense that, regardless of their race and zip code, 20% of applicants with bad credit scores go on to default, and 10% of applicants with good credit scores do so. By ignoring credit score and basing risk scores purely on applicants' zip codes, the algorithm seems to treat black applicants unfairly in comparison to white applicants. However, it's easy to see that the algorithm satisfies the weak formulation of the calibration within groups criterion. For, the proportion of white applicants assigned a risk score of $\frac{1}{4}$ who actually default is $\frac{1}{8}$, which is equal to the proportion of black applicants

assigned a risk score of $\frac{1}{4}$ who actually default, and the proportion of white applicants assigned a risk score of $\frac{3}{4}$ who actually default is $\frac{3}{20}$, which is equal to the proportion of black applicants assigned a risk score of $\frac{3}{4}$ who actually default. This means that both risk scores have the same evidential import for both groups, and hence that the algorithm satisfies the weak formulation of calibration within groups. This in turn establishes that calibration within groups is *not* a sufficient condition for algorithmic fairness, and that even if one still thinks that calibration is a necessary condition for algorithmic fairness, one would still need further criteria in order to diagnose unfairness in cases like this.

But before going on to identify those alternative criteria, it will be prudent to briefly clarify exactly what aspects of the Redlining 1 example generate the obvious unfairness. First of all, if, as in the actual historical case, the creators of the algorithm crafted it with the intention of disadvantaging black applicants, then it's obvious that the designer's actions in designing and constructing the algorithm themselves constitute a source of injustice and unfairness. Secondly, even if the designers of the algorithm did not explicitly intend to disadvantage black applicants, one could argue that the correlations between race, zip code and default rates are themselves the product of unjust social economic historical trends, and hence that it is unjust to apply an algorithm that exploits those correlations without recognising, and in some way compensating for, their unjust historical origin. It is important to recognise that these obvious sources of unfairness are both in some sense external to the algorithm itself. The unjust intentions of the designers demonstrate that the algorithm was the product of an unjust and unfair algorithmic design process. The fact that the correlations between race, zip code and default rates are products of unjust and unfair historical conditions shows that even if the predictions of the algorithm are not intrinsically unfair, it may still be unfair to actually use the algorithm to make important decisions without attempting to appropriately address those historical injustices. But I don't think that either of these observations show that the algorithm's predictions are *themselves* intrinsically unfair. While this might seem like a rather pedantic observation, it will have important implications further down the

line. More generally, I think that, in discussions of algorithmic fairness, it is crucial to keep track of distinctions between different kinds of unfairness, since the mechanisms that are best employed to combat or compensate for one kind of unfairness (e.g. the unjust historical origins of the correlations exploited by an algorithm) may not be effective in dealing with another kind of unfairness (e.g. an unfair statistical imbalance in the predictive tendencies of an algorithm).

So it's clear that the Redlining 1 example above involves extreme unfairness in terms of the historical conditions surrounding the design and application of the algorithm. But is there anything intrinsically unfair about the algorithm or its predictions in and of themselves? Perhaps the most obvious thing to say here is that the algorithm is intrinsically unfair simply in virtue of its using zip codes as a proxy for race. But as I intimated above, I do not take this to be a promising response. Firstly, there is good reason to think that fairness sometimes *requires* predictive algorithms to explicitly base their predictions on group membership traits like gender and race. For instance, as Corbett-Davies and Goel (2018) write,

> ...[I]t is often necessary for equitable risk assessment algorithms to explicitly consider protected characteristics. In the criminal justice system, for example, women are typically less likely to commit a future violent crime than men with similar criminal histories. As a result, gender-neutral risk scores can systematically overestimate a woman's recidivism risk, and can in turn encourage unnecessarily harsh judicial decisions. Recognizing this problem, some jurisdictions, like Wisconsin, have turned to gender-specific risk assessment tools to ensure that estimates are not biased against women.(Corbett-Davis and Goel, 2018: p2).

