Journal of Artificial Intelligence Research (2023) Submitted 07/23; published

Individual Fairness, Base Rate Tracking and the Lipschitz Condition

Benjamin Eva BENJAMIN.EVA@DUKE.EDU
Duke University, Department of Philosophy,
Durham, NC 27712 USA

Abstract

In recent years, there has been a proliferation of competing conceptions of what it means for a predictive algorithm to treat its subjects fairly. Most approaches focus on explicating a notion of *group fairness*, i.e. of what it means for an algorithm to treat one group unfairly in comparison to another. In contrast, Dwork et al (2012) attempt to carve out a formalised conception of *individual fairness*, i.e. of what it means for an algorithm to treat an *individual* fairly or unfairly. In this paper, I demonstrate that the conception of individual fairness advocated by Dwork et al is closely related to a criterion of group fairness, called 'base rate tracking', introduced in Eva (2022). I subsequently show that base rate tracking solves some fundamental conceptual problems associated with the Lipschitz criterion, before arguing that group level fairness criteria are at least as powerful as their individual level counterparts when it comes to diagnosing algorithmic bias.

1. Introduction

How do we determine whether the predictions made by an algorithm are fair and unbiased? According to a highly influential recent strand of literature, unfairness can often be diagnosed by means of *statistical criteria of algorithmic fairness*, i.e. necessary conditions that must be satisfied by the statistical profile of the algorithm's predictions if those predictions are to count as 'fair'. Typically, these criteria are *group-based*, meaning that they are articulated in terms of how the algorithm's predictions track distinctions between different subgroups of the relevant population. For instance, if we are interested in determining whether the algorithm treats women unfairly in comparison to men, then we can check how the algorithm's predictions for men differ from its predictions for women, and compare the results against the relevant criteria. Below are four prominent examples of group based statistical criteria of algorithmic fairness that have been widely advocated in the literature. For illustration, I suppose that the algorithm is trying to predict whether subjects instantiate a particular trait T, and that we are interested in whether the algorithm's treatment of a group G_1 is fair in comparison to its treatment of a second group G_2.

> **Statistical Parity:** The percentage of subjects from G_1 that are predicted to have T should be equal to the percentage of subjects from G_2 that are predicted to have T.

> **Equal False Positive Rates:** The percentage of subjects from G_1 that are falsely predicted to have T should be equal to the percentage of subjects from G_2 that are falsely predicted to have T.

©2023 The Authors. Published by AI Access Foundation under Creative Commons Attribution License CC BY 4.0.

> **Equal False Negative Rates:** The percentage of subjects from G_1 that are falsely predicted to lack T should be equal to the percentage of subjects from G_2 that are falsely predicted to lack T.
>
> **Equal Error Rates:** The percentage of subjects from G_1 for which the prediction is incorrect should be equal to the percentage of subjects from G_2 for which the prediction is incorrect.[1]

There has been a great deal of debate focusing on the relative merits and shortcomings of criteria like these. But one thing that almost all such debates have shared is the presupposition that whatever the true statistical criteria of algorithmic fairness end up being, they will be articulated in terms of group prediction profiles. One notable exception to this general pattern comes from Dwork et al (2012), who propose a statistical criterion of algorithmic fairness that is articulated not in terms of the differential treatment of groups, but rather in terms of the differential treatment of *individuals*. Very roughly summarised, their criterion, the 'Lipschitz condition', requires that *similar individuals should receive similar treatment*. Dwork et al argue both that the Lipschitz condition is capable of diagnosing a range of important types of fairness violation, and that the condition can be usefully operationalised in real world settings.[2]

As it turns out, the Lipschitz condition shares a close conceptual affinity with a recently proposed group based fairness criterion called 'base rate tracking' (Eva, 2022). Whereas the Lipschitz condition (roughly) requires that similar individuals should be treated similarly, base rate tracking (roughly) requires that similar groups should be treated similarly. In this paper, I argue that base rate tracking (i) avoids some basic conceptual and operational problems associated with the Lipschitz condition whilst respecting the spirit of its motivation, and (ii) performs at least as well (when combined with other group level criteria) as the Lipschitz condition in diagnosing unfairness in the kinds of cases for which the Lipschitz condition was designed. The upshot is that we can capture the main advantages of the Lipschitz criterion without moving from group based to individual based criteria for algorithmic fairness.

The structure of the article is as follows. §2 introduces and motivates the Lipschitz condition and the base rate tracking criterion and formally illustrates the strong structural analogy between them. §3 identifies an important conceptual problem for the Lipshitz criterion, evaluates (and ultimately rejects) some possible solutions, and observes that the base rate tracking criterion completely avoids the problem. §4 compares the ability of base rate tracking and the Lipschitz criterion to diagnose unfairness in the types of cases for which the Lipschitz condition was designed. Finally, §5 discusses the prospects for individual based fairness criteria in general and concludes.

[1] For discussion of these and other group level statistical fairness criteria, and their relationship to one another, see e.g. Corbett-Davies and Sharad (2018), Hedden (2021), Klein et al (2016), Miconi et al (2017), Pleiss et al (2017).

