Journal of
APPLIED PHILOSOPHY

Journal of Applied Philosophy
doi: 10.1111/japp.12778

Broomean(ish) Algorithmic Fairness?

CLINTON CASTRO

ABSTRACT *Recently, there has been much discussion of 'fair machine learning': fairness in data-driven decision-making systems (which are often, though not always, made with assistance from machine learning systems). Notorious impossibility results show that we cannot have everything we want here. Such problems call for careful thinking about the foundations of fair machine learning. Sune Holm has identified one promising way forward, which involves applying John Broome's theory of fairness to the puzzles of fair machine learning. Unfortunately, his application of Broome's theory appears to be fatally flawed. This article attempts to rescue Holm's central insight – namely, that Broome's theory can be useful to the study of fair machine learning – by giving an alternative application of Broome's theory, which involves thinking about fair machine learning in counterfactual (as opposed to merely statistical) terms.*

1. Introduction

Recently, there has been much discussion of 'fair machine learning': fairness in data-driven decision-making systems (which are often, though not always, made with assistance from machine learning systems).[1] The type of decision making under discussion and the reasons for the controversy surrounding it are exemplified by COMPAS,[2] software that predicts the likelihood that a defendant will commit a crime in the future on the basis of 137 data points about them.[3] These predictions factor into high-stakes decisions, such as those relating to setting bond amounts and determining criminal sentences.[4]

A bombshell report by ProPublica in 2016 found that COMPAS, despite not collecting data directly about defendants' race, 'was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants'.[5] ProPublica concluded from this that COMPAS is biased against Black defendants. Northpointe (now equivant), the company that developed COMPAS, saw things differently. They responded that it is wrong to infer that the system is biased against Black defendants from the fact that it has a higher false positive rate for Black defendants. This, they claim, is because unequal false positive rates[6] between groups is not proper evidence of bias.[7] According to them, we should instead look at predictive values – such as *positive predictive value*, that is, the rate at which those who are predicted to reoffend actually go on to reoffend[8] – which COMPAS does equalize across racial lines.[9]

Ever since, one major theme of the discussion over fair machine learning has been whether fairness demands that data-driven decision systems satisfy ideals of *error parity* (ProPublica's preferred standard) – that is, having similar false positive and/or negative rates across groups of interest (e.g. racial and ethnic groups)[10] – or *predictive parity* (Northpointe's preferred standard) – that is, having similar positive and/or negative predictive values across groups of interest.[11] As various impossibility results have shown, it is typically impossible for a system to simultaneously satisfy both types of these ideals, so we seem to have to choose between them.[12]

Sune Holm suggests that we can deploy a well-established theory of fairness – namely, John Broome's theory of fairness[13] – to see our way through this thicket.[14] Holm is right that Broome's theory has much to offer. However, Holm's application of Broome's theory has significant shortcomings.[15] That is where this article comes in: it provides an application of Broome's theory that avoids the pitfalls that Holm's falls into. Among the upshots of this new application is the idea that proper thinking about fair machine learning will involve resisting the tendency to think about fair machine learning in merely statistical terms, which is typically how *parity criteria* – that is, both error parity and predictive parity – are employed. Instead, we must, as some have proposed,[16] think about these issues in different – for instance, counterfactual – terms.

The article begins with an introduction to Broome's theory of fairness and Holm's application of it. It then presents two criticisms of Holm's application of Broome's theory. It next formulates an application of Broome's theory that evades these problems. The article concludes by taking stock of what this new application of Broome's theory has to offer to our understanding of fair machine learning.

2. Ground Clearing

2.1. Broome on Fairness

The heart of Broome's theory[17] is the:

> *Fairness Principle*: 'fairness requires that claims should be satisfied in proportion to their strength'.[18]

While the bit of his theory that applies to our discussion has to do with circumstances where the Fairness Principle *cannot* be satisfied, it will be helpful to first clarify what it demands.

Begin with the idea of a claim. It is helpful to think of 'claims', as it appears in the Fairness Principle, as a placeholder for any reasons to give a candidate a good which are grounded in 'duties *owed to the candidate herself*'.[19] Broome does not offer a general theory of claims but does give some examples: needs, general rights, and debts of gratitude.[20] At the level of abstraction we are working at, the general theory of claims – whatever it might turn out to be – is largely irrelevant to applying Broome's theory to the matters at hand, so it does not need to be discussed any further for present purposes.

