On Statistical Criteria of Algorithmic Fairness

Abstract

Predictive algorithms are playing an increasingly prominent role in society, being used to predict recidivism, loan repayment, job performance, and so on. With this increasing influence has come an increasing concern with the ways in which they might be unfair or biased against individuals in virtue of their race, gender, or, more generally, their group membership. Many purported criteria of algorithmic fairness concern statistical relationships between the algorithm's predictions and the actual outcomes, for instance requiring that the rate of false positives be equal across the relevant groups. We might seek to ensure that algorithms satisfy all of these purported fairness criteria. But a series of impossibility results shows that this is impossible, unless base rates are equal across the relevant groups. What are we to make of these pessimistic results? I argue that none of the purported criteria, except for a calibration criterion, are necessary conditions for fairness, on the grounds that they can all be simultaneously violated by a manifestly fair and uniquely optimal predictive algorithm, even when base rates are equal. I conclude with some general reflections on algorithmic fairness.

1 Introduction

Predictive algorithms are playing an increasing role in our lives, and with this increasing influence has come an increasing concern with the ways in which they might be unfair or biased. The most famous case study of algorithmic fairness concerns the COMPAS algorithm used to predict recidivism. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a proprietary tool which assigns risk scores to individuals on the basis of a questionnaire asking about prior arrests, criminal behavior among friends and family, employment, housing, substance use, and personality traits. Importantly, it does not ask about race. COMPAS is used in jurisdictions across the United States to make decisions about pretrial release, sentencing, and parole.

In a major ProPublica report, Angwin et al. (2016) argued that COMPAS is 'biased against blacks,' since it yielded a higher rate of false positives (non-recidivists incorrectly labeled high-risk) for blacks than for whites, and a higher rate of false negatives (recidivists incorrectly labeled low-risk) for whites than for blacks. Northpointe, the company behind COMPAS, responded that the algorithm is not biased, in part because its predictions were equally accurate for the two groups (Dietrich et al. 2016). And Flores et al. (2016) rebutted ProPublica's charge, noting that for each possible COMPAS risk score, the percentage of blacks assigned that risk score who recidivated was (approximately) the same as the percentage of whites assigned that risk score who recidivated, meaning that the algorithm was in that sense equally well-calibrated. What are we to make of this? It seems that ProPublica focused on one set of fairness criteria (equal false positive and false negative rates), while Northpointe and Flores et al. focused on different ones (equal predictive accuracy and calibration).

A recent explosion of research has investigated different purported criteria of fairness and how to ensure that they are satisfied by predictive algorithms (see e.g., Hardt et al. 2016, Agarwal et al. 2018). And a number of impossibility theorems have shown that some of the most intuitively compelling statistical criteria of fairness are not jointly satisfiable except in marginal cases (Kleinberg et al. 2016, Chouldechova 2017, Miconi 2017). Importantly, the purported criteria of fairness and the impossibility theorems apply to all predictions, whether generated by machine learning, questionnaires, human judgment, or any other mechanism.

What lesson should we draw from these impossibility theorems? Many scholars interpret them pessimistically, as showing that fair prediction is impossible and that moral dilemmas are inevitable. Kleinberg et al. (2016) write that 'any assignment of risk scores can in principle be subject to natural criticisms on the grounds of bias.' Johnson (forthcoming) concludes, 'There is no such thing as an unbiased program.' Mayson (2019) interprets them as showing that 'Race neutrality is not attainable,' and Berk et al. (2018) similarly conclude, 'The implications of the impossibility results are huge...The goal of complete race or gender neutrality is unachievable.' At best, we can make optimal trade-offs between different kinds of unfairness.

More optimistically, perhaps we need to look again and determine which of the purported criteria of fairness are genuine and which are specious (see e.g., Long ms). I argue that among the statistical criteria of fairness discussed in the literature, none except perhaps calibration is a genuine necessary condition on fairness for predictive algorithms, on the grounds that all other such criteria can be violated by a manifestly fair and uniquely optimal predictive algorithm. Moreover, and surprisingly, they can be violated by this fair algorithm *even when base rates are equal* across the relevant groups. My conclusion is compatible with any or all of these (non-calibration) conditions being such that their violation provides some *prima facie* evidence that the algorithm is unfair or biased. But their violation neither constitutes nor entails unfairness in the algorithm.

2 Criteria of Fairness and Impossibility Theorems

The predictive algorithms I consider aim to predict whether an individual falls into one of two classes, positive or negative, on the basis of a set of known features. Some predictive algorithms make only a binary prediction of positive or negative. Some also, or instead, give a more fine-grained risk score. For simplicity, I assume these risk scores fall in the interval [0, 1] and can be interpreted as probabilities that the individual falls into the positive class.

I will consider algorithms that give both a risk score and a binary prediction.

Let me also emphasise that I am here concerned primarily with *predictive* algorithms, and only secondarily with the *decisions* that might be made on the basis of their predictions. A predictive algorithm might be perfectly fair, even though its predictions are put to subtly unfair or even blatantly nefarious uses. Moreover, a single predictive algorithm might be put to multiple uses, some benign and some not, or it might not feed into any decisions at all, being used instead just to satisfy one's curiosity. This distinction between predictions and decisions is important, and I return to it briefly below.