Secondly, it is difficult to define exactly what it means for a predictive feature to be used as a proxy for a group membership trait. On what grounds can one say that zip code counts as a proxy for race in the above case, while other variables that are also correlated with race do not count as proxies? This problem is further compounded when we recall that the predictive algorithms whose fairness we hope

to assess are often proprietary, meaning that we don't actually know exactly which predictive features are being employed by the algorithm. Thirdly, I think it's clear that merely citing the use of a proxy variable does not helpfully identify what's intrinsically wrong with the algorithm in the Redlining 1 example. In this respect, I think that the inner workings of the algorithm are largely irrelevant. If the algorithm used some other features rather than zip code to obtain the same predictions, it would still be just as unfair. It seems to me that there is something *intrinsically* unfair in the predictions themselves, and that we should not need to refer to the predictive features used by the algorithm in order to diagnose that unfairness. That is to say, we should be able to diagnose the intrinsic unfairness of the algorithm's predictions using statistical criteria alone. But as we've just seen, the most popular statistical criterion of algorithmic fairness from the literature, calibration within groups, is unable to identify any unfairness in this case. We need a new criterion to help us clearly diagnose the sense in which the predictions of the algorithm in the Redlining 1 example are intrinsically unfair.

Before introducing this new criterion, I will briefly pause to clarify what I mean when I say that the algorithm in Redlining 1 is `intrinsically unfair'. It's clear that the fairness of an algorithm is a function of many factors, including e.g. the intentions of its designers, the social and historical context of its design and application, the historical origins of the correlations that it exploits and the statistical profile of its predictions. In order to get a full picture of the (un)fairness of the algorithm, we generally need to have access to all of these features and many more. And since many of these factors concern properties of the social/historical situation inhabited by the algorithm, it seems clear that the algorithm's overall (un)fairness is not an intrinsic property. But it's possible to acknowledge that the (un)fairness of an algorithm is generally far from an intrinsic property whilst also recognizing that there are some intrinsic properties of algorithms such that any algorithm that has those intrinsic property is bound to be unfair to some degree, regardless of its other non-intrinsic properties. For instance, if an algorithm assigns radically different average risk scores to two groups with the same long run expected base rates, then there is

something intrinsically unfair about the way that the algorithm makes its judgements, in the sense that no algorithm with this property can be perfectly fair, regardless of the details of its social/historical context etc. This does not entail that there's nothing else to say about the nature and degree of the algorithm's unfairness (there always will be). But it does illustrate that there is a meaningful sense in which (the predictions of) algorithms can be intrinsically unfair. Specifically, when algorithms make predictions that systematically favor one group over another, we can conclude that those algorithms are unfair, at least to some degree, in a way that is independent of their social/historical context (in the sense that any algorithms with the same statistical profiles will be similarly unfair, regardless of their internal workings and social context). When I argue that the algorithm in Redlining 1 is intrinsically unfair, I mean that its intrinsically unfair in this sense, and we need a statistical criterion of algorithmic fairness that properly capture this fact.

4 Base Rate Tracking

To identify this criterion, let's look more closely at the Redlining 1 example. Note first that the overall average risk score for white applicants is $\frac{9}{20}$, while the overall average risk score for black applicants is $\frac{11}{20}$. Next, note that the overall default rate for white applicants is $\frac{27}{200}$, while the overall default rate for black applicants is $\frac{28}{200}$. So while the difference between the average risk scores of white and black applicants is $\frac{2}{20}$, the difference between the overall default rates of white and black applicants is only $\frac{1}{200}$. The difference between the average risk scores of the two groups is *twenty times* as great as the difference between their actual default rates. This, it seems to me, is a clear indication of unfairness. If an algorithm assigns one group a higher average risk score than another, that discrepancy has to be justified by a corresponding discrepancy between the base rates of those two groups, and the magnitudes of those discrepancies should be equivalent. In slogan form: an algorithm should only treat one groups as much more risky than another if it *really is* much more risky. We can formalise this idea with the following criterion,

Base Rate Tracking: The difference between the average risk scores assigned to the relevant groups should be equal to the difference between the (expected) base rates of those groups.[7]