Turning now to the idea of satisfying claims in proportion to their strength, it is helpful for what comes later to understand this as involving the following four intuitive ideas. First, claims can be stronger or weaker. For instance, if some people are hungry and similar in all relevant respects except that one has more need than the others, then it makes intuitive sense to say that this particularly needy person has a stronger claim to food that is being given out. Second, when it comes to claims of similar strength, similarly strong claims should – from the perspective of fairness – enjoy similar levels of claim satisfaction. Third, those with stronger claims should get higher levels of claim satisfaction than those with weaker claims. Fourth, higher levels of satisfaction should be proportionate to the strength of claims (though we should perhaps not take 'proportion' too precisely).[21]

Importantly for our discussion, there are cases where hewing to the fairness principle is not, all things considered, the right thing to do. Some goods, such as kidneys, are non-divisible. And if we have a kidney and two individuals with equally strong claims to it, it is clear that we should distribute the kidney, thereby violating the fairness principle.

Broome is happy to admit this, as it is not the case that the only normatively loaded thing to say about claims is that they should be satisfied *in proportion to their strength*. He thinks that *fairness* demands this. But he also thinks that *claims should be satisfied*; this is what *goodness* demands.[22] In the case at hand, it is clear that goodness and fairness do not coincide. Indeed, Broome readily admits that there can be tension between goodness and fairness: sometimes we must sacrifice some goodness to be fair, and sometimes we must sacrifice some fairness to be good.

In situations such as the kidney case – where fairness must yield to goodness – we can still 'meet … the requirement of fairness to some extent'.[23] Indeed, we can achieve a 'surrogate'[24] satisfaction of fairness by, for example, holding a lottery where the objective chances of receiving the good is proportional to the strengths of claims.[25] If we distribute the good, we will, by the lights of the fairness principle, act unfairly. But to act in accordance with the fairness principle would be wrong anyhow. In situations like these, the surrogate satisfaction of fairness is to be our guiding light in the compromise between fairness and goodness.

2.2. *Holm on Broome's Theory*

Holm sees the idea of the surrogate satisfaction of fairness as ripe for application to debates in fair machine learning. His sense is that the parity criteria are on to something, as both can be seen as drawing on the intuition embodied in the fairness principle.

Ideals of error parity can be seen as demanding that those with similar claims to a good (say, the innocent with respect to being found not guilty) should not have different chances of receiving the good due to, for example, belonging to different racial groups. To illustrate further, one's chance of a false positive – say, being found guilty when one is innocent – should not be different for a Black defendant and a White defendant just because one is Black and the other White.

Something similar can be said of ideals of predictive parity: actuarial predictions of, say, being 10% likely to be in a car crash in the next year should not more closely track the frequency with which one gets into car accidents simply because one belongs to one group (say, Black drivers) and not another (say, White drivers).

Holm goes on to argue that Broome's theory favors ideals of error parity over those of predictive parity. His reasons will not concern us here; the debate over which parity criterion is favored by Broome's theory is beyond the scope of this article. Instead, what will concern us is *how* Holm interprets the criteria while adopting a Broomean point of view.

Holm's account relies on the crucial assumption that, as Castro and Loi put it, the 'parity criteria (applied to the proper subset of candidates and goods) ensure that candidates have equal chances of getting the goods'.[26] That is to say, Holm takes it that, for example, if Black defendants and White defendants get falsely convicted at similar rates, then *each* Black and White defendant has the same chance of a false conviction. Now, were this true – were it the case that, for example, the false positive rate for one's group just was their chance of getting a false positive – then we could seamlessly use Broome's theory to underwrite the use of certain criteria (e.g. as showing that having unequal false positives across racial groups is (surrogate) unfair) once we can sort out who, in any given case, is the true subset of claimants.

2.3. Two Problems for Holm

Unfortunately for Holm, we cannot seamlessly apply Broome's theory in this way. Here, I will present two problems for Holm's application of Broome's theory: the equal probabilities talk problem,[27] and the claim-strength sensitivity problem.[28]

The equal probabilities talk problem is the observation that group-level summary statistics such as the false positive rate do not generally represent each group member's chance of, say, receiving a false positive. To paint an illustrative example, suppose that determinism is true. And suppose we use COMPAS – which, let us assume, is highly accurate but imperfectly so (namely, it will falsely flag some people as future offenders) – to generate some predictions about who will go on to commit a crime. In this case, each individual's objective chance of receiving the prediction 'will go on to commit a crime' is either zero or one. But the false positive rate will be neither zero nor one, as it is calculated by taking the number of members of the group who received a false positive (which will be greater than zero) and dividing by the number of members of that group who did not go on to commit a crime (which will be much, much greater than zero). So, the false positive rate will not align with *any* individual's objective chance of receiving a false positive: it will be *between* zero and one, but their chance of getting a false positive will be *either* zero or one.

While this stylized example is extreme, it is representative of many systems under consideration in the discussion about fair machine learning, as the systems are deterministic in the following sense: once a subject's inputs to the system are assigned, the output is – for all intents and purposes – already decided (as the system just performs a function that maps inputs onto outputs). Given that Holm relies on an inference from, for example, group-level false positive rates to individual-level chances of receiving false positives, the equal probabilities talk problem seems to identify a fatal flaw in his application of Broome's theory. It means that we can no longer use Broome's theory in the elegant way that Holm suggests we can (i.e. affirming – with Broome – that (surrogate) fairness can be satisfied by equalizing certain objective chances and using group-level summary statistics (e.g. false positive rate) as describing the relevant chances).