Our focus is on fairness, but we need to be more specific. I want to focus not on whether an algorithm is unfair to individuals, or whether it is unfair to groups. Rather, I want to focus on whether it is unfair to individuals *in virtue of their membership in a certain group.* We might spell out the notion of being unfair to someone *in virtue of* some trait of theirs in different ways. For instance, we might cash it out in terms of motivating reasons, so that one is unfair to someone in virtue of their having some trait only if their having that trait is one's motivating reason for treating them in that way. It is not clear, however, that this way of cashing it out could extend to the case of algorithms, for it is doubtful whether it makes sense to talk about an algorithm's motivating reasons for action. We might instead cash it out in causal or explanatory terms, so that one is unfair to someone in virtue of their having some trait only if their having that trait is a cause of, or explains, one's treating them in that way.[1] Such a causal or explanatory gloss is probably more appropriate in the present context. Even if it does not make sense to talk about algorithms acting for reasons, it does make sense to talk about an algorithm's actions being caused or explained by various factors. This is no doubt more to be said here, but I hope that this makes the notion of unfairness to

[1] This gloss is actually rather broad, for it allows that one might count as treating someone unfairly in virtue of their group membership even if one is unaware of their group membership, for instance if their group membership causes or explains some other trait of theirs, which in turn causes or explains why one treated them differently. In response, we might narrow our gloss by adding the restriction that one's treatment of the other must be caused or explained by one's *knowledge* that they have the trait in question in order to count as unfair treatment in virtue of their having that trait. For our purposes, however, we can stick with the original gloss, for the algorithm I focus on below is not unfair to individuals in virtue of their group membership, even under this broader understanding of the notion.

individuals in virtue of their group membership clear enough for present purposes.[2]

How does this notion of fairness differ from the others? One can be unfair to an individual without being unfair to them in virtue of their group membership, for instance if one treats them worse for no reason at all or for some reason unrelated to their group membership, such as their poor eye contact during an interview (assuming that this is in fact unrelated to their membership in various demographic groups). As for unfairness to groups, it is not obvious that fairness is owed to groups, as opposed to individuals. And even granting the notion of unfairness to groups, one can perhaps be unfair to an individual in virtue of their membership in a certain group without being unfair to that group itself, for instance if one treats a single individual worse because of their race or gender but at the same time takes other actions that are to the net benefit of that group. Going forward, whenever I talk about unfairness or bias, I will mean it in this sense, unless otherwise noted.

How can we determine whether a predictive algorithm is unfair? We might check whether its risk scores are based in part on group membership; that is, whether membership in one or the other group is part of the feature vector upon which predictions are based. If it is, that might constitute unfairness or bias (though see Corbett-Davies and Goel (2018) and Hellman (2020) for arguments that fairness might actually be improved by not blinding algorithms to individuals' group membership). We might also check whether it uses different thresholds for different groups in going from a continuous risk score to a binary prediction. If it does, that might also constitute unfairness or bias. But beyond these straightforward (if not uncontentious) conditions, what other conditions might be necessary for an algorithm to qualify as fair or unbiased?

I will consider a class of purported fairness criteria that require that certain relations between predictions and actuality be the same for each of the groups in question. I call these 'statistical criteria of fairness.'[3] They are attractive in part because we can determine whether

[2] See Eidelson (2015, ch. 1) for further discussion, though he focuses on discrimination rather than unfairness. Eidelson favors an explanatory account of discrimination, on which one discriminates against another on the basis of some trait if and only if one treats that other worse than some actual or counterfactual other, and where this differential treatment is explained by one's perception that the other has the trait in question.

[3] Hardt et al. (2016) describe these criteria as 'oblivious,' in the sense that they are oblivious to the 'func-

they are satisfied without actually looking at the inner workings of the algorithm, which may be proprietary or otherwise opaque. Instead, we just have look at the results—what the algorithm predicted and what actually happened.

There are eleven natural statistical criteria (three for risk scores, and eight for binary predictions) that have been proposed as potential necessary conditions[4] for algorithmic fairness. I will first state all the criteria and then describe their motivations.

> **Statistical Fairness Criteria for Continuous Risk Scores**
>
> (1) Calibration Within Groups: For each possible risk score, the (expected[5]) 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.[6]
>
> (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.
>
> **Statistical Fairness Criteria for Binary Predictions**
>
> (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

tional form of the score and underlying training data,' and are instead 'based only on the joint distribution, or joint statistics, of the true target Y, the predictions $\hat{Y}$, and the protected attribute A.'

[4] Many leading researchers refer to these and other criteria as possible 'definitions of fairness.' This is the case, for example, in Agarwal et al. (2018), Berk et al. (2018), Corbett-Davies et al. (2017), Dwork et al. (2012), and Kleinberg et al. (2016). Since these criteria are clearly not definitions, strictly speaking, we can charitably interpret these authors as proposing them as necessary conditions for fairness.

[5] I formulate all of these criteria in terms of probabilistic expectations rather than actual relative frequencies, since an algorithm can satisfy the expectational version but violate its actual relative frequency-based analogue simply due to the vagaries of chance. By the same token, an algorithm can violate the expectational version of a criterion while satisfying its actual relative frequency-based analogue, again due to the vagaries of chance. Probabilities needn't match actual relative frequencies (Hàjek 1996), and when they come apart, it is the former rather than the latter which matter for the fairness of an algorithm. How should we understand these probabilities? In some cases, as in the example involving coin flips below, we can understand them as objective chances. In others, where no objective chanciness is involved, I suggest that they are best understood as epistemic probabilities, and more specifically as the subjective probabilities that would be assigned by a reasonable individual who is familiar with the workings of the algorithm in question.

[6] See Kleinberg et al. (2016) for these first three fairness criteria. Chouldechova (2017) also considers a weakened variant of Calibration Within Groups, which requires that, for each possible risk score, the percentage of individuals assigned that risk score who are actually positive be the same for each relevant group, but not necessary equal to that risk score.

group.[7]

(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.[8]

(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]

(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]

(10) Statistical Parity: The (expected) percentage of individuals predicted to be positive is the same for each relevant group.[11]

(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.[12]

Having stated these criteria, let me give a brief overview of the motivations behind them. I will do so in a different order than that in which I presented them above, since some criteria for risk scores are analogues of other criteria for binary predictions, and *vice versa*. Start with criteria (1), (6), and (7), which form a natural family. Criterion (1) is motivated by the thought that fairness requires a given risk score to 'mean the same thing' for each relevant group (to use Corbett-Davies and Goel's (2018) phrase). It would be better to put it in

[7] The criteria of Equal False Positive Rates and Equal False Negative Rates are discussed by Hardt et al. (2016), who refer to their conjunction as 'equalized odds' and to Equal False Negative Rates alone as 'equal opportunity,' Corbett-Davies et al. (2017), who refer to Equal False Positive Rates alone as 'predictive equality,' Chouldechova (2017), who refers to their conjunction as 'error rate balance,' and Berk et al. (2018), who refer to their conjunction as 'conditional procedure accuracy equality.'