I argue that, unlike calibration within groups, base rate tracking really is as a statistical criterion of algorithmic fairness, i.e. a necessary condition that any fair algorithm must satisfy. As I've just shown, base rate tracking (unlike calibration within groups) allows us to accurately diagnose the intrinsic unfairness of the predictions given by the algorithm in Redlining 1. Since the difference between the average risk scores assigned to white and black applicants is twenty times greater than the corresponding difference between their base rates, we can say that the algorithm treats black applicants unfairly in comparison to white applicants. If we were to rely only on calibration within groups, then we would need to refer to the designers' intentions, or the unjust historical origins of the relevant correlations, or the internal workings of the algorithm, in order to diagnose the unfairness in this case. But base rate tracking allows us to directly identify the algorithm as intrinsically unfair on the basis of its predictions alone. Given the lack of information that is generally available regarding the design process and internal architecture of predictive algorithms, this is important, since it shows that base rate tracking allows us to identify algorithmic unfairness in many cases where we would otherwise be unable to do so. Furthermore, base rate tracking is motivated by a natural philosophical intuition regarding the nature of fairness: that any difference in the way that an algorithm treats two groups needs to be justified by a corresponding difference in the relevant behaviours/properties of the two groups. It is unfair to treat white loan applicants as if they have a much lower average risk of defaulting compared to black applicants if they do not actually have a much lower default rate. It is also easy to see that base rate tracking, unlike the 10 influential criteria we

[7] It is worth noting here that base rate tracking is logically entailed by the strong formulation of calibration within groups, but is logically independent of the weak formulation. So there is a sense in which one can think of base rate tracking as trying to capture the aspect of strong calibration that goes beyond weak calibration but is still relevant to evaluations of fairness.

discussed in Section 2, is not undermined by Hedden's counterexample. Since the base rates for the two rooms in Hedden's coin flip example are equal to the average risk scores assigned to the people in those rooms, base rate tracking is trivially satisfied by the optimal predictive algorithm. Finally, note that base rate tracking, unlike calibration within groups, is not undermined by the insurance pricing example from Section 3.1, since that algorithm satisfies base rate tracking with respect to age groups. Whereas calibration within groups mistakenly identifies age bias where there is none, base rate tracking does not identify any unfairness in the way that the algorithm treats the two age groups, which seems intuitively correct. Overall, base rate tracking (i) is motivated by a simple and powerful philosophical intuition about the nature of fairness, (ii) is not undermined by Hedden's coin flipping example or the insurance pricing example, and (iii) significantly expands the diagnostic scope of calibration within groups in some important cases.

At this stage, a few clarifications are in order. Firstly, it should be noted that one might plausibly reformulate base rate tracking in terms of the ratios of averages risk scores and base rates, rather than differences. The resultant formulation is clearly and importantly distinct from the formulation I gave above, although it has the same motivation and is equally able to diagnose the intrinsic unfairness of the predictions in Redlining 1. For now, I am happy to defer the thorough comparative evaluation of these two formulations to later work, since the philosophically crucial insight is that the disparity in average scores should mirror the disparity in base rates, and both formulations try to capture that insight, although they do so using different mathematical functions.

Secondly, note that one could naturally try to construct an analogue of base rate tracking for binary classification algorithms. Specifically, the following principle is also motivated by the idea that any difference in the way that an algorithm treats two groups should be justified by a corresponding difference in their actual behaviours/properties.

Binary Base Rate Tracking: The difference between the percentage of members of each relevant group that are classed as 'positive' should be equal to the (expected) difference between the base rates of those groups.

To illustrate: binary base rate tracking says that it is unfair for a binary classification algorithm to class 50% of loan applicants from Group 1 as 'high risk' while classing only 30% of applicants from Group 2 as 'high risk' if it's not the case that the (expected) percentage of Group 1 applicants who actually default is not exactly 20% greater than the percentage of Group 2 applicants who actually default. While binary base rate tracking seems to be motivated by the same compelling motivation as standard base rate tracking, it's easy to see that it's actually prone to powerful counterexamples to which the original formulation is immune. To see this, imagine that twenty people are split evenly between two rooms, A and B. The A-people are all assigned coins with bias 0.6 and the B-people are assigned coins with bias 0.4. A binary classification algorithm predicts whether people's coins will land heads when tossed on the basis of their coin's bias. If the bias is 0.6, it predicts that the coin will land heads, and if the bias is 0.4, it predicts that it will land tails. Then the algorithm will predict that all A-people will toss heads, and that no B-people will toss heads, which seems perfectly fair. But the difference in the base rates of the two groups is only 20%, which is five times less than the 100% difference between the percentages of each population that are predicted to toss heads by the algorithm. This example illustrates that there is no obvious and plausible analogue of base rate tracking for binary classification algorithms. As it stands, base rate tracking can only be legitimately applied as a necessary condition for the fairness of risk scoring algorithms.[8]

Thirdly, note that base rate tracking is intended to act as a necessary condition for an algorithm to count as perfectly fair. In practice, few real algorithms will fully