The claim-strength sensitivity problem is based on another simple observation. Often, the subjects of decision-making systems will have claims of different strengths to the good being distributed. For example, suppose that we use COMPAS for more ameliorative purposes than have been discussed above. Suppose that we are allocating vouchers for therapy services and that the number of crimes one will go on to commit is relevant to one's need-based claim to those services (with greater numbers constituting greater need). The parity criteria advocated by Holm won't apply here, as they purport to tell us how to properly equalize chances among those with *similar* claims. This observation needn't be disastrous for fans of the parity criteria: we could always supplement these with criteria that are sensitive to the fact that we also need an account of what to do with *dissimilar* claims. However, extending Holm's account in the most natural way will not work. That would involve saying, for instance, that we want lower false negative rates among those who would go on to commit more crimes, as this would be constitutive of giving those individuals a greater objective chance of the good. However, in light of the equal probabilities talk problem, this extension of Holm's account will not work. To have a satisfactorily complete Broomean account, then, we need to address the claim-strength sensitivity problem, and we must do this in a way that does not run afoul of the core issues at play in the equal probabilities talk problem.

3. An Alternative Application of Broome's Theory

3.1. Evading the Equal Probabilities Talk Problem

I will begin my development of an alternative application of Broome's theory by focusing on a method for coping with the equal probabilities talk problem.

For the time being, I will bracket the question of satisfying claims in proportion to their strength (we will return to this later). Let us, then, focus on the modest question of how we can guarantee the surrogate satisfaction of claims *among those with claims of similar strength*, without availing ourselves of the idea that group-level statistics are a direct guide to individual-level chances.

Taking inspiration from Kusner *et al.*,[29] we can find what we are looking for by thinking in counterfactual terms.[30] The basic idea here is that we can say that similar claims have been treated similarly if predictions (e.g. of innocence or guilt) do not 'listen to' (i.e. are not influenced by) the wrong sorts of variables (e.g. one's race).[31] And we can test for this by considering certain counterfactuals.

The following case might help in understanding the general idea being proposed. Return to the kidney example and suppose that we implement some mechanism to randomly[32] choose a recipient. But suppose also that determinism is true. Does this mean that the lottery (and every other lottery) is *ipso facto* (surrogate) unfair? I don't think so, though it could have been: for example, someone could have rigged it to guarantee a certain outcome (e.g. that their friend would win). How, then, could determined outcomes and (surrogate) fairness be compatible? The key detail is that in the fair but determined lottery, anyone who lost did not lose *because of*, for example, who they know. Relating this to the above idea of thinking counterfactually, suppose that a patient who has friends in the hospital wins that hospital's lottery for a kidney, and suppose that because of this there are concerns about bias. We could demonstrate that the system wasn't biased in the patient's favor in virtue of who he knew by showing that, for all who lost, it is not the case that they would have won were they, instead, to have been friends with the people at the hospital. More generally, it seems that we can have a fair lottery even if the *chances* are extreme (all zeros and ones), so long as the distribution of chances is not sensitive to the wrong sorts of attributes. Further, one way to see that a distribution of extreme chances is not sensitive to the wrong sorts of attributes is to consider certain counterfactuals.

Before delving further into this thought, let me pause to make a few notes about what it means for an attribute to be of the wrong sort. Assume that in the kidney lottery example, the potential kidney recipients all have an equal claim to the kidney. Assume further that in this particular context the only relevant type of claim is need. Thus, who someone is friends with is irrelevant from the perspective of fairness. This is what makes who someone is friends with the wrong kind of attribute to be sensitive to. In some cases, it might, in fact, be fair to consider special relationships. But, by stipulation, it does not ground a claim to the kidney in this case.

Now, there are many attributes – call them ‘protected attributes’ (e.g. race and gender) – that are typically assumed as being the wrong sorts of attributes to be sensitive to in the context of data-driven decision making. Before moving on, let me say a few words about how I think the Broomean framework that we are working with interacts with this assumption. On the Broomean framework, these attributes will, indeed, *often* be the wrong sorts of attributes to be sensitive to. When these are the wrong sorts of attributes to be sensitive to, it will – per Broome’s theory – be that they are, for example, not relevant to claims in the given context, much in the way that friendship is irrelevant in the kidney case. So, in those cases, to be sensitive to race is to be sensitive to the wrong sort of attribute.