[2] Since its introduction by Dwork et al, the notion of individual level algorithmic fairness has been further investigated by e.g. (Fleisher (2021), Friedler et al (2016), Ilvento (2020), Mukherjee et al (2020), Wang et al (2019)).

§2: Base Rate Tracking and the Lipschitz Condition

I begin by briefly establishing some basic terminology and notation. Throughout the paper, I assume that there exists a population $N = \{x_1, ..., x_n\}$ of n individuals and a property y such that every individual in N either has the property y or lacks it. The goal of a predictive algorithm is to predict, for each individual in N, whether or not they have the property y. This is accomplished by assigning each individual a 'risk score' between 0 and 1 that can intuitively be conceived as the probability that the algorithm assigns to that individual possessing the property y.[3] We formalise this by representing a predictive algorithm as a function $h : N \to [0, 1]$ that takes individuals to risk scores.

§2.1: The Lipschitz Condition

What does it mean for a predictive algorithm to treat its subjects fairly? Dwork et al (2012) answer this question starting from the following basic premise.

> We capture fairness by the principle that any two individuals who are similar with respect to a particular task should be classified similarly. (Dwork et al 2012: p214).

To illustrate the idea here, imagine that a loan application algorithm aims to predict whether a subject will default on their loan payments. It would seem to be unfair if two subjects who were similar in all relevant respects (credit history, age, income etc) were assigned dissimilar predictions by the algorithm. In order to formalise this intuitive idea, Dwork et al assume that it is possible to define a *similarity metric* $d : N \times N \to [0, 1]$ over the space N of individuals. Intuitively, the lower the value of $d(x_1, x_2)$, the more similar individuals x_1 and x_2 are to one another. In order to formalise the Lipschitz condition, we also need to establish a statistical distance metric D between risk scores that quantifies the similarity between the risk scores assigned to different individuals. Intuitively, the lower the value of $D(h(x_1), h(x_2))$, the more similar the risk scores assigned to x_1 and x_2 are to eachother.[4] Following Dwork et al, we remain agnostic for now about exactly what form the distance metrics d and D must take, and treat them as primitives (we'll return to this in §3). We can now define the Lipschitz condition as follows,

The Lipschitz Condition (LC): For any $x_i, x_j \in N$, $D(h(x_i), h(x_j)) \leq d(x_i, x_j)$

[3] Note that my focus here is on quantitative risk scoring algorithms rather than qualitative classification algorithms that produce categorical predictions like 'yes' or 'no' (regarding whether the subject has the target property), as opposed to numerical probabilities. Most of the discussion can be easily translated to the qualitative setting, but I focus only on risk scoring algorithms since doing so simplifies the presentation.

[4] If we want to be formally rigorous, note that every risk score corresponds to a probability distribution over the partition whose two cells correspond to the case where the individual has and lacks property y, respectively. The statistical distance measure represents (some formalisation of) the distance between these probability distributions. One might use, for instance, an f-divergence such as the Kullback-Leibler divergence or the Hellinger distance (see e.g. Diaconis and Zabell (1982), Eva et al (2020)), or some other kind of statistical distance measure, like the squared Euclidean distance.

Informally, LC just says that the extent to which the algorithm can fairly treat individuals as dissimilar (by assigning them dissimilar risk scores) is bounded by the extent to which the individuals really are dissimilar (according to the given similarity metric). In slogan form, LC requires the algorithm to treat similar individuals similarly.

§2.2: Base Rate Tracking

Like LC, base rate tracking is motivated by the idea that a fair algorithm can only treat subjects differently when that differential treatment is justified by relevant and proportional differences in the behaviour/traits of those subjects. But whereas LC formalises this idea in terms of similarities and differences between *individuals*, base rate tracking formalises it in terms of similarities and differences between *groups*. In fact, we can formalise base rate tracking using an analogous schema to that used for LC. Let $d : \mathcal{P}(N) \times \mathcal{P}(N) \to [0, 1]$ be a function that takes subsets of N (i.e. groups) and returns a number representing the similarity of the groups (where lower values represent more similar groups). Similarly, let $D : \mathcal{P}(N) \times \mathcal{P}(N) \to [0, 1]$ be a function that takes a pair of groups and returns a number representing the distance between the distributions of risk scores across the two groups (where lower numbers represent more similar distributions). Again, we take these distance metrics as primitive for now. Then we can formalise a group level generalisation of LC as follows,

The Group-Level Lipschitz Condition (GLC): For protected groups $X_i, X_j \subseteq N$,

$$D(X_i, X_j) \leq d(X_i, X_j)$$

We can think of Eva's (2022) base rate tracking criterion as a particular instantiation of GLC. Specifically, we get half of the logical strength of base rate tracking if we conceive of the similarity and differences between groups in terms of their base rates (for instantiating the target variable y). To see this, consider Eva's original formulation of base rate tracking,

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

Trivially, BRT can be divided into two separate conditions,

BRT1: The difference between the average risk scores assigned to the relevant groups should be no greater than the difference between the (expected) base rates of the groups.

