Post Hoc Interventions and Machine Bias

Kasper Lippert-Rasmussen

In Gunnemyr, Mattias & Jönsson, Martin L. (2023) *Post Hoc Interventions: Prospects and Problems.*
Lund: Department of Philosophy, Lund University. https://doi.org/10.37852/oblu.184

ISBN: 978-91-89415-60-7 (print)
978-91-89415-61-4 (digital – pdf)
978-91-89415-62-1 (digital – html)

DOI: https://doi.org/10.37852/oblu.184.c508

Post Hoc Interventions
Prospects and Problems

Published by the Department of Philosophy, Lund University.
Edited by: Mattias Gunnemyr and Martin L. Jönsson
Cover layout by Cecilia von Arnold, Pufendorf Institute for Advanced Studies

8

Post Hoc Interventions and Machine Bias

Kasper Lippert-Rasmussen[1]

Abstract. In a US context, critics of court use of algorithmic risk prediction algorithms have argued that COMPAS involves unfair machine bias because it generates higher false positive rates of predicted recidivism for black offenders than white offenders. In response, some have argued that algorithmic fairness concerns calibration across groups – roughly, that a score assigned to different individuals by the algorithm involves the same probability of the individual having the target property across different groups of individuals – only. I argue that in standard non-algorithmic contexts, such as hirings, we do not think that lack of calibration entails unfair bias, and that it is difficult to see why algorithmic contexts, as it were, should differ fairness-wise from non-algorithmic ones. Hence, we should reject the view that calibration is necessary for fairness in an algorithmic context and be open in principle to post hoc interventions counteracting differential false positive rates.

1. Introduction

It is widely acknowledged that certain groups of people are disadvantaged across a wide range of contexts as the result of unfair biases working to their disadvantage. Traditionally, it has often been assumed that the biases are known to the bearers. However, much recent research focuses on "implicit biases" involving automatic dispositions of which, sometimes, the agent is

[1] Kasper Lippert-Rasmussen, Professor of Political Science, Department of Political Science, Aarhus University. Head of The Centre for the Experimental-Philosophical Study of Discrimination (CEPDISC).

unaware. Indeed, some implicitly biased agents will strongly disavow the biases their behavior manifests when questioned about them.[2]

While most people, at least at some point of their lives, will belong to at least one group with biases working against it, some people belong to many such groups all their lives. Biases are stronger against some groups than others. Some are active across a wider range of contexts than others, and sometimes biases are mutually enforcing. Because biases often result in undesirable, e.g., because unjust, outcomes, it is generally agreed that sometimes, at least, we ought to intervene to mitigate or prevent their effects.[3] Such interventions can be ante hoc or post hoc. *Ante hoc* interventions concern biases themselves or their manifestation in behavior such as decision-making. The aim is to make the biases less common or to reduce their influence on behavior. *Post hoc* interventions take biases and their manifestation in behavior as parametric and aim to reduce the degree to which these result in undesirable outcomes.[4]

Post hoc interventions, which form a big family, differ in various ways. First, they differ in terms of the means adopted to avoid the relevant undesirable outcomes, e.g., quotas, or an adjustment of qualification scores to counteract evaluators' known biases. Second, they differ in terms of the sort of undesirable outcome they seek to avoid. Thus, in a series of recent articles, Martin Jönsson (2022), and Martin Jönsson and co-authors (2017, 2022), have focused on post hoc interventions seeking to reduce the inaccuracy of rankings produced by biased evaluators. By contrast, traditional affirmative action interventions such as quotas are sometimes intended to reduce unjust inequality of opportunity (Lippert-Rasmussen 2020, 72-102). The undesirable outcome I shall focus on here – differential false positive rates – is neither of these, but it is one that has received considerable attention in recent discussions of algorithmic fairness.

I begin, in Section 2, by describing the well-known controversy over COMPAS. There, critics have argued that black offenders are victims of

[2] For philosophically informed overviews of the implicit bias literature, see Beeghly and Madva (2020); Brownstein 2019; Brownstein and Saul (2016).

[3] To agree with this is not to say that we would have no reason to counteract such biases if they did not result in undesirable outcomes.

[4] Ante hoc and post hoc interventions can supplement each other, and perhaps in many cases the aim informing either intervention can be achieved only by adopting both. However, doubts about how successful an ante hoc intervention is introduce doubts about what the correct post hoc intervention is (but see Jönsson and Bergman 2022, 12, 21). Nothing in what I say below hangs on whether the two interventions supplement each other.

machine bias in that the recidivism risk prediction algorithms burden them with a higher rate of false positives (roughly: inaccurate predictions that an offender will reoffend) than white offenders face.[5] An obvious post hoc intervention in which judges are instructed or advised to draw different conclusions from a given risk score depending on whether the offender is black or white could, potentially, mitigate that problem. Yet, many think such an intervention, resulting in a deviation from the guidance provided by a well-calibrated risk assessment would be unfair to white offenders.[6] Section 3 briefly explores the implications of a commonly held view about unfair bias on the job market in light of audit studies and the conceptual apparatus introduced in Section 2 in relation to COMPAS. The section explains that in a job market where, because of past sexist discrimination, men are more likely to be qualified for certain jobs, deeming an applicant to be qualified means different things across male and female applicants, since there is greater chance of being qualified for the former. Many, this author included, would see no fairness-based reason in this situation for a post hoc intervention to secure a well-calibrated hiring process.[7] Thus, Section 3 ends with a trilemma consisting of three claims: 1) Lack of calibration does not amount to unfair bias in job markets; 2) Job markets and sentencing do not differ as regards whether a lack of calibration amounts to unfair bias; 3) Lack of calibration amounts to unfair bias in sentencing. Plainly, we must reject at least one of these claims, so the following sections (4-6) go through each of them in turn, asking which should be abandoned. Section 7 concludes.