Being sensitive to the wrong sorts of attributes isn’t, of course, the only way for a system to be unfair. It can also be unfair in *how* it relates to those attributes. I mention this because there could very well be cases in which protected attributes are relevant to claims. The ultimate deciding factor for whether an attribute is the wrong sort to be sensitive to is whether it is grounds for a claim, and I see no reason protected attributes couldn’t be, from the perspective of fairness, the right ones to be sensitive to in a range of cases. Exactly when and how race and gender might relate to claims in such cases is, of course, an extremely complicated topic, one that would take us beyond the scope of the present article (and, further, one that Broome’s theory does not give us the resources to settle fully, as Broome does not provide us with a complete theory of claims). This is worth mentioning because I will, mostly for reasons of simplicity, make the typical assumption that protected attributes will be the wrong sorts of attributes to be sensitive to. But, to be clear, this is not because it is categorically wrong to be sensitive to them from a Broomean point of view. The key thing, for us, is that it is unfair for a system to be sensitive to the wrong sorts of attributes, and in many cases (though not always) this will align with the common assumption that we do not want data-driven decision-making systems to be sensitive to protected attributes.

Relating this now to the technical side of the literature on fair machine learning, one example of how we might implement these ideas can be found in the methods proposed by Kusner *et al.*, which they describe as *counterfactual fairness*. The main idea at play in counterfactual fairness is that systems’ predictions about an individual shouldn’t change if that individual’s protected attributes were otherwise.[33]

The specific procedure they propose as a test for counterfactual fairness can be illustrated by considering a stylized case.[34]

> *Law School Success*: You are building a system to decide who to admit to your law school. First-year law grades are predictive of professional success, and via the magic of big data, you are able to predict first-year law grades. You therefore decide to admit students on the basis of their projected first-year grades. Your system is able to predict first-year grades on the basis of an applicant’s undergraduate GPA and their LSAT score.

How might we test the system for counterfactual fairness?

One way is to appreciate the causal effects that race, gender, and any other variables of concern have on the variables we are considering when making our predictions (e.g. undergraduate GPA and LSAT score).[35] Let us imagine that the causal nexus as it relates to the variables can be modeled as shown in Figure 1.[36]

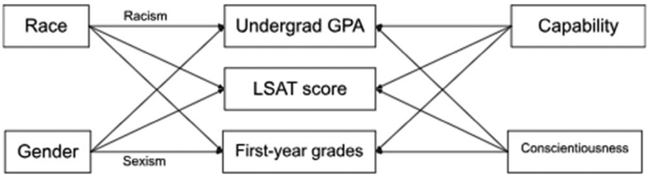

Figure 1. Causal Model for Law School Success

This Figure 1 encodes the fact that the system in Law School Success is counterfactually unfair.

Consider two applicants counterparts of each other *iff* they have the same values for variables *not* influenced by protected attributes (while keeping in mind that this is a highly simplified case and account, which is being used to illustrate key features of taking a counterfactual – as opposed to purely statistical – approach to fair machine learning). In this case, to be one's counterpart is to be equally conscientious and capable as them.

Given this definition of counterparthood, there will (at least hypothetically) be cases where two counterparts get different results. For instance, we can imagine two applicants, one White and the other Black, who are counterparts but have different projections of first-year grades because the Black applicant has a lower GPA (as a result of, say, stereotype threat).

To give a clearer sense of how we might use a causal model to show this, let us assume that the edges and nodes (i.e. the boxes and arrows) depicted in Figure 1 represent precise equations that would allow us to solve for certain unknown variables when others are known. With these equations, we could show that the Law School Success system is unfair in the following way. Suppose in the first instance that we consider the file of some applicant. Call them Applicant One. The Law School Success algorithm takes as inputs LSAT score and undergraduate GPA. So we feed it that information. Suppose the LSAT score is 150 and the GPA 3.5, and that, on the basis of these inputs, the system predicts a first-year law grade average of 3.8, which is above the desired cutoff of 3.7. Thus, Applicant One is admitted.

But now suppose that the following is true. Suppose that Applicant One is a White man, and suppose he is of average capability and conscientiousness, say 7.5 on a scale of one to ten for each. Suppose further that, given the equations that the edges and nodes represent, the following is true. If we enter 'Black' for race, 'woman' for gender, '7.5' for capability, and '7.5' for conscientiousness, we could solve the equations for GPA and LSAT, yielding 3.0 (as opposed to 3.5) and 145 (as opposed to 150), respectively. Suppose further that if we then put these figures (i.e. 3.0 and 145) into the algorithm, we get a projection of 3.5, yielding a rejection. This is what the system's being counterfactually unfair consists in: it treats counterparts differently.

As this simplified example hopefully makes clear, we can use counterfactual reasoning to meaningfully ask and answer questions such as, 'Would this algorithm admit this applicant's counterparts?' In the case of Law School Success, we saw what it looks like in some detail when the answer is 'no'. As the example is intended to demonstrate, one way to do this is to have a rich understanding of how certain variables relate to each other causally and a sense of what you want to use as a definition of counterparthood (in this case, being equally capable and conscientious).