[8] See Chouldechova (2017), who refers to this criterion as 'predictive parity.' Berk et al. (2018) discuss the conjunction of Equal Positive Predictive Value and Equal Negative Predictive Value, referring to it as 'conditional use accuracy equality.'

[9] See Hellman 2020, though she is clear that she regards violation of this criterion as potential evidence of unfairness, rather than as a necessary condition on fairness. Berk et al. (2018) consider a related criterion, dubbed 'treatment equality,' which requires that the ratio of the actual number of false positives to false negatives (as opposed to their rates) be equal across the relevant groups. Note that for (8) and (10), we regard the conditions as being satisfied if the ratio is undefined for both groups due to division by zero.

[10] See Berk et al. (2018), who refer to this as 'overall accuracy equality.'

[11] See Dwork et al. (2012), Hardt et al. (2016) (who refer to it as 'demographic parity'), Chouldechova (2017), and Berk et al. (2018). Note that Hardt et al. and Chouldechova explicitly reject it.

[12] This is Miconi's (2017) 'Measure 3.'

evidential rather than semantic terms; the idea is that we want the assignment of a given risk score to have the the same evidential value, regardless of the group to which the individual belongs. Criteria (6) and (7) can be seen as generalizations of (1) from the case of risk scores to the case of binary predictions. Criterion (6) is motivated by the idea that fairness requires a prediction of positive to mean the same thing, or to have the same evidential value, regardless of the group to which the individual belongs; similarly, *mutatis mutandis*, for (7).

Next, consider criteria (2), (3), (4), and (5), which form another natural family. Start with the latter two. Criteria (4) and (5) incorporate the thought that fairness requires individuals from different groups who exhibit the same behavior to, on balance, be treated the same by the algorithm in terms of whether they are predicted to be positive or negative. It would be unfair, for instance, if individuals from one group who are actually negative tended to be predicted to be positive at higher rates than actually negative members of the other group. Criteria (2) and (3) can be seen as generalizations of (5) and (4), respectively, from the case of binary predictions to the case of risk scores, and they can be motivated in the same way. Indeed, Pleiss et al. (2017) call the quantities referred to in criteria (2) and (3) the 'generalized false negative rate' and the 'generalized false positive rate,' respectively.

Criterion (8) requires that the ratio of the false positive rate to the false negative rate be equal across each relevant group. It can be motivated by the idea that fairness requires assigning equal relative weights to the two main error types, false positives and false negatives, for the various groups (see e.g., Hellman 2020, 24-5). It would be unfair, for instance, if the algorithm tended to err on the side of caution for one group while tending to do the reverse for the other group. Criterion (9) incorporates the thought that it would be unfair if an algorithm were simply less accurate for one group than for another.[13]

Criterion (10) requires that the percentage of individuals predicted to be positive be the

[13] We could also consider analogues of (8) and (9) for risk scores, though they have not, to my knowledge, been discussed or defended in the literature. A risk score analogue of (8) would say that the (expected) ratio of the average risk score assigned to actually negative individuals to the average risk score assigned to actually positive individuals should be the same for each relevant group. A risk score analogue of (9) would require that some measure of the distance between the risk scores assigned to individuals and the truth (perhaps based on the Breier score or some other scoring rule) be the same for each relevant group.

same for each relevant group. If we think of a prediction of positive as corresponding to, or tending to yield, a certain beneficial outcome, like receiving a loan, or gaining admission or employment, then (10) can be motivated by the thought that fairness requires the same percentage of applicants from each group to receive that beneficial outcome. Put another way, (10) requires that the demographic makeup of those receiving that beneficial outcome match the demographic makeup of the population as a whole, or at least that subset of the population that applies for that benefit. But (10) is in fact widely rejected, because it is insensitive to differences in base rates (ratios of actual positives to actual negatives) across groups. Indeed, when base rates differ across groups, criterion (10) will be violated by an omniscient algorithm which perfectly predicts people's behavior. But a perfect algorithm would, presumably, not be unfair simply in virtue of differing base rates. Criterion (11) is better. Satisfaction of (11) entails satisfaction of (10) when base rates are equal across groups. But when base rates differ, it requires that differences in base rates yield corresponding differences in the rates at which individuals are predicted to be positive.[14]

Stepping back, each of these criteria requires that some function of the algorithm's predictions and the actual outcomes be the same for each group (at least in expectation). They differ in their motivations, and some are more attractive than others. But each has been considered as a possible necessary condition for a predictive algorithm to be fair or unbiased. It might be hoped, then, that some algorithms would satisfy all, or nearly all, of these criteria.

But that hope cannot be met except in marginal cases, as demonstrated by a series of impossibility theorems.[15] Kleinberg et al. (2016) prove that no algorithm can satisfy (1), (2), and (3) unless either (i) base rates are equal across the relevant groups, or (ii) the algorithm makes perfect predictions (assigning risk score 1 to all actual positives and risk score 0 to

[14] Again, we might also consider analogues of (10) and (11) for risk scores, though no such criteria have been discussed so far. An analogue of (10) would say that the (expected) average risk score assigned to individuals (whether positive or negative) should be the same for each relevant group. And an analogue of (11) would say that the (expection of) the average risk score assigned to individuals (whether positive or negative), divided by the number of individuals who are actually positive, should be the same for each relevant group.

[15] Because they do not actually show that their co-satisfiability is *impossible*, but rather that it can happen only in marginal—or trivial—cases, these results should perhaps be known as 'triviality results,' in a nod to Lewis, rather than 'impossibility results,' in a nod to Arrow.

all actual negatives). Chouldechova (2017) proves an analogous result for binary predictions, showing that no algorithm can satisfy (4), (5), and (6), again unless base rates are equal or the algorithm makes perfect predictions. Miconi (2017) proves that no algorithm can satisfy more than one of (4)&(5), (6)&(7), and (11), again unless base rates are equal or the algorithm makes perfect predictions. No published impossibility results target (8), (9), and (10). But it is easy to see that (9) and (10) are inconsistent when base rates are unequal, and likewise for (8) and (10), unless both groups have no false positives or no false negatives.