BRT2: The difference between the average risk scores assigned to the relevant groups should be no less than the difference between the (expected) base rates of the groups.

It is easy to see that BRT1 is an instance of GLC. Specifically, in the special case in which d is the function that takes a pair of groups and returns the difference between their base rates and D is the function that takes a pair of groups and returns the difference between their average risk scores, BRT1 is equivalent to GLC. This observation clearly demonstrates that BRT and LC are, in a precise sense, motivated by the same basic idea

– that any difference in treatment needs to be justified by a corresponding proportional difference in the relevant traits/behaviour.

Of course, readers will also note that BRT2 is logically independent of GLC. Since BRT is the conjunction of BRT1 and BRT2, this shows that BRT is intuitively 'stricter' than LC in the sense that it imposes a fairness constraint whose individualised analogue is not implied by LC. However, we can easily formalise the individualised analogue of BRT2, as follows,

The Inverse Lipschitz Condition (ILC): For any $x_i, x_j \in N$,

$$D(h(x_i), h(x_j)) \geq d(x_i, x_j)$$

ILC is to BRT2 as LC is to BRT1. That is to say, the group analogue of ILC is equivalent to BRT2 in the case where the group metrics d and D are interpreted as the difference between the base rates and the difference between the average risk scores of the relevant groups, respectively. Let SLC (the 'strong Lipschitz condition') be the conjunction of LC and ILC. Trivially, BRT is equivalent to the group analogue of SLC when BRT1 and BRT2 are equivalent to the group analogues LC and ILC, respectively.

Since Dwork et al (2012) only defend LC, I will focus primarily on BRT1 and LC (rather than BRT2 and ILC) in what follows. However, it worth noting that there are good reasons to think that BRT2 and SLC are just as well motivated as BRT1 and LC in most cases. While BRT1/LC (roughly) require that similar groups/individuals receive similar treatment, BRT2/ILC (roughly) require that dissimilar groups/individuals receive dissimilar treatment. Regardless of whether we're talking about groups or individuals, the latter requirement seems to be at least as compelling as the former. For instance, it would be manifestly unfair if, in a pretrial setting, women had a much lower base rate than men for recidivism, but were predicted to reoffend at a similar rate to men (see e.g. Corbett-Davies Goel (2018)). In what follows, I will focus on BRT1/LC and argue that while BRT1 should be viewed as a genuine statistical criterion of algorithmic fairness, LC should not. However, as intimated above, I think that the motivations for BRT2/ILC are just as strong as the motivations for BRT1/LC. As it turns out, all of my arguments for preferring BRT1 to LC also license a corresponding preference for BRT2 over ILC.

Before moving on, it is important to make one further clarification. Eva (2022) formulates BRT in terms of differences between base rates and average risk scores. But he also explicitly acknowledges that one might rather use some other distance functions, as long as they take the relevant base rates and average risk scores as their arguments. For instance, one might reformulate BRT in terms of ratios of base rates and ratios of average risk scores, as follows

Ratio Base Rate Tracking (RBRT): The ratio of the average risk scores assigned to the relevant groups should be equal to the ratio of the (expected) base rates of the groups.

Like Eva (2022), I remain agnostic about which formulation of BRT to prefer, since the arguments regarding the comparative advantages of BRT1 over LC are independent of the specific preferred functional form of BRT.

§3: The Problem of Similarity

§3.1: Individual Similarity

I turn now to highlighting a basic conceptual problem with LC. Specifically, LC presupposes the existence of a distance metric d over the space of individuals, which is supposed to encode facts about the comparative similarity of individuals to one another. But how should determinations of similarity be made in order to fix the values of this metric? Dwork et al state that they are interested in similarity *with respect to the particular task in question.* But that doesn't help to answer the question. By way of illustration, consider a predictive algorithm that aims to predict whether a subject will like the next James Bond movie. If we hope to apply LC to help gauge the fairness of the algorithm, we need to determine which subjects are similar to which other subjects with respect to the question of whether or not they will like the new James Bond movie. But what does that mean? Which features are relevant to determinations of similarity in this context and how do we determine the weights of their contributions?[5]

One possible answer is to say that the weight of a characteristic's contribution to the similarity metric should be a function of the extent to which two individuals having the same profile for that characteristic is predictive of their having the same profile for the target variable (in this case, liking the new James Bond movie). Under this interpretation of similarity, two individuals being relevantly similar essentially equates to them being roughly the same with respect to the traits that are predictive of the target variable. Of course, there will generally be very many traits that are in some way predictive of the target variable, and one would always need to identify some optimal subset of those to defer to when fixing the values of the similarity metric. As Dwork et al point out, this is the kind of task for which many machine learning methods have been designed,[6] so one might hope that the set of similarity determining characteristics (and their weights) can be determined via a suitable machine learning method that identifies an optimal set of predictors for the target variable. I take this to be the proposal when Dwork et al write 'The construction of a suitable metric can be partially automated using existing machine learning techniques' (Dwork et al (2012): p223).[7]