In this case, we can go from a judgment of counterfactual fairness to a judgment of Broomean fairness if we add the assumption that someone's claim to a slot in law school depends only on how capable and conscientious they are. To be clear, this is a simplifying assumption meant to illustrate how one might use the tools of counterfactual fairness to operationalize Broomean fairness. It is not intended as a serious proposal of what the relevant claims actually are or anything like substantive guidance for the real world.[37]

A realistic application of Broome's theory, then, would have to determine which claims are relevant to the determination being made (which, in turn, would help determine which features – e.g. race, sex, etc. – to control for). We will not explore those details here, as they are beyond the scope of this article. As mentioned earlier in this article, 'claims', as it appears in the Fairness Principle, is a placeholder. And Broome, recall, does not offer a general theory of claims. Further, it is not the business of this article to determine, for example, what fairness in law admissions amounts to (which is a very difficult question in its own right). Instead, it is the task of this article to understand how to assess a data-driven decision-making system for fairness, whatever the relevant claims turn out to be.

Hopefully none of this obscures the important contours of this aspect of the proposal, which is that even in situations with extreme (i.e. 0, 1) objective chances, there might still be a surrogate for fairness (which has to do with whether outcomes are caused in the right way) and there are technical approaches to fair machine learning that at least in principle can track this.

3.2. Evading the Claim-Strength Sensitivity Problem

Let us now turn to the claim-strength sensitivity problem, the problem of coping with the fact that claims come in varying degrees of strength. To make matters concrete, let us return to Law School Success. The preceding discussed some of the initial moves that would be involved in treating similar cases similarly. How would we, even in principle, see to it that different cases get treated differently and in (rough) proportion to their difference in a way that would satisfy Broome's theory?

Here we can turn to a modalized account of risk. To fix ideas, we could understand modal risk as follows:

> *Modal Risk*: Risky events are potential unwanted events where the degree of risk involved relates to how modally close, on average, the unwanted event is with respect to nearby possibilities.[38]

While Broomeans needn't accept this account of risk – certainly, there are different ways to build a modalized account of risk, sorting out the fine details of which is a task for another article – the proposal makes clear the advantages that a modalized account of risk has to the Broomean: it directs us towards a hospitable surrogate for (or interpretation of)[39] 'objective chance' in surrogate fairness.

This idea can be adapted to the claim-strength sensitivity problem in the following fashion.

First, we can understand systems as being geared towards judgments that are more or less risky with respect to certain outcomes, such as false positives. To see this, we can compare two systems for finding defendants culpable. One instructs jurors to base their decision on the preponderance of the evidence, that is, to rule against the defendant *iff* they deem that, on the basis of the available evidence, it is more likely than not that the accused is culpable. Another instructs jurors to base the decision on the 'beyond a reasonable doubt' standard, that is, to rule against the defendant *iff*, on the basis of the available evidence, no reasonable person would think the defendant isn't culpable. With respect to the risk of falsely being found culpable, the 'beyond a reasonable doubt' standard is the less risky of the two: adopting it means adopting a system that errs on the side of mistakenly finding defendants innocent.[40]

Second, we can understand the strength of claims as involving claims to judgments that are more or less risky. To see this, consider a large difference between criminal and civil cases. In losing a civil suit, one might be ordered to pay a sum of money. In losing a criminal suit, one might be ordered to go to prison. In virtue of this difference, it is natural to think of innocent defendants in criminal cases as generally having a stronger claim to not being falsely found culpable than innocent defendants in civil cases. Relating this now to the burdens of proof example in the previous paragraph, these observations make it fitting for criminal cases to be held to the higher burden of proof (i.e. 'beyond a reasonable doubt').

How these ideas relate to Broome's theory and fair machine learning can be clarified by considering a simple lottery case.

> *Lottery*: We are distributing an indivisible good among people with claims of different strengths. Some people have claims of strength 1*x*, others of strength 2*x*. We give the 1*x*-ers each one ticket and the 2*x*-ers two. We then randomly select tickets to determine winners of the lottery.

In Lottery, we should be on track to being (surrogate) fair by Broomean standards. We might even be on track if determinism is true. Even though the chances would in that case be extreme, 2*x*-ers get more goods in most nearby worlds than 1*x*-ers. This, it seems to me, goes at least some way towards being fair in the same (or nearly the same) way that getting a *chance to a good* as opposed to a *good itself* does.[41]

How might we account for this in machine learning systems? One way is to take care to ensure that, for example, among those who got false negatives, they receive a true positive in more nearby worlds (conditional on the system working the way that it does in the actual world), than those who shouldn't have received a positive (also conditional on the system working the way that it does in the actual world). For instance, let us suppose that we are building the Law School Success algorithm in a much fairer world. Namely, let us suppose that no protected attribute influences LSAT scores. Suppose further that we can consider LSAT scores in one of two different ways: we could consider applicants' one best LSAT score (with students being required to take the exam exactly three times) or we could consider applicants' two best scores (while still requiring students to take the exam exactly three times). Assuming that errors (misleading test scores) are randomly distributed and that qualified applicants will more reliably achieve higher LSAT scores and have stronger claims to admission, we might, in the name of fairness, choose the 'two best' testing regime: this would put our false negatives closer to more worlds where they are true positives.