Together, these impossibility results are striking, showing that a number of intuitively attractive statistical criteria of fairness are not jointly satisfiable except in marginal cases. We might interpret these results as showing that fairness dilemmas are inevitable: whatever we do, we cannot help being unfair or biased. Alternatively, we might interpret them as showing that not all of these statistical criteria are necessary conditions for an algorithm to be fair or unbiased. Which criteria, then, are genuine conditions of fairness?

3 People, Coins, and Rooms

One way to determine whether some criterion is a genuine necessary condition on algorithmic fairness is to find a perfectly fair algorithm and see whether it is possible for it to violate that criterion. If so, then the criterion is not in fact a necessary condition on algorithmic fairness. However, this methodology may seem impractical, since in the messy real-world of predicting human behavior, there are perhaps no algorithms that are clearly and uncontroversially fair and unbiased. But we can do better by considering coin flips.

Suppose that there are a bunch of coins of varying biases. Each individual in the population is randomly assigned a coin. Then those individuals are randomly assigned to one of two rooms, A and B. Our aim is to predict, for each person, whether that person's coin will land heads or tails. That is, our aim is to predict, for each person, whether they are a heads person or a tails person. Luckily, each coin comes labeled with its bias, with a real number in the interval [0, 1] indicating its bias, or its objective chance of landing heads.

Here is a perfectly fair and unbiased predictive algorithm: For each person, take their coin and read its label. If it says 'x,' assign that person a risk score of x. And if $x > 0.5$, make the binary prediction that they are a heads person (positive), while if $x < 0.5$, make the binary prediction that they are a tails person (negative). (What if $x = 0.5$? We might arbitrarily predict heads in that case, or randomize our prediction. I will sidestep this issue by assuming that none of the coins are labeled '0.5.')

This algorithm is perfectly fair and unbiased, and in particular, it is not unfair to any people in virtue of their room membership. Its predictions are not sensitive to individuals' room membership. And the sole feature on which its predictions are based (the labeled bias of the coin) is clearly the relevant one to focus on and is neither a proxy for, nor caused or explained by, room membership. Indeed, it is not just that the algorithm is in no way unfair to individuals in virtue of their membership in a certain room; there is seemingly no unfairness of any kind anywhere in this situation. Moreover, this algorithm is uniquely optimal; no alternative can be expected to do as well or better at predicting whether individuals are heads people or tails people.

It might be objected that there can trivially be no unfairness if an algorithm's risk scores and predictions are divorced from action. Let me make three points in response. First, this objection concedes that my algorithm is fair, and hence that any criteria it violates are not necessary for fairness. However, this is compatible with those criteria being necessary for fairness, conditional on the algorithm being used to make decisions. But rather than saying that some criterion governing risk scores and predictions is necessary for fairness only when the latter feed into some decision-making procedure, it would seem more natural to hold that it is the decisions, not the risk scores and predictions, that are potentially unfair. Insofar as fairness requires anything of risk scores and predictions, it should require it whether or not they are used to make decisions. Second, recent discussion of moral encroachment (Basu 2019a,b, Moss 2018) and epistemic injustice (Fricker 2007) suggests that there may be distinctively epistemic forms of harm and injustice. For example, if one disregards the testimony of a witness due to their race or gender, or forms the belief that an African American man is staff rather than a guest due to base rate information,[16] there seems to be something morally objectionable about those beliefs themselves, even if they do not issue in any objectionable behavior. Third, we could add to my example that there are various payoffs given to individuals depending on the risk score and binary prediction they are given, and on how their coin actually lands. Even so, despite there being plenty of randomness, there is still no unfairness in the sense of treating individuals differently in virtue of their group membership.[17]

Here, now, is the key observation: While (1) Calibration Within Groups cannot be violated in this setup, all of the other criteria (2)-(11) can by violated by my fair algorithm, given a suitable assignment of coins to people and people to rooms. Moreover, criteria (2)-(11) can be violated *simultaneously* and *even when base rates are equal between the two rooms*. This is important, since much of the discussion of algorithmic fairness has focused on unequal base rates as the root of our problems: it is because base rates are unequal that we are bound to be unfair or biased. My discussion shows that all these purported criteria of fairness can and should be violated even in some cases where base rates are equal.

To see this, let us consider an easy case where actual relative frequencies match coin biases.[18] Suppose that room A has 12 people with coins labeled '0.75' and 8 people with coins labeled '0.125.' The former are all assigned risk score 0.75 and predicted to be heads people, and 9 of them in fact are heads people (since we're assuming that relative frequencies match biases). The latter are all assigned risk score 0.125 and predicted to be tails people, and 1 of them in fact is a heads person. Room B contains 10 people with coins labeled '0.6' and 10 people with coins labeled '0.4.' The former are all assigned risk score 0.6 and predicted

[16] This is an allusion to a case introduced to philosophy by Gendler (2011) and discussed by Basu (2019b) and Moss (2018). In 1995, the African American historian John Hope Franklin was hosting a dinner at the Cosmos Club the night before receiving the Presidential Medal of Freedom, when a woman called to him, presented him with her coat check, and asked him to retreive her coat.

[17] Moreover, for each criterion, whether a violation will be to the overall net detriment of one group or the other, or to neither, will depend on the details of these payoffs.

[18] See footnote 5. The distinction between formulating the criteria in terms of expectations versus actual relative frequencies is irrelevant when we consider a case where the expected relative frequencies match the actual relative frequencies. In such cases, an expectation-based criterion is violated if and onlyif its actual relative frequency-based analogue is violated.

to be heads people, and 6 of them are in fact heads people. The latter are all assigned risk score 0.4 and predicted to be tails people, and 4 of them are in fact heads people. Note that the base rates are equal; in each room 10 out of the 20 people are heads people.