Unfortunately, there are some significant issues with this strategy. Recall that our basic problem is to construct a method for evaluating the fairness of a predictive algorithm h. The current proposal suggests that we do this by way of a similarity metric whose values are determined by a second external predictive algorithm h^* (the machine learning method that identifies the optimal set of predictors). But then how do we ensure that h^* operates in a fair manner that outputs a similarity metric that is genuinely relevant to the fairness of h? Clearly, we can't appeal to LC again, on pain of regress. The alternatives are (i) to impose some further fairness constraints on h^* (not including LC), or (ii) to argue that h^* does not

[5] Should we include protected characteristic such as race, gender and sexual orientation in determinations of similarity? If so, then we seem to allow that it can be fair for the algorithm to make different predictions for two subjects who are completely alike in every respect except for their race, which (it seems) is exactly the kind of thing that LC was designed to preclude. If we don't allow protected characteristics to enter into determinations of similarity, then what about non-protected characteristics that are strongly correlated with protected ones? And how much correlation is too much?

[6] Think, for instance, of minimising R^2, eliminating colinearity etc in regression analysis.

[7] I also consider some alternative interpretations of their proposals later in this subsection.

need to satisfy any fairness criteria whatsoever. The problem with (i) is that the advocate of LC owes us an account of what these further conditions should be. And whatever account they give, it seems somewhat convoluted and artificial to impose LC as a fairness criterion, and then to impose some distinct fairness criteria as constraints on the algorithms that we defer to in order to apply LC. Furthermore, this response begs the question 'if LC is really a fairness criterion, then why doesn't it apply to h^* as well as h, since both are predictive algorithms?'. Remember, we are conceiving of statistical criteria of algorithmic fairness as *necessary conditions for algorithmic fairness*, and necessary conditions don't admit of exceptions. Thus, (i) does not look like a promising response. (ii) holds that h^* doesn't need to satisfy any fairness criteria whatsoever. But again, this seems strange. How can we expect LC to enforce the fairness of predictive algorithms if the application of LC relies on using predictive algorithms that are not constrained by any fairness considerations whatsoever? Overall then, it's difficult to see how this approach to fixing the similarity metric can work without relying on blind faith in the fairness of the algorithm that identifies the set of similarity determining characteristics.[8]

Another problem for this approach is that there will, in general, be multiple different ways of identifying the optimal set of predictors for the target variable that are all roughly on a par with respect to accuracy, but which ultimately yield highly divergent similarity metrics. This observation is a direct corollary of the 'Rashomon effect' (see Breiman, 2001) – a common phenomenon that arises in situations where 'there are many models that satisfy predictive accuracy criteria equally well, but process information in the data in different ways' (D'amour 2001, p1). In situations like this, it becomes completely unclear how we should determine the similarity metric (even if we forget about the fairness related worries outlined above). If multiple models that are all equally accurate identify different sets of characteristics as predictive of the target variable, then which model should we defer to when fixing the similarity metric? It does not seem that there could be a principled answer to this question.

At this stage, the idea that we can fix the similarity metric purely by using a machine learning algorithm to identify (and weigh) an optimal set of predictive characteristics seems dead in the water. But Dwork et al do also tentatively suggest some additional mechanisms for fixing the similarity metric, and one might expect those to prove helpful here. One interesting suggestion (see Dwork et al (2012), p 224) is that the similarity metric may be determined, at least in part, by how the subjects *want* to be compared to the rest of the population. In determining whether the algorithm treated a particular subject S fairly, we should take into account how S would like to be compared to their peers. For instance, in the case where there are multiple models that are equally accurate but which yield different sets of predictive characteristics and similarity metrics, the subject might choose the metric that they like best from amongst the set generated by the most accurate models. While this might look like an elegant solution to the under-determination of the similarity metric, it doesn't actually help much. First of all, if (as seems likely) subjects are often indifferent between multiple different similarity metrics, then the under-determination remains.[9] Secondly, one can imagine a case where a subject, for whatever reason, chooses

[8] This point is related to Fleisher's (2021) second objection to individual fairness, discussed below.

[9] One can even imagine a case where the subject, for whatever reason, is *incapable* of choosing between the candidate similarity metrics.

to be evaluated on a metric according to which the algorithm treats them fairly, when there are many other metrics that deem their treatment to be unfair. For instance, the subject might belong to an ethnic minority and have an optimistic perspective on the racial politics of their society, which leads them to choose a similarity metric that's not sensitive to racial bias. And it could turn out that this metric deems the subject's treatment to be fair, while other metrics that are sensitive to racial bias would deem their treatment to be unfair. It seems strange to say that the algorithm is fair in this case, just because the subject chose to be evaluated by a metric that was blind to racial bias. Thirdly, even if this strategy were able to solve the under-determination problem in a satisfactory manner (which it doesn't seem to be), the problem of ensuring the fairness of the algorithms that yield the candidate similarity metrics remains unsolved.