[8] One could propose a weaker version of binary base rate tracking that requires only that the sign of the difference in the base rates be the same as the sign of the corresponding difference between the percentages of the groups that are classed as positive. This formulation avoids straightforward counterexamples, but it also robs the criterion of its bite and will be too weak to diagnose many paradigmatic instances of algorithmic unfairness.

satisfy this criterion.[9]However, we can still use the criterion to assess the scale and significance of an algorithm's unfairness by evaluating how far away it is from satisfying base rate tracking. If the difference between the average risk scores is far greater than the difference between the base rates, then the algorithm is very unfair, but if the divergence between those quantities is small, then the unfairness may be slight. As with any evaluative standard, perfection is a rare exception at best, and the fact that the standard is rarely fully satisfied does not undermine its claim to normative significance. Of course, one might think that the notion of `perfect fairness' is a red herring here, and claim that all we ever have are pragmatically determined standards of what counts as `fair enough'. When we're dealing with judgements that have life or death outcomes, the standard is much higher than when we're dealing with judgements that, at worst, lead to minor inconveniences for those affected. If one prefers to eschew the general ideal of perfect fairness and focus rather on context dependent notions of sufficient fairness, then one can interpret my arguments as supporting the idea that in order for an algorithm to be `fair enough' in a given context, the divergence between the base rates and the average risk scores should not be `too great', where what counts as `too great' (like what counts as `sufficiently fair') is determined by a range of pragmatic contextual variables. However, I prefer to think of statistical criteria of algorithmic fairness as imposing necessary conditions for an algorithm to count as perfectly fair, where the extent of an algorithm's unfairness tracks the extent of its violation of those criteria, and I will stick to this conception in what follows.

Fourth, note that as well as requiring that the average risk scores be equal when the base rates are, base rate tracking also requires the converse, i.e. that when the risk scores are equal, the base rates should be too. So as well as stipulating that a fair algorithm only treats groups differently when there is a suitable difference in their base rates, base rate tracking also requires that groups should only be treated similarly to the extent that their base rates are similar. Again, this is motivated by a natural intuition: that it would be unfair to treat two groups as equally risky if one

[9] The same can be said of all the statistical criteria discussed in Section 2

was in fact more risky than another. Recalling the correlation between gender and recidivism, an algorithm would seem to be unfair if it assigned males and females similar risk scores even though females had a significantly lower actual rate of recidivism. But this observation gives rise to a possible objection to the base rate tracking criterion. Going back to Redlining 1, base rate tracking successfully identifies the fact that the algorithm is unfair to black applicants, because the difference between the average risk scores of white and black applicants is far greater than the difference between their base rates. However, base rate tracking still *requires* that white applicants should be assigned a lower average risk score than black applicants, since black applicants have a higher overall default rate. And one might plausibly object that this is obviously unfair, since black applicants have the same default rate as white applicants within any given zip code. This in turn implies that base rate tracking is not a plausible statistical criterion of algorithmic fairness.

In order to respond to this concern, let's alter the Redlining 1 algorithm so that it accords with base rate tracking, as below.

Redlining 2

Race	**Zip**	**Credit**	**Number**	**Default Rate**	**Risk Score**
White	TR10	Good	90	$\frac{1}{20}$	$\frac{9}{40}$
White	TR10	Bad	30	$\frac{1}{10}$	$\frac{9}{40}$
White	TR11	Good	40	$\frac{1}{10}$	$\frac{1}{4}$
White	TR11	Bad	40	$\frac{1}{5}$	$\frac{1}{4}$
Black	TR10	Good	60	$\frac{1}{20}$	$\frac{9}{40}$
Black	TR10	Bad	20	$\frac{1}{10}$	$\frac{9}{40}$
Black	TR11	Good	60	$\frac{1}{10}$	$\frac{1}{4}$
Black	TR11	Bad	60	$\frac{1}{5}$	$\frac{1}{4}$

In this case, the algorithm still assigns risk scores based purely on zip code, but instead of uniformly assigning risk scores of $\frac{1}{4}$ and $\frac{3}{4}$ to all applicants from TR10 and TR11, respectively, it rather assigns scores of $\frac{9}{40}$ and $\frac{1}{4}$, which ensures that the difference between the average risk scores assigned to white and black applicants is equal to the difference in the default rates for white and black applicants, as required by base rate tracking. But again, one might think that this algorithm is still unjust, since it assigns white applicants a lower average risk score (0.235) than black applicants (who receive an average risk score of 0.24), and does so purely on the basis of their zip code.