Let's look at our statistical criteria (2)-(11). The average risk score assigned to room A people who are actually heads people (positive) is $(9 \times 0.75 + 1 \times 0.125)/10 = 0.6875$, while the average risk score assigned to room B people who are actually heads people is $(6 \times 0.6 + 4 \times 0.4)/10 = 0.52$. This violates (2) Balance for the Positive Class. The average risk score assigned to room A people who are actually tails people (negative) is $(3 \times 0.75 + 7 \times 0.125)/10 = 0.3125$, while the average risk score assigned to room B people who are actually tails people is $(4 \times 0.6 + 6 \times 0.4)/10 = 0.48$. This violates (3) Balance for the Negative Class.

The false positive rate for room A is 3/10, while the false positive rate for room B is 4/10, thus violating (4) Equal False Positive Rates. The false negative rate for room A is 1/10, while for room B it is 4/10, thus violating (5) Equal False Negative Rates. The positive predictive value for room A is 3/4, while for room B it is 3/5, thus violating (6) Equal Positive Predictive Value. The negative predictive value for room A is 7/8, while for room B it is 3/5, thus violating (7) Equal Negative Predictive Value. The ratio of the false positive rate to the false negative rate for room A is 3, while for room B it is 1, thus violating (8) Equal Ratios of False Positive Rate to False Negative Rate. The overall error rate for room A is 4/20, while for room B it is 8/20, thus violating (9) Equal Overall Error Rates. The proportion of room A people predicted to be heads people is 12/20, while for room B it is 10/20, thus violating (10) Statistical Parity. And the ratio of predicted heads people to actual heads people for room A is 12/10, while for room B it is 10/10, thus violating (11) Equal Ratios of Predicted Positives to Actual Positives.

It should be clear that these facts do not show that the predictive algorithm was unfair or biased against any individuals in virtue of their being members of one room or the other. The algorithm is perfectly fair; it is the statistical criteria that must go, with the sole exception of Calibration Within Groups.[19]

[19] I have argued that my example shows that no criterion except Calibration Within Groups is necessary

We can see that the argument does not depend on the neat and tidy nature of coin flips. Take any sort of behavior you like, say recidivism or defaulting on a loan. And take any set of features and any algorithm which makes predictions based on those features. Suppose that this algorithm is perfectly calibrated in the sense that, for any individual to which the algorithm asigns risk score x, the probability that that individual will display that behavior is x. In our original case, there was no unfairness in the way in which individuals were assigned features; coins were distributed by a random process. But here we can imagine that there is some background unfairness in the process by which individuals came to have the relevant features that feed into our algorithm.

Now suppose that we randomly assign these individuals to one of rooms A and B. As before, depending on the distribution of features among the individuals, and depending on who gets assigned to which room, our algorithm may violate any of these statistical criteria, except for Calibration Within Groups. But clearly, even though there may be unfairness somewhere in the system (e.g., in what determined which individuals would have which features), the algorithm is not unfair to individuals *in virtue of their room membership.*

This shows that the argument does not depend on the chanciness of coin flips, or on the fact that the coins were assigned to individuals by a random process. Even if the thing we are predicting is a non-chancy event, and even if the process by which individuals come to have the relevant features is non-chancy, our algorithm can violate any and all of our statistical criteria, with the exception of Calibration Within Groups, without being unfair to individuals in virtue of their membership in a certain group.

Let me emphasize the limited nature of my argument. I am not claiming that the case of people, coins, and rooms is realistic or completely analogous to cases like COMPAS. Of course it is not. In my example, room membership is not socially constructed, is not the basis of historical oppression, and does not influence what features people have or how they ‘behave’ (whether their coins land heads). By contrast, race, gender, and various other demographic

for fairness. One might worry that this is compatible with those other criteria being such that their violation is a *pro tanto* contributor to unfairness. But if you agree that the algorithm in my example is not merely fair, but *perfectly* fair, as I believe, then the case also shows that their violation is not even *pro tanto* unfair.

groups are at least partly socially constructed, are the bases of historical oppression, and influence both how people behave and what features they have which can feed into our predictions. In addition, as just noted, human behavior is not normally chancy, at least not in the same way that coin flips are. Finally, in my example, the two groups differ dramatically in their underlying risk distributions, despite having equal base rates. By contrast, in real life cases, groups with equal base rates likely have at least similar underlying risk distributions.[20]

But my argument does not depend on my example being realistic. First, simplifications and idealizations can help clarify issues by abstracting away from messy complicating factors. In real-life cases, group membership influences what features individuals have, thereby raising the thorny issue of basing predictions on 'proxies' for group membership. And even aside from the issue of proxies, it is typically controversial which features we should be basing predictions on, and how they are evidentially relevant. My example idealizes away from these complications, since it is clear what features we should be basing prediction on, namely the labeled biases of people's coins, which are not influenced by, or robustly correlated with, what room they belong to, and hence are not proxies for room membership. Second, and more importantly, I am only arguing that none of the above criteria (except Calibration Within Groups) are *necessary* for fairness. And to conclude that some criterion is not necessary for fairness, all you need is a single case where fairness is satisfied but the criterion violated. That is what I have sought to provide.

4 Marginality and Evidence

Many of the statistical criteria considered above fail to be necessary for reasons related to so-called 'infra-marginality,' first discussed by Ayres (2002) and recently introduced into discussions of algorithmic fairness by Corbett-Davies and Goel (2018). In their view, to be fair

[20] Indeed, it would be interesting to investigate under what assumptions about the nature of the underlying risk distributions can various of our statistical criteria be violated. For instance, which ones can be violated if base rates differ across groups but underlying risk has the shape of a normal distribution for each group? What if each group's underlying risk distribution is unimodal but not necessarily normal? And so on. I leave these questions for future research.

or unbiased, an algorithm must treat marginal cases the same across the relevant groups. For instance, the least suspicious African American person who would be predicted to be involved in crime should be exactly as suspicious as the least suspicious white person who would be predicted to be involved in crime. The least promising woman who would be predicted to succeed in some job or course of study should be exactly as promising as the least promising man who would be predicted to succeed. But many of the criteria considered above partly concern how things turn out with non-marginal, or infra-marginal, cases.