Apart from this specific suggestion, Dwork et al also suggest that 'human insight and domain information' may also be useful in fixing the similarity metric. But it's difficult to see how any general appeal to human insight or domain specific information can be of use here. Even when we narrow our domain to a very specific prediction task, there will generally be many possible similarity metrics, all motivated by different predictive models that are indistinguishable in terms of accuracy. It's fanciful to suppose that human judgement can generally choose between these metrics in a principled, reliable and non-arbitrary way. Furthermore, Fleisher (2021) identifies some more general and fundamental conceptual limitations to this strategy.

Firstly, when Dwork et al propose using 'human insight' to fix the similarity metric, the idea seems to be that we identify some group of human arbiters who provide feedback on the extent to which different individuals should be considered relevantly similar (for a given task) and/or whether those individuals are treated fairly by some predictive algorithms, and that we use that feedback to determine the values of the similarity metric. In fact, various recent attempts to implement individual fairness criteria explicitly rely on this kind of appeal to the judgements of human arbiters (see e.g. Gillen et al (2018), Ilvento (2020), Lahoti et al (2019), Mukherjee et al (2020), Wang et al (2019)). However, as Fleisher notes, relying on the judgements of human arbiters to fix the similarity metric raises more problems than it solves.

> [T]he appeal to human arbiters to learn a similarity metric suffers from a difficulty stemming from human biases. It is well known that humans exhibit pernicious, discriminatory biases in their judgments. Moreover, these biases need not be explicit. A large body of psychological research collected over the past five decades provides significant evidence that human judgment and decision-making suffer from systematic biases that individuals are not aware of... Much of this bias concerns rational belief and decision-making quite generally, e.g., failures to respect basic principles of probability and rational choice... The evidence of these biases have raised significant difficulties for using classical decision-theory for descriptive purposes in economics. And unfortunately, these implicit and systematic biases are not limited to prudential rationality. Implicit bias is a significant factor in perpetuating oppressive structures involving race, gender, and other sensitive categories. (Fleisher 2021: 11)

The objection is a deep one. By uncritically appealing to the judgements of human arbiters to determine the relevant similarity metrics, we run the very real risk of further entrenching the (implicit or explicit) systematic biases of those arbiters by enshrining them in our conception of algorithmic fairness. Thus, the proposal to simply rely on the judgements of human arbiters does not seem to be a promising one. Of course, it's true that we need to rely on human judgement and intuition when we systematise our conception of algorithmic fairness at a general level, but uncritically integrating the similarity judgements of humans into our methods for diagnosing unfairness without a principled method of evaluating or filtering those judgements is clearly prone to exacerbate the issue.[10]

Fleisher (2021) also raises another, more general, objection against LC and its similarity based conception of individual fairness. Specifically, Fleisher notes that, however we try to fix the similarity metric, doing so will require us to make substantive moral judgements about what considerations can fairly be treated as relevant to determinations of individual similarity. For example, in the context of predicting an applicant's performance at a university, one might ask whether race can fairly be treated as a relevant factor when determining the similarity of two applicants. Any answer to this question represents a substantive moral presupposition that lies beyond the purview of the Lipschitz criterion. So whenever we apply LC, we are implicitly relying on antecedent moral judgements regarding which characteristics can fairly be incorporated into determinations of individual similarity. But LC is completely silent about how we should arrive at those moral judgements.[11] Thus, there is a gaping moral lacuna at the heart of LC that needs to be filled before the condition can be meaningfully applied to any real world problem.

In sum then, it seems that the problem of fixing the similarity metric to be used in LC is a deep one that admits of no straightforward resolution. In order for LC to be a substantive and coherent criterion of algorithmic fairness, we would need to have access to a principled general procedure for determining the appropriate similarity metric for every prediction task, and for settling the substantive moral questions pertaining to which characteristics can fairly be treated as relevant to judgements of individual similarity. We have just seen that we do not have access to any such procedure.

§3.2: Group Similarity

The fundamental problem with LC, then, is that it lacks a principled and robust way of determining the similarity of individuals. But what about BRT1? Recall that BRT1 can be seen as a particular instantiation of the group level analogue of LC, GLC. As such, it relies on a notion of similarity between groups. Is this notion any less problematic than its individual level analogue? It's not hard to see that it is. In fact, BRT1 encodes a particular conception of how to rigorously determine group level similarity relations in a principled way, namely via *base rates*. According to BRT1, two groups are similar (with respect to a given prediction task) to the extent that their base rates for the target variable are similar (where similarity between base rates is assessed using e.g. differences/ratios etc). Returning to an earlier example, in the context of predicting whether or not subjects will like the new

[10] For a recent discussion of the conditions under which human judges can reasonably be replaced by AI judges, see Afrouzi (forthcoming).

[11] Just as it was silent about how to assess the fairness of the machine learning methods that are used to determine the similarity metric on the original proposal.

James Bond film, we can consider two group G_1 and G_2 to be similar to the extent that the proportions of members of those two groups who do like the movie are similar. This conception of group level similarity is compelling – the extent to which two groups are relevantly similar with respect to exhibiting a trait T is nothing other than the extent to which the proportion of members of the first group who exhibit the trait is similar to the proportion of members from the second group who exhibit the trait.