In response to this criticism, it is important to recognise first that, like Redlining 1, Redlining 2 suggests some obvious sources of unfairness concerning the historical origins of the algorithm and the correlations it exploits to make its predictions. If the algorithm was designed to disadvantage black applicants, or if the correlations upon which it relies are the product of unjust historical conditions, then those constitute independent sources of unfairness which need to be appropriately recognised and taken into account in the application of the algorithm. Of course, statistical criteria like base rate tracking are unable to directly diagnose these kinds of unfairness, since they concern the historical origins of the algorithm and the relevant correlations, rather than predictive properties of the algorithm itself. However, as I stressed above, I think that one can recognise these sources of injustice without thinking that the algorithm and its predictions are themselves intrinsically unfair. In contrast, when the algorithm in Redlining 1 assigned black applicants a risk score that was higher than their white counterparts in a manner that could not be justified by a comparable disparity in their base rates, that was a case in which the algorithm's predictions were themselves intrinsically unfair, and could be identified as such on the basis of purely statistical criteria. To illustrate the point further, consider the following example. We want to predict how likely premier league forwards are to score at least 10 goals over the course of a season. Towards this end, our algorithm simply looks at whether or not the player takes more than 5 shots a game on average and assigns a risk score of $\frac{9}{40}$ to those players that don't and a risk score of $\frac{1}{4}$ to those

that do, i.e. it views players that take more than 5 shots a game as having $\frac{1}{40}$ more chance of scoring 10 goals in the season than those that don't.

Goal Predictor

Height	**> 5 Shots Per Game**	**> 50% Accuracy**	**Number**	**Base Rate**	**Risk Score**
Tall	No	No	90	$\frac{1}{10}$	$\frac{9}{40}$
Tall	No	Yes	30	$\frac{1}{5}$	$\frac{9}{40}$
Tall	Yes	No	40	$\frac{1}{10}$	$\frac{1}{4}$
Tall	Yes	Yes	40	$\frac{1}{5}$	$\frac{1}{4}$
Short	No	No	60	$\frac{1}{10}$	$\frac{9}{40}$
Short	No	Yes	20	$\frac{1}{5}$	$\frac{9}{40}$
Short	Yes	No	60	$\frac{1}{10}$	$\frac{1}{4}$
Short	Yes	Yes	60	$\frac{1}{5}$	$\frac{1}{4}$

As it turns out, short players tend to take more shots than tall players, which means that short players are assigned a higher average risk score than tall players. It's also the case that whether a player in fact scores more than 10 goals in a season is perfectly predicted by whether their shooting accuracy is greater than 50%. Regardless of the player's height and whether they shoot more than 5 times a game, 20% of players with a shot accuracy greater than 50% do score at least 10 over the season, while only 10% of those players with less than 50% accuracy do so. The proportion of tall players who are accurate is the same as the proportion of short players who are accurate amongst both the subset that take a lot of shots, and the subset that don't, although short players are slightly more accurate overall. Similarly, the proportion of tall players that take a lot of shots who actually score at least 10 is the same as the proportion of small players that take a lot of shots who score at least 10, and the proportion of tall players that don't take a lot of shots who actually score

at least 10 is the same as the proportion of small players that don't take a lot of shots who score at least 10. But it doesn't seem to me that the algorithm is treating tall players unfairly in comparison to short players by assigning them a lower average risk score. Certainly, the algorithm is far from predictively optimal, and it would be much better if it based predictions on accuracy of shots rather than volume. Still, it seems wrong to say that the algorithm treats tall players *unfairly* in comparison to shorter players. There is a difference of $\frac{1}{200}$ between the average risk scores assigned to tall and short players, but that difference is justified by a corresponding difference of the same magnitude between their base rates. But of course, the algorithm in this example is structurally isomorphic to the algorithm used in Redlining 2, both in terms of the statistical properties of its predictions, and in terms of how it uses what looks like a statistical proxy for group membership in order to determine risk scores, while ignoring more accurate predictive features like shot accuracy. I think this shows that the unfairness apparent in Redlining 2 (unlike the unfairness in Redlining 1) is not intrinsic to the algorithm, but stems rather from facts regarding the unjust historical conditions that gave rise to the correlations exploited by the algorithm, together with facts about the unjust intentions of the algorithm's designers.