Now, one might think that this is no different from focusing on error rates in the way that Holm would. But there is an important difference. The modalized risk account will be sensitive to factors that go beyond error rates. This is because we will need to have some assurance that mistakes are not modally robust in the wrong ways. Imagine, for instance, that it turns out that 'two best' and a system that only consults parental levels of education are equally accurate at the group level when it comes to predicting capacity and conscientiousness.[42] One reason (among many others) to adopt the 'two best' regime is that under it, students who are misclassified could have more easily been properly classified: in more nearby worlds their performance matches their level of qualification, as errors were assumed to be randomly distributed. But under the parental education regime, this is not the case: we can assume that for most applicants, their capacities, their level of conscientiousness, and their parents' level of education are more tightly coupled across nearby worlds. In other words, errors are more modally robust.

4. Conclusion

This has been the opening salvo in developing a framework for thinking about a fair machine from a Broomean point of view. Much space has been dedicated to the fundamentals of applying Broome's theory to questions of fair machine learning. Less has been given to relating this approach to the parity criteria. Let us briefly turn to that now.

On the Holmian interpretation, we can see the rate at which a system assigns false positives to, say, innocent Black defendants as defining the chance that any given innocent Black defendant has of being falsely viewed as guilty. On this interpretation, were it to work, the thing that we want (fair chances) and something we can observe (group-level ratios) are one and the same thing, making the detection of fairness a straightforward task (at least in principle). Unfortunately, group-level ratios and individual-level chances don't relate to each other in the way needed for this to work.[43]

On my alternative interpretation, we want each individual-level outcome to be caused in the right way (i.e. by systems that manage risk in the right way and are not inappropriately influenced by, for example, race). Unfortunately, this is something that – unlike group-level ratios – we cannot see. This means that the detection of fairness will have to be indirect. For instance, we most certainly *can* use group-level ratios, such as false positive rates, to make inferences about whether a system seems to be 'listening to', for example, individuals' race. This will be an inexact science, but the hope of this article is that if we better understand what we are after – that is, if we realize that we are not after certain ratios *themselves* but are instead interested in understanding what they are evidence *for* (i.e. causal structures that violate Broomean standards of (surrogate) fairness) – we will be better positioned to effectively use the evidence that we can gather (e.g. group-level statistics).

Clinton Castro, The Information School and Department of Philosophy, University of Wisconsin–Madison, Madison, WI, USA. clinton.g.m.castro@gmail.com

Acknowledgements

I am grateful to Jimmy Goodrich, David Gray Grant, Milo Phillips-Brown, Annette Zimmerman, participants at the Oxford Workshop on Algorithmic Bias, the Philosophy of Data & AI work in progress group at UW–Madison, Jain Family Institute Digital Ethics Work-in-Progress group, the 2024 Pacific Division meeting of the American Philosophical Association, and the Research Center in Minority Institutions at FIU, and two anonymous reviewers for helpful feedback that greatly improved the quality of this article.

NOTES

[1] See e.g. Castro *et al.*, "Egalitarian Machine Learning"; Castro and Loi, "Fair Chances"; Eva, "Algorithmic Fairness"; Fleisher, "What's Fair"; Grant, "Equalized Odds"; Hedden, "On Statistical Criteria"; Hellman, "Measuring"; Holm, "Fairness"; Johnson, "Algorithmic Bias"; Kusner *et al.*, "Counterfactual Fairness"; Long, "Fairness"; Wong, "Democratizing." See Fazelpour and Danks, "Algorithmic Bias."

[2] COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions.

[3] Angwin *et al.*, "Machine Bias."

[4] Ibid.

[5] Ibid.

[6] The false positive rate (FPR) for a group is the number of individuals in that group who are *falsely predicted* as *having* the trait that is being predicted (e.g. future criminal behavior), divided by the total number of individuals in that group who *actually* do *not* have the trait.

[7] Northpointe Inc., "COMPAS Risk Scales."

[8] The positive predictive value (PPV) for a group is the number of individuals in that group who are *correctly predicted* as *having* the trait that is being predicted (e.g. future criminal behavior), divided by the total number of individuals in that group who are *predicted* to *have* the trait.

[9] Northpointe Inc., "COMPAS Risk Scales."