Furthermore, it is clear that formalising group similarity in terms of base rates allows us to completely bypass the problems associated with determining an individual level similarity metric. There is no need for a second algorithm to identify an optimal set of predictors to ground the similarity metric, which means that the concerns surrounding the fairness of the second algorithm and the under-determination of the metric (arising from the Rashomon effect) simply don't apply to BRT1. Similarly, there is no need to appeal to the judgements of human arbiters, which means that BRT1 does not run the risk of further entrenching any systematic biases that may be prevalent in the judgements of those arbiters.

The crucial observation here is that groups posses a statistical structure that warrants principled task dependent similarity judgements, while individuals lack any such structure. What this all shows is that if we want to respect the ideal that similar subjects should receive similar treatment, then we need to encode that ideal not in an individual based fairness criterion, but rather in a group based criterion, like BRT1.

§4: Diagnostic Power

I've argued both that LC and BRT1 are motivated by the same basic philosophical ideal, and that LC faces a deep and intractable conceptual problem that BRT1 completely avoids. I turn now to comparing the ability of the two criteria to diagnose unfairness in the kinds of cases that Dwork et al cite as motivations for LC. It's important to stress here that both LC and BRT1 are supposed to represent necessary, not sufficient, conditions for algorithmic fairness.[12] So even if there are some cases of unfairness that the algorithms are unable to diagnose, that doesn't speak against their legitimacy as necessary conditions for fairness (for that, we'd need examples of fair algorithms that violate the conditions, rather than unfair algorithms that satisfy the conditions). Nevertheless, it is obviously desirable for our statistical criteria of algorithmic fairness to be capable of diagnosing as many different flavours of unfairness as possible, and the following discussion examines their abilities to do exactly that. In an appendix, Dwork et al (2012) list six salient varieties of algorithmic bias that they intend to target with LC. I go through each in turn.

1: 'Blatant explicit discrimination' – Dwork et al describe this as a case in which membership in a protected group S is explicitly tested for and unfavourable outcomes (for our purposes, high risk scores) are given to members of S compared to nonmembers. LC will diagnose this variety of unfairness as long as the similarity metric does not consider members of S to be dissimilar to nonmembers (as Dwork et al recognise). However, if S membership (or one of its correlates) is one of the properties that

[12] Dwork et al sometimes refer to LC as a 'definition' of fairness, which suggests that they also consider it to be a sufficient condition for the fairness of an algorithm. But in the present context, I am concerned only with the weaker claim that LC constitutes a necessary condition for fairness (see Fleisher (2021) for an argument against the sufficiency of LC).

the metric uses to determine similarity, then S members will typically be regarded as dissimilar from nonmembers, which will result in LC failing to diagnose explicit discrimination as unfair.

What about base rate tracking? Again, base rate tracking will accurately diagnose this form of algorithmic bias so long as the group S is regarded as similar to its complement, i.e. as long as the relevant base rate for S is similar to the corresponding base rate for S's complement. If the base rates for the two groups are in fact dissimilar, then base rate tracking won't always identify any bias or unfairness here (similarly to how LC fails to diagnose explicit discrimination as unfair when members of S are deemed to be relevantly dissimilar to nonmembers).

Now, one might worry that even if the relevant base rate for S is significantly different from the corresponding base rate for S's complement, it might be unfair for the algorithm to assign higher risk scores to members of S than it does to nonmembers. There is already a large literature surrounding the question of whether it is unfair for predictive algorithms to assign higher risk scores to members of groups with higher base rates when the base rates are largely a result of large scale bias and discrimination. While that question lies well beyond the scope of the current discussion, I will make two relevant observations here. Firstly (as has been widely noted), statistical criteria of algorithmic fairness are supposed to diagnose the fairness of predictive algorithms, not the fairness of the decision making processes that make use of those algorithms, or the fairness of the processes by which the data that is fed to the algorithms is generated or collected. One might argue (for example) that it can be perfectly fair for an algorithm to assign higher risk scores to members of disadvantaged groups with higher base rates, but subsequently contend that the disadvantaged groups should be given preferential treatment when it comes to making decisions on the basis of the algorithm's recommendations (so that a member of a disadvantaged group with the same risk score as a nonmember might get assigned a more favourable outcome). Secondly, it should be noted that even if one thinks that the predictive algorithm itself is unfair in cases where it assigns higher risk scores to disadvantaged groups with higher base rates, LC fares no better than BRT1 in diagnosing that alleged unfairness. For, if S has a higher base rate than its complement, then S membership (as well as many of the characteristics that are correlated with S-membership) will be predictive of the target variable. So if (as discussed in §3) we try to use relevantly predictive characteristics to fix the similarity metric for LC, we will very likely end up basing our similarity judgements on statistical proxies for S-membership, which will yield the result of regarding S-members as dissimilar to nonmembers, which in turns makes it impossible for LC to diagnose the unfairness.

2: 'Discrimination based on redundant encoding' – they describe this as a case where the explicit test for group membership that occurs in explicit discrimination is replaced by another test that is in practice equivalent. For instance, rather than explicitly considering sexual orientation, the algorithm might rather consider another feature that is strongly correlated with sexual orientation and make its recommendations based on that. Dwork et al claim that LC is also able to successfully diagnose discrimination based on redundant encoding.