I take these observations to provide an adequate response to the earlier objection that base rate tracking would require unfair differences in average risk scores in cases like Redlining 2. More generally, it is worth reiterating that base rate tracking allows us to directly diagnose unfairness on the basis of predictions alone in many cases in which calibration within groups is blind to the relevant unfairness. So even if one hopes to resist my arguments for the non-necessity of calibration within groups, there is still good reason to consider base rate tracking as an additional criterion that extends our diagnostic tool kit for identifying algorithmic injustice. Furthermore, base rate tracking is justified by the intuitive idea that groups should only be treated differently to the extent that their behaviours/properties actually reflect that difference, and, unlike most other influential statistical criteria for algorithmic fairness, it is not undermined by Hedden's counterexample.

5 The Role of Statistical Criteria

I conclude with some general remarks concerning the proper role of statistical criteria of algorithmic fairness in ensuring the equitable and fair use of predictive algorithms in society. The greatest ostensible benefit of statistical criteria is that they provide us with concrete diagnostic tools for identifying algorithmic unfairness in cases where the internal mechanisms, design processes and historical origins of the relevant algorithm are opaque or controversial. Unfortunately, Hedden's counterexample undermined 10 of the 11 most promising extant statistical criteria, and I have offered a novel counterexample that, by my lights, undermines the 11th remaining criterion. However, I have also posited another novel criterion, base rate tracking, that avoids both of these counterexamples and codifies a natural pre-theoretic intuition about the nature of algorithmic fairness. I've also shown that, even if one hopes to hold on to the calibration within groups criterion, adding base rate tracking to our toolkit allows us to diagnose unfairness that would otherwise go undetected in many cases (e.g. Redlining 1). However, it is crucial to recognise that there are some instances of the applications of predictive algorithms (e.g. Redlining 2) that involve grave injustices that simply *cannot* be properly diagnosed by purely statistical criteria. In cases like these, one can reasonably contend that the injustice is not an intrinsic property of the algorithm itself, but rather arises from historical conditions pertaining to the development and application of the algorithm. This is made most clear by cases in which two algorithms with isomorphic predictions and internal structures differ in terms of their fairness (e.g. Redlining 2 and Goal Predictor).

So while statistical criteria like base rate tracking can play an important role in the fight against algorithmic unfairness, the hardest problem will be to develop mechanisms that properly identify and compensate for the way in which algorithms exploit correlations which themselves arise from unfair historical conditions. It is important that we recognise this problem as distinct from the problem of diagnosing unfairness that is intrinsic to the way that a given algorithm makes predictions, since

the tools we use to address the latter problem (statistical criteria of algorithmic fairness) are not well suited to addressing the former.

Bibliography

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J. and Wallach, H. (2018). A Reductions Approach to Fair Classification. *Proceedings of the 35th International Conference on Machine Learning*.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias. *ProPublica*, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Chouldechova, A. (forthcoming). Fair Predictions with Disparate Impact: a Study of Bias in Recidivism Prediction Instruments. *Big Data* 5(2):153–163.

Corbett-Davies, S., and Sharad, G. (2018). The Measure and Mismeasure of Fairness: a Critical Review of Fair Machine Learning, arXiv:1808.00023.

Hedden, B. (2021). On Statistical Criteria of Algorithmic Fairness. *Philosophy and Public Affairs*, 49(2): 209–231.

Hillier, A. (2003). Redlining and the Home Owners' Loan Corporation. *Journal of Urban History.* 29 (4): 394-420.

Lippert-Rasmussen, K. (2013). *Born Free and Equal?* Oxford University Press.

Kleinberg, J, Mullainathan, S. and Raghavan, M. (2016). Inherent Tradeoffs in the Fair Determination of Risk Scores. arXiv:1609.05807

Long, R. (forthcoming). Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness. *Manuscript*.

Miconi, T. (2017). The Impossibility of 'Fairness': a Generalized Impossibility Result for Decisions. arXiv:1707.01195

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. and Weinberger, K. (2017). On Fairness and Calibration. *Proceedings of the 31st Conference on Neural Information Processing Systems.*

Thomsen, F. (2017). Direct Discrimination. In Lippert-Rasmussen (ed.) *The Routledge Handbook of Discrimination*, Routledge: 19—30.