[10] The false negative rate (FNR) for a group is the number of individuals in that group who are *falsely predicted* as *not* having the trait that is being predicted (e.g. future criminal behavior), divided by the total number of individuals in that group who *actually do* have the trait.

[11] The negative predictive value (NPV) for a group is the number of individuals in that group who are *correctly* predicted as *not* having the trait that is being predicted (e.g. future criminal behavior), divided by the total number of individuals in that group who are *predicted* to *not* have the trait.

[12] See Chouldechova, "Fair Prediction," and Kleinberg *et al.*, "Inherent Trade-Offs," for proofs of this claim.

[13] See Broome, "Selecting"; Broome, "V*—Fairness."

[14] Holm, "Fairness."

[15] Castro and Loi, "Fair Chances."

[16] For example, Kusner *et al.*, "Counterfactual Fairness."

[17] Let me here flag that my interpretation of Broome's theory has benefited greatly from Piller, "Treating Broome Fairly."

[18] Broome, "V*—Fairness," 95.

[19] Ibid., 115.

[20] Broome, "Selecting," 44.

[21] Here is the full quotation on which this is based: 'I do not mean "proportion" to be taken too precisely. But I do mean that equal claims require equal satisfaction, that stronger claims require more satisfaction than weaker ones, and also – very importantly – that weaker claims require some satisfaction. Weaker claims must not simply be overridden by stronger ones' (Broome, "V*—Fairness," 98).

[22] Ibid.

[23] Ibid., 98.

[24] Ibid., 98.

[25] Note, in keeping with the above, that the surrogate satisfaction of fairness does not demand that the chances be 50/50. Indeed, surrogate fairness is satisfied so long as their chances are *the same*. On Broome's view we should opt for 50/50 – as opposed to 49/49 and a 2% chance that we simply destroy the kidney – not for reasons of fairness, but for reasons of goodness.

[26] Castro and Loi, "Fair Chances," 334. This reading of Holm is supported by passages such as the following, in which Holm is discussing his running example of a hypothetical algorithm ('DR-A') that general practitioners might use to decide whether their patients should see a specialist for diabetic retinopathy (DR): 'As I have interpreted the statistical fairness criteria, they claim that an algorithm is fair in virtue of ensuring that all claim holders have the same *chance* of a decision to grant them a good. E.g., Equal FNR [i.e. Equal False Negative Rate] states that DR-A is *fair* if and only if all DR-patients have the same *chance* of a positive decision' (Holm, "Fairness," 274). Elsewhere in the paper – when he is speaking about the four ideals discussed above (i.e. equal false positive rate, equal false negative rate, equal positive predictive value, and equal negative predictive value) – he says, 'The four criteria agree that fairness requires that all members of a certain subgroup of the population … should have an equal chance of a decision based on a true prediction across socially salient groups. However, the criteria disagree about which sense of equal chances is morally relevant' (ibid., 268). He

makes similar claims (i.e. claims that conflate group-level ratios and individual-level chances) elsewhere in the paper (e.g. on pages 273, 275, 277).

[27] Castro and Loi, "Fair Chances." For a response to the equal probabilities talk problem that is fundamentally different than the one pursued here – instead of trying to shore up the application of Broome, the authors approach the problems Holm is concerned with but from a Rawlsian vantage point – see Castro and Loi, "Representative Individuals."

[28] Original to this article.

[29] Kusner *et al.*, "Counterfactual Fairness."

[30] Though, to be clear, I don't think that this is the only way to go. For instance, there might be a purely information-theoretic way to go as well.

[31] Important caveat: these examples aren't meant to imply that race and claim satisfaction always should or can be independent. For an incisive discussion of this point as it relates to fair machine learning, see Hu, "What is 'Race'?"

[32] Or 'randomly' if, strictly speaking, being random and determined is incompatible.

[33] Cf. Kusner *et al.*, "Counterfactual Fairness."

[34] What follows is based on their 'Law School Success' example, though some inessential details have been slightly modified for presentational purposes. Please note that the case is incredibly simplified and not at all intended to model the complexities of the real world.

[35] Kusner *et al.*, "Counterfactual Fairness."

[36] The nodes (i.e. boxes) in this graph denote variables. The edges (i.e. arrows) denote causal influences, flowing in the direction of the arrow. So, for example, race causally influences GPA (via racism) but GPA does not causally influence race.

[37] For helpful discussion of these issues more fit for contemplation of real-world cases, see Zimmermann and Lee-Stronach, "Proceed with Caution."

[38] This formulation closely resembles that of Pritchard, "Risk," but it inserts some key differences (namely it adds the idea of an average over nearby possibilities). This is in no way meant as a challenge to Pritchard's account. The insertion is there simply to facilitate the application of Broome's theory to this particular setting. Whether this formulation or Pritchard's (or some other) is correct is beside the point for present purposes. The key idea here is to illustrate a family of options for Broomeans: modality (as opposed to – or as an interpretation of (see the following footnote) – probability).