Now, since LC and BRT1 are both statistical criteria of algorithmic fairness, they look only at the statistical profile of the algorithm's predictions, not at the internal architecture or design of those algorithms. Thus, from the perspective of LC and BRT1, discrimination based on redundant encoding is essentially equivalent to explicit discrimination, in the sense that both (by stipulation) have roughly equivalent effects on the algorithm's predictive tendencies. This being the case, everything I said above regarding LC and BRT1's abilities to diagnose explicit discrimination applies directly to their ability to diagnose discrimination based on redundant encoding.

3: 'Redlining' – they describe this as 'a well known form of discrimination based on redundant encoding' (Dwork et al (2012): p224). Roughly, redlining occurs when institutions use a subject's area of residence as a proxy for their race as a way to deny financial services to racial minorities without engaging in explicit racial discrimination. Again, since redlining is an instance of discrimination based on redundant encoding, which in turn is statistically indistinguishable from explicit discrimination, there is not much more to say here about the respective abilities of BRT1 and LC to diagnose this type of unfairness. However, it should be noted that Eva (2022) demonstrates that BRT1 is better placed than many alternative group based statistical criteria of algorithmic fairness when it comes to diagnosing the unfairness of predictive profiles generated by redlining practices, and is able to identify redlining style biases in realistic cases.

4: 'Cutting off business with a segment of the population in which membership in the protected set is disproportionately high' – Dwork et al describe this as 'a generalization of redlining, in which members of S need not be a majority of the redlined population; instead, the fraction of the redlined population belonging to S may simply exceed the fraction of S in the population as a whole.' (Dwork et al (2012), p224). Again, this case does not seem to require separate treatment, since it is structurally analogous to standard redlining cases which, as noted above, are well diagnosed by BRT1.

5: 'Self fulfilling prophecy' – this is a case where the algorithm intentionally creates a bad track record for S by giving favourable treatment to the least qualified members of S. For example, imagine an algorithm that predicts mortgage defaults and that assigns a low risk score to a poorly qualified applicant from S and a high risk score to a well qualified applicant from S. If there are also well qualified applicants that achieve low risk scores and poorly qualified applicants that achieve high risk scores, then the algorithm will violate LC as long as the similarity metric deems the well (poorly) qualified applicant that is assigned a high (low) risk score to be similar to some well (poorly) qualified applicant that is assigned a low (high) risk score, since that would amount to treating two similar individuals in a dissimilar fashion. So LC is, in principle, capable of diagnosing this type of unfairness (given a suitable similarity metric).

In contrast, it's not hard to see that BRT1 cannot so straightforwardly diagnose unfairness in this type of case. Suppose we have a fair and accurate algorithm h that assigns reasonable risk scores to all members of S. Then take an algorithm h^* that is identical to h with the exception that it swaps the risk scores of a highly qualified applicant and a poorly qualified applicant. The average risk scores assigned to members of S by h and h^* are the same. So from the perspective of BRT1, h^* seems to treat S just as fairly (in comparison to the complement of S) as h does. While this might look like a victory for LC over BRT1, that impression is premature, for the following reasons.

Firstly, recall that in order for LC to be meaningfully applied, we first need to solve the problem of determining the similarity metric. Until that problem is solved, any promise regarding the diagnostic power of LC will ring hollow. And in the previous section, I argued that there is simply no promising way to solve that problem. So while proponents of LC can reasonably claim that, given a recipe for determining the similarity metric, the condition allows us to diagnose this species of unfairness, the fact that no such recipe exists renders this a Pyrrhic victory.

Secondly, note that BRT1 aims to diagnose when one group is treated unfairly in comparison to another. While it does seem reasonable to say that there is some kind of unfairness at play when the risk scores of a well qualified applicant from S and a poorly qualified applicant from S are switched (the well qualified applicant is being treated unfairly compared to the poorly qualified applicant), it doesn't seem that this unfairness necessarily counts as group level unfairness against S. For, S members are not being treated poorly (in comparison to non-members) in any kind of a systematic way. Some S members are being treated better than is warranted, and some are being treated worse than is warranted. Now, if the switching of risk scores were the result of a nefarious plan to produce negative data for members of S, then that would of course count as a group level injustice, but it's not the kind of injustice that could possibly be diagnosed by any statistical fairness criterion, since it is grounded not in the pattern of ill treatment, but rather in the ill intentions of the algorithm's proprietors.

Thirdly, I stress that I do *not* consider BRT (let alone BRT1) to constitute a sufficient condition for algorithmic fairness. Even if we restrict ourselves to talking about purely statistical criteria[13], I strongly suspect that BRT is not the only criterion that we should employ. And combining BRT1 with other group level statistical criteria will allow us to diagnose the relevant type of unfairness in many realistic cases. For instance, imagine that the risk scores (derived from an accurate and fair algorithm) of many qualified members of S are swapped with the risk scores of unqualified members of S. While this won't change the average risk score assigned to S overall, it will have a major impact on the overall profile of predictions for S. And this impact will be readily picked up by many alternative group level fairness criteria. For instance, the equal error rates, equal false positive rates and equal false negative rates crtieria from the introduction will all identify unfairness in this type of case, as would the influential calibration within groups criterion.[14] So while BRT1 alone may not be able to diagnose unfairness in this type of case, there is no need to resort to individual

[13] I think it's clear that there are some kinds of unfairness that cannot be diagnosed by any purely statistical criteria.