In real-life, it will always be controversial whether a given case ought to be considered a marginal one, since it is controversial which features we ought to be considering and what predictions we ought to make on that basis. But in my example, it is clear which cases are marginal and which are not; the former are those people with coins whose labeled biases are close to 0.5, while the latter are those with coins whose labeled biases are far from 0.5. In my example, our algorithm treated people from room A exactly like people from room B, assigning a risk score matching the labeled bias of their coins, and predicting heads just in case it was greater than 0.5. But it violated all of our criteria except Calibration Within Groups. This is because the people from room A were all relatively 'clear' or non-marginal cases, while those from room B were all relatively unclear or marginal cases. For this reason, our fair algorithm made more mistakes for people from room B than for people from room A. That is why the people from room A who were actually heads (tails) people had higher (lower) average risk scores than the people from room B who were actually heads (tails) people. And it's why room A had a lower false positive rate, lower false negative rate, higher positive predictive value, higher negative predictive value, lower overall error rate, and higher ratio of false positive rate to false negative rate than room B (this last inequality stemming partly from the asymmetry in the room A population).

So far, I have given no positive argument in favour of Calibration Within Groups. It is just the only one left standing in the case of people, coins, and rooms. But it is intuitively compelling and easily motivated. If it were violated by some algorithm, that would mean that the same risk score would have different evidential import for the two groups. Our probability that an individual is positive, given that they received a given risk score, would have to be different depending on the group to which the individual belongs. A given risk score, intended to be interpreted probabilistically, would in fact correspond to a different probability of being positive, depending on the individual's group membership. This seems to amount to treating individuals differently in virtue of their differing group membership.[21]

Even if Calibration Within Groups is necessary for fairness, it is likely not sufficient, for there may be other necessary conditions for fairness that concern the algorithm's inner workings and so do not count as statistical criteria (cf. Long (ms)). For instance, fairness may require that the algorithm be blinded to protected class membership, and to any proxies for protected class.[22] There may also be other conditions I have not considered.

It is compatible with my conclusion that some or all of the criteria that I have denied are *necessary* for fairness are such that their violation provides some *prima facie* evidence of unfairness. For example, a violation of one of (2)-(11) might provide evidence that the algorithm is basing its predictions on group membership or on proxies thereof, or that it violates some other necessary condition on fairness that I have not considered. But how strong this evidence is will depend on what these other necessary conditions are, and also on assumptions about the background population, including not only base rates but also the distribution of 'clear' and 'unclear' cases.

[21] Actually, fairness may only motivate a slightly weaker condition, namely that, for each risk score, the expectation of the proportion of people assigned that risk score who are actually positive should be the same for each group, *but not necessarily equal to that risk score.* Calibration Within Groups in effect combines this fairness condition with an accuracy condition which requires the proportion of people assigned that risk score who are actually positive to be equal to that risk score itself.

[22] Depending on what information we possess on which to base predictions, it may be impossible to satisfy both Calibration Within Groups and the requirement that the algorithm be blinded to group membership. And while this will often motivate acquiring more information, so that we can satisfy both calibration and blinding, this may not always be possible. To modify my example, suppose that people's coins do not come labeled with their bias. But we know that all room A people have coins with bias 0.1, while all room B people have coins with bias 0.3. Here, it is impossible to satisfy Calibration Within Groups except by basing predictions on room membership, assigning someone risk score 0.1 if they come from room A and 0.3 if they come from room B. What to make of this tension? Perhaps this is a fairness dilemma, in which we cannot help but be unfair in some respect. Or perhaps this means that Calibration Within Groups and the requirement of blindness to group membership are not both necessary for fairness. If so, I would be inclined to reject the latter rather than the former.

5 Implications

I have argued that none of the statistical criteria considered in the literature are necessary conditions for algorithmic fairness, except Calibration Within Groups. This is because all of the others can be violated by a manifestly fair and unbiased algorithm. And this can happen even when base rates are equal across the relevant groups.

Let me conclude by making a conceptual point and a practical one. The conceptual point is this: When a predictive algorithm is used to make decisions with distributional consequences or other effects that we deem unfair or unjust, this does not mean that the algorithm itself is unfair or biased against individuals in virtue of their group membership. The unfairness or bias could instead lie elsewhere: with the background conditions of society, with the way decisions are made on the basis of its predictions, and/or with various side effects of the use of that algorithm, such as the exacerbation of harmful stereotypes. The practical point is that, as a result, the best response may sometimes be not to modify the predictive algorithm itself, but to instead intervene elsewhere, by changing the background conditions of society (e.g., through reparations, criminal justice reforms, or changes in the tax code), by modifying how we act on the basis of the algorithm's predictions (e.g., by adopting different risk thresholds for different groups, above which we deny bail, or reject a loan application, and so on),[23] or by attempting to mitigate the other negative side effects of the algorithm's use.

Consider an analogy. Suppose we face two problems: traffic and inequality. We are deciding whether to adopt congestion pricing, which reduces traffic through extra fees for driving in the city during rush hours. One might worry that this is unfair to poorer people, who may have to drive farther to work and (since they earn less) would pay a higher percentage of their incomes on congestion fees. In response, one might be tempted to abandon congestion pricing altogether, or to shift to a more complicated scheme which exempts low income drivers. But a better solution is available: institute the original congestion pricing scheme along with, say, a reduction in the income tax for all lower earners. We have multiple goals (reducing

[23] Huq (2019) suggests using different risk thresholds for blacks and whites for determining whether to implement coercive interventions, albeit only for the case of less serious, non-violent crimes.

congestion and reducing inequality), but we also have multiple points where we can intervene. We shouldn't think that fairness demands that the congestion pricing be scrapped or that lower earners be exempted. We shouldn't ask the congestion pricing scheme to do all the work, addressing congestion and inequality at the same time.[24] Of course, if it is politically or otherwise infeasible to enact this optimal policy, where congestion and inequality are addressed simultaneously but separately, it may be second best to enact the more complex congestion pricing scheme that tries to address congestion and inequality at the same time. But we should not be misled into thinking that fairness itself requires this second-best solution.