[39] Whether this is better understood as a surrogate or interpretation is immaterial for the purposes of this article; I mention both simply to be non-committal: I am indifferent as to whether we understand this as an interpretation or surrogate.

[40] Lillquist, "False Positives."

[41] Though, admittedly, I think there is more room for discussion of this matter to be had.

[42] Importantly, assume that these two systems *do not* pick out the same individuals as qualified. They are simply predictively equivalent in terms of overall accuracy at the group level.

[43] Cf. Castro and Loi, "Fair Chances."

References

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

Broome, John. "Selecting People Randomly." *Ethics* 95, no. 1 (1984): 38–55.

Broome, John. "V*—Fairness." *Proceedings of the Aristotelian Society* 91, no. 1 (1991): 87–102.

Castro, Clinton, and Michele Loi. "The Fair Chances in Algorithmic Fairness: A Response to Holm." *Res Publica* 29, no. 2 (2023): 231–37.

Castro, Clinton, and Michele Loi. "The Representative Individuals Approach to Fair Machine Learning." Unpublished Manuscript. (2024).

Castro, Clinton, David O'Brien, and Ben Schwan. "Egalitarian Machine Learning." *Res Publica* 29, no. 2 (2023): 237–264.

Chouldechova, Alexandra. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5, no. 2 (2017): 153–163.

Eva, Benjamin. “Algorithmic Fairness and Base Rate Tracking.” *Philosophy and Public Affairs* 50, no. 2 (2022): 239–266. https://doi.org/10.1111/papa.12211.

Fazelpour, Sina, and David Danks. “Algorithmic Bias: Senses, Sources, Solutions.” *Philosophy Compass* 16, no. 8 (2021): e12760. https://doi.org/10.1111/phc3.12760.

Fleisher, Will. “What’s Fair about Individual Fairness?” In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (July 2021)*, 480–490. Virtual Event: Association for Computing Machinery, 2021. https://doi.org/10.1145/3461702.3462621.

Grant, David G. “Equalized Odds is a Requirement of Algorithmic Fairness.” *Synthese* 201, no. 3 (2023): 101.

Hedden, Brian. “On Statistical Criteria of Algorithmic Fairness.” *Philosophy and Public Affairs* 49, no. 2 (2021): 209–231. https://doi.org/10.1111/papa.12189.

Hellman, Deborah. “Measuring Algorithmic Fairness.” *Virginia Law Review* 106, no. 4 (2020): 811–866.

Holm, Sune. “The Fairness in Algorithmic Fairness.” *Res Publica* 29, no. 2 (2023): 265–281.

Hu, Lily. “What is ‘Race’ in Algorithmic Discrimination on the Basis of Race?” *Journal of Moral Philosophy* 1 (2023): 1–26.

Johnson, Gabbrielle M. “Algorithmic Bias: On the Implicit Biases of Social Technology.” *Synthese* 198, no. 10 (2020): 9941–61. https://doi.org/10.1007/s11229-020-02696-y.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. “Inherent Trade-Offs in the Fair Determination of Risk Scores.” In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. *Leibniz International Proceedings in Informatics SchlossDagstuhl–Leibniz-Zentrum für Informatik*, 43:1–43:2. Germany: Dagstuhl Publishing, 2016.

Kusner, Matt, Joshua Loftus, Chris Russell, and Ricardo Silva. “Counterfactual fairness.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*, 4069–4079. Red Hook, NY: Curran Associates Inc, 2017.

Lillquist, Erik. “False Positives and False Negatives in Capital Cases.” *Indiana Law Journal* 80 (2005): 11.

Long, Robert. “Fairness in Machine Learning: Against False Positive Rate Equality as a Measure of Fairness.” *Journal of Moral Philosophy* 19, no. 1 (2021): 49–78.

Northpointe Inc. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity Performance of the COMPAS Risk Scales in Broward County 2016. https://www.semanticscholar.org/paper/COMPAS-Risk-Scales-%3A-Demonstrating-Accuracy-Equity/cb6a2c110f9fe675799c6aefe1082bb6390fdf49

Piller, Christian. “Treating Broome Fairly.” *Utilitas* 29, no. 2 (2017): 214–238. https://doi.org/10.1017/S0953820816000303.

Pritchard, Duncan. “Risk.” *Metaphilosophy* 46 (2015): 436–461. https://doi.org/10.1111/meta.12142.

Wong, Pak-Hang. “Democratizing Algorithmic Fairness.” *Philosophy and Technology* 33, no. 2 (2020): 225–244. https://doi.org/10.1007/s13347-019-00355-w.

Zimmermann, Annette, and Chad Lee-Stronach. “Proceed with Caution.” *Canadian Journal of Philosophy* 52, no. 1 (2022): 6–25.