[14] For discussion of this criterion, see e.g. Hedden (2021), Kleinberg et al (2016), Miconi (2017), Pleiss et al (2017).

based fairness criteria, since other group level criteria are well equipped to handle such cases.[15]

6: 'Reverse tokenism' – Dwork et al describe this as a type of unfairness whereby the algorithm's proprietors assign an unjustifiably harsh prediction to a qualified member of S's complement (a 'token rejectee') in order to refute complaints that members of S are treated harshly in virtue of their membership of that group. Again, if LC had access to a suitable similarity metric, it could plausibly diagnose this type of unfairness as long as the token rejectee is deemed similar to some other highly qualified applicants that are not given unusually poor treatment. And again, it's not clear that BRT1 has such an easy time diagnosing unfairness in this type of case. However, the case is sufficiently analogous to the self fulfilling prophecy case discussed above that the three counterarguments listed there all apply equally to this case. Most pertinently I note that, as with the previous case, (i) LC's claim to diagnostic power is hollow in the absence of a solution to the problem of fixing the similarity metric, and (ii) in realistic instances of this type of unfairness, BRT1 can be paired with other group level criteria to successfully identify the algorithm's bias.

§5: Conclusion

Overall then, we've seen that LC suffers from a fundamental conceptual problem that admits of no promising solution, and that the kinds of unfairness that LC was designed to identify can all be diagnosed at least as well by group level statistical criteria. Where does this leave the project of developing an individual based conception of algorithmic fairness?

One lesson from the preceding analysis is that group based fairness criteria have the major advantage that they can utilise the rich internal statistical structure of groups, while individual based criteria cannot. From a statistical perspective, individuals are essentially atomic in the sense that they have no internal structure (i.e. no non-trivial base rates, average risk scores, false positive rates etc). LC attempts to compensate for this lack of structure by relying on a similarity metric, but, as we've seen, this strategy is ultimately unsuccessful since there is no principled and rigorous way to determine the similarity metric in concrete cases. Thus, the challenge for advocates of an individual based conception of algorithmic fairness is to find alternative ways of compensating for individuals' lack of internal statistical structure. While I don't have any idea of how this could be accomplished, I also don't see any fundamental conceptual obstacle that precludes the possibility of success. At the current juncture, however, group based criteria look like a more promising tool for the policing of bias and injustice in algorithmic prediction.

Bibliography

Afrouzi, A. On AI Judges. *Journal of Criminal Law and Criminology*, Forthcoming.

[15] Of course, the question of which other group level criteria should be adopted in conjunction with BRT is a difficult one.

Breiman, L. Statistical modeling: The two cultures. *Statistical science*, 16(3): 199–231, (2001).

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

D'Amour, A. (2021).Revisiting Rashomon: A Comment on 'The Two Cultures' *Observational Studies*, 7 (1): 59–63.

Diaconis, P. and S. Zabell (1982). Updating Subjective Probability. *Journal of the American Statistical Association* 77: 822-830.

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

Eva, B. (2022). Algorithmic Fairness and Base Rate Tracking. *Philosophy and Public Affairs*, 50(2): 239–266.

Eva, B., Hartmann, S. and Rafiee Rad, S. Learning from Conditionals. *Mind*, 129(514):461-508

Fleisher, W. (2021). What's Fair About Inidividual Fairness? *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics and Society*: 480–290.

Friedler, S. A.; Scheidegger, C.; and Venkatasubramanian, S. (2016). On the (Im)Possibility of Fairness. arXiv:1609.07236.

Gillen, S., Jung, C., Kearns, M. and Roth, A. (2018). Online Learning with an Unknown Fairness Metric. In *Advances in Neural Information Processing Systems, NeurIPS 2018*: 2605–2614.

Ilvento, C. (2020). Metric Learning for Individual Fairness. In *1st Symposium on Foundations of Responsible Computing, FORC20*, 3:1–20.

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

Lahoti, P., Gummadi, K. P., and Weikum, G. (2019). iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 1334–1345.

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

Mukherjee, D., Yurochkin, M., Banerjee, M., and Sun, Y. (2020). Two Simple Ways to Learn Individual Fairness Metrics from Data. In III, H. D. and Singh, A. eds. *Proceedings of the 37th International Conference on Machine Learning, Volume 119 of Proceedings of Machine Learning Research*: 7097–7107.

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.*

Wang, H. Grgic-Hlaca, N. Lahoti, P. Gummadi, K. P. and Weller, A. (2019). An Empirical Study on Learning Fairness Metrics for COMPAS Data with Human Supervision. arXiv:1910.10255.