Similarly with predictive algorithms. We have multiple aims: fair and accurate predictions, as well as just decisions and a just overall society. And we should not put excessive responsibility on the predictive algorithm itself for achieving these multiple ends. We should, of course, ensure that the predictive algorithm achieves the first aim. But insofar as we can, we should use additional interventions elsewhere in the system to achieve the others.

It may, however, be politically or otherwise infeasible to address unfairness through separate interventions. First, some policy measures that would be most effective in rectifying inequality or historical injustices may be too controversial to have a realistic chance of enactment. Reparations may be a case in point, and similarly for using different risk thresholds for different groups when making decisions. Second, those in charge of designing predictive algorithms are often not empowered to make some of these other interventions. They might be private companies which sell or license predictive algorithms to other bodies, such as police departments. In this case, they can offer guidance on how to make decisions on the basis of the algorithm's predictions, but their recommendations may not be binding. Or they might be local government bodies which are in a position to determine which predictive algorithm to use and how to use it, but which are unable to enact major policy changes which require legislation at the national level. Third, the use of a given predictive algorithm may have negative side effects that simply cannot be avoided. For example, even if a given predictive algorithm is itself fair, its use might exacerbate harmful stereotypes about certain groups, simply by

[24] Compare Schelling (1984, ch. 1).

making salient certain statistical information about those groups.[25] We might attempt to mitigate these effects, for instance by emphasizing that higher risk scores for members of certain groups do not reflect inherent deficiencies, but are instead the result of socioeconomic and other disadvantages. But such a disclaimer may not be fully effective.

Now, in cases where the use of an algorithm has deleterious consequences for some disadvantaged group, and where it is infeasible to mitigate these consequences through interventations elsewhere in the system, this does not necessarily mean that the predictive algorithm itself is unfair. Rather, it highlights the fact that the fairness of the predictive algorithm itself is not all that matters.

This observation, in turn, raises a number of difficult and important questions, which I can only touch on here. Let me briefly discuss two such questions. First, how should we think about these deleterious consequences? Do they constitute unfairness of some kind, or are they otherwise morally objectionable but not unfair? Consider the notion of disparate impact in anti-discrimination law. In contrast to disparate treatment, cases of disparate impact do not necessarily involve a conscious or unconscious policy or practice of treating individuals differently on the basis of their membership in a certain group, but the policy or practice nonetheless imposes disproportionate burdens on the membership of that group (and also fails some particular standard of justification). Scholars disagree about whether disparate impact should be counted as a form of discrimination. It is regarded as discrimination in many legal systems and, indeed, is often referred to as a form of 'indirect discrimination.' But some scholars argue that it is not a form of discrimination, even if it is objectionable and ought sometimes to be prohibited. Eidelson (2015, 39), for instance, argues that prohibitions on practices that result in disparate impact 'are best seen as redistributive programs to expand

[25] Other algorithms may exacerbate stereotypes more directly. For instance, some users reported that FaceApp's 'hot' filter made their skin appear lighter (Plaugic 2017), which both reflected and may have exacerbated stereotypes linking lighter skin with attractiveness. Similarly for the case of Google Translate, which translated gender-neutral pronouns into masculine English pronouns in the context of stereotypically masculine words like 'strong' or 'doctor,' and into feminine English pronouns in the context of stereotypically feminine words like 'beautiful' and 'nurse' (Lee 2018). These do not qualify as predictive algorithms on my usage, since they do not output a risk score or prediction. But the point still stands that algorithms can exacerbate stereotypes in unexpected ways.

access to opportunity, akin in many respects to a moderate but compulsory practice of positive discrimination or affirmative action.'

Similarly, we might debate whether deleterious consequences for a disadvantaged group that result from the use of an algorithm qualify as unfair or instead as perhaps morally objectionable, but not on grounds of unfairness. If they do qualify as unfair, the unfairness might be unfairness to groups, rather than unfairness to individuals in virtue of their group membership (this might also be the right way to think about disparate impact). Or they might constitute unfairness to individuals in virtue of their group membership, even if they do not implicate the predictive algorithm itself as unfair, but rather the way in which it is used or the society in which it is embedded.

In any case, even if such deleterious consequences qualify as unfair, this would not rehabilitate the statistical criteria that I have rejected. For whether the violation of a given statistical criterion has deleterious consequences for the members of some group depends on the 'payoff structure' specifying the harmful or beneficial effects of false positives, true negatives, and so on, both for those whose behavior is being predicted and for third parties, and this payoff structure will vary depending on how the predictive algorithm is used to feed into decisions, as well as the background conditions of society. Therefore, none of the statistical criteria are such that their violation necessarily harms a given group, or members thereof, and so none are necessary for fairness on this basis.

Second, regardless of whether such deleterious consequences count as unfair, how should those charged with designing or utilizing algorithms respond? Is it permissible for them to simply ensure that the algorithm itself is fair and then let the chips fall where they may, or are they sometimes required to abandon or modify an intrinsically fair algorithm so as to mitigate those deleterious consequences? How does this depend on the exact nature and severity of those consequences? What obligations do different types of institutions, such as governments and corporations, have to combat social inequities both among members (citizens and employees) and in the broader society? One might think, for instance, that public entities have stricter obligations in this regard than private ones.

Addressing these questions would go beyond the scope of this paper. Still, I think it is clear that in at least some cases where an algorithm that is itself fair is used in such a way that it unavoidably has deleterious consequences, there is good reason to abandon use of the algorithm, or modify it so as to mitigate these consequences and, more generally, to combat social inequities. While it is an ethical matter what principles of fairness, justice, and value apply here, what ought to be done in such cases will also be a complex empirical matter. It will depend on how others are likely to put the algorithm to use in making decisions, along with what broader policies might be put into practice. It will also depend on the likely side effects of different possible algorithm designs, for instance whether certain setups will exacerbate harmful stereotypes and how the use of the algorithm might change the behavior of individuals, say by encouraging or discouraging members of certain groups from applying for loans or jobs. And it will depend on whether attempts to design predictive algorithms so as to combat injustice might generate a backlash. In light of these various complexities, it is highly unlikely that the practical aim in designing predictive algorithms should be to ensure the satisfaction of some simple statistical criterion.

Summing up, I have argued that no statistical criteria, except perhaps Calibration Within Groups, are necessary conditions on fairness for predictive algorithms. But how we should actually design predictive algorithms depends on more than just the fairness of the algorithm itself. In some cases, we may be able to get the results we want by just ensuring the fairness of the algorithm while making suitable interventions elsewhere. But in other cases, we ought to design the algorithm so as to achieve certain distributional and other results. How to go about this, however, will depend both on ethical considerations and on complex, multidimensional empirical factors not reducible to a simple formula.

References

Agarwal, Alekh; Beygelzimer, Alina; Dudìk, Miroslav; Langford, John; and Wallach, Hanna. 2018. 'A Reductions Approach to Fair Classification.' *Proceedings of the 35th International Conference on Machine Learning.*

Angwin, Julia; Larson, Jeff; Mattu, Surya; and Kirchner, Lauren. 2016. 'Machine Bias.' *ProPublica*: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Ayres, Ian. 2002. 'Outcome Tests of Racial Disparities in Police Practices.' *Justice Research and Policy* 4: 131-42.

Basu, Rima. 2019a. 'The Wrongs of Racist Beliefs.' *Philosophical Studies* 176 (9): 2497-2515.

Basu, Rima. 2019b. 'Radical Moral Encroachment: The Moral Stakes of Racist Beliefs.' *Philosophical Issues* 2019 (1): 9-23.

Berk, Richard; Heidari, Hoda; Jabbari, Shahin; Kearns, Michael; and Roth, Aaron. 2018. 'Fairness in Criminal Justice Risk Assessments: The State of the Art.' *Sociological Methods and Research* OnlineFirst.

Chouldechova, Alexandra. 2017. 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.' *Big Data* 5 (2): 153-63.

Corbett-Davies, Sam and Sharad Goel. 2018. 'The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.' *arXiv preprint: https://arxiv.org/pdf/1808.00023.pdf.*

Corbett-Davies, Sam; Pierson, Emma; Feller, Avi; Goel, Sharad; and Huq, Aziz. 2017. 'Algorithmic Decision Making and the Cost of Fairness.' *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797-806.

Dietrich, William; Mendoza, Christina; and Brennan, Tim. 2016. 'COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.' Technical Report, Northpointe, July 2016. https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html

Dwork, Cynthia; Hardt, Moritz; Pitassi, Toniann; Reingold, Omer; and Zemel, Richard. 2012. 'Fairness Through Awareness.' *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214-26.

Eidelson, Benjamin. 2015. *Discrimination and Disrespect.* Oxford: Oxford University Press.

Flores, Anthony; Bechtel, Kristin; and Lowenkamp, Christopher. 2016. 'False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks.' *Federal Probation* 80 (2): 38-46.

Fricker, Miranda. 2007. *Epistemic Injustice: Power and the Ethics of Knowing.* New York: Oxford University Press.

Gendler, Tamar. 2011. 'On the Epistemic Costs of Implicit Bias.' *Philosophical Studies* 156: 33-63.

Hàjek, Alan. 1996. 'Mises Redux-Redux: Fifteen Arguments Against Finite Frequentism.'

Erkenntnis 45 (2-3): 209-27.

Hardt, Moritz; Price, Eric; and Srebro, Nathan. 2016. ‘Equality of Opportunity in Supervised Learning.’ *Advances in Neural Information Processing Systems 29.*

Hellman, Deborah. 2020. ‘Measuring Algorithmic Fairness.’ *Virginia Law Review* 106 (4): 811-66.

Huq, Aziz. 2019. ‘Racial Equity in Algorithmic Criminal Justice.’ *Duke Law Journal* 68 (6): 1043-1134.

Johnson, Gabbrielle. forthcoming. ‘Algorithmic Bias: On the Implicit Biases of Social Technology.’ *Synthese.* http://philsci-archive.pitt.edu/17169/

Kleinberg, Jon; Mullainathan, Sendhil; and Raghavan, Manish. 2016. ‘Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv preprint: https://arxiv.org/pdf/1609.05807.pdf.*

Lee, Dami. 2018. ‘Google Translate Now Offers Gender-Specific Translations for Some Languages.’ https://www.theverge.com/2018/12/6/18129203/google-translate-gender-specific-translations-languages

Long, Robert. manuscript. ‘Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness.’ NYU. Available:
http://robertlong.online/wp-content/uploads/2020/05/Fairness_in_machine_learning_5May.pdf

Mayson, Sandra. 2019. ‘Bias In, Bias Out.’ *Yale Law Journal* 128 (8): 2218-2300.

Miconi, Thomas. 2017. ‘The Impossibility of “Fairness”: A Generalized Impossibility Result for Decisions.’ *arXiv preprint: https://arxiv.org/pdf/1707.01195.pdf.*

Moss, Sarah. 2018. ‘Moral Encroachment.’ *Proceedings of the Aristotelian Society* 118 (2): 177-205.

Plaugic, Lizzie. 2017. ‘FaceApp’s Creator Apologizes for the App’s Skin-Lightening ‘Hot’ Filter.’ https://www.theverge.com/2017/4/25/15419522/faceapp-hot-filter-racist-apology

Pleiss, Geoff; Raghavan, Manish; Wu, Felix; Kleinberg, Jon; and Weinberger, Killian. 2017 ‘On Fairness and Calibration.’ *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017).* Available: http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf

Schelling, Thomas. 1984. *Choice and Consequence.* Cambridge, MA: Harvard University Press.