[5] False positive rates are defined as: False Positives (FP)/Actual Negatives=FP/True Negatives (TN) + FP. False negative rates are: False Negatives (FN)/True Positives (TP) + FN. See also Table 1 below.

[6] As will become clearer shortly, the post hoc intervention in question here might not be one that presumes people are psychologically biased and then seeks to mitigate the degree to which that bias translates into differential outcomes for different groups (cp. Jönsson and Sjödahl 2017, 500). Rather, it may seek to mitigate the extent to which differential recidivism base rates, through what in the literature is referred to as machine bias, are turned into differential unjust outcomes by seemingly – so the criticism goes – unfair algorithms. This shows that there can be a rationale for exploring post hoc interventions even in the absence of implicit psychological biases that are difficult, very costly, or even impossible to eliminate. In short, the justification for exploring post hoc interventions is robust regarding the manipulability of implicit psychological biases.

[7] This point relates to a point made in Thore Husfeldt's article (this volume) that if we give everyone equal odds, we will not get demographic parity unless we have equal base rates across different groups. However, Husfeldt is agnostic on the implications of this for concerns about fairness.

In a nutshell, I argue, *first*, that we should, as it were, bring what we think of algorithmic fairness into line with what we think about job market discrimination in an ordinary non-algorithmic setting. That result I am quite confident of. How we should do it, i.e., how we should resolve the trilemma, I am less clear about. However, I offer some reasons suggesting, *second*, that in certain cases involving differential base rates – in principle, at least – we should allow post hoc interventions to equalize false positive/negative rates even if that means violating calibration. These are the two main claims in this article.

2. COMPAS and Calibration

I start, then, with a thumbnail sketch of the COMPAS debate. COMPAS, which stands for Correctional Offender Management Profiling for Alternative Sanctions, uses information about, among other things, an offender's employment and housing status, personality traits and criminal record to arrive at a risk of recidivism score – basically, a number from 1 (least likely) to 10 (most likely) indicating how likely it is the offenders will recidivate relative to other offenders – which is used by the courts in sentencing. It does not use information about race. Higher scores, indicating a greater likelihood that the offender will reoffend, will generally lead the courts to sentence offenders to longer periods of incarceration than they would be given with lower scores.[8] Hence, a false positive is a bad thing for an offender and a false negative is a good thing.[9]

In a renowned article entitled "Machine Bias" in *ProPublica*, Angwin et. al. (2016) suggested that COMPAS is unfair because it is racially biased. Like other ways of assessing the risk of recidivism, e.g., simply relying on the judge's impression of the offender and a statement from a psychiatrist, COMPAS is far from perfectly accurate.[10] In some cases, it predicts it to be

[8] Some might object to this sentencing practice on the grounds that it involves sentencing offenders on bases other than the crime committed. I set aside the issues raised by this complaint, noticing though that in most jurisdictions assessments of an offender's dangerousness can play a lawful role in sentencing. In any case, COMPAS is also used for other purposes than sentencing, e.g., decisions about bail.

[9] For a useful and insightful description and analysis of the case, see Hellman (2020).

[10] According to ProPublica, COMPAS was only "somewhat more accurate than a coin flip". Whether it is more accurate than standard assessments of risk of recidivism is an important question given that such assessments, in some form or another, play a role in determining the level of punishment.

highly likely that an offender will reoffend and in fact they do not (false positives).[11] In other cases, it deems it highly unlikely that the offender will reoffend and in fact they do (false negatives). What is striking is that even though, overall, COMPAS is equally accurate in making correct predictions across black and white offenders, its false positive and false negative rates differ across white and black offenders.[12] COMPAS is more likely to misclassify a non-recidivating black offender (44.9%) than a non-recidivating white (23.5%) offender as dangerous, and it is more likely to misclassify a recidivating white offender (47.7%) than a recidivating black (28.0%) offender as not being dangerous. This seems unfair, because it seems that sentencing based on COMPAS treats black offenders (upon whom it imposes a greater risk of an unduly long incarcerations) worse than white offenders (whom it privileges with a greater prospect of an unduly short period of incarceration).[13] At any rate, this was the intuitively forceful complaint set out in the "Machine Bias" paper.

In response to this criticism, Northpointe – the company that sells COMPAS to US courts – conceded the factual basis of Angwin et. al.'s criticism. However, it replied that COMPAS is well calibrated across black and white offenders. Essentially, in the case at hand this means that, for any given risk score, the probability that the offender will recidivate is the same whether the offender is black or white. Or, to put this in more general terms, which will be helpful later in Section 3: for each possible score, the (expected) percentage of individuals assigned this score who are positive is the same for each relevant group.[14] Calibration across groups, Northpointe submitted, is necessary and sufficient for algorithmic fairness.

[11] Strictly speaking, COMPAS' risk scores are ordinal, not cardinal. A high-risk score simply indicates that the offender belongs to a percentile of offenders who are more likely to reoffend than offenders from most other percentiles, not that the offender is very likely to recidivate (though, as a matter of fact, they do).

[12] In fact, Angwin et. al. used a finer-grained taxonomy of racialized groups, but for present purposes this makes no difference.

[13] What, exactly, (un)fair treatment amounts to is complex. Here I shall simply assume that differential treatment of the sort involved here is unfair. I return to these issues in Section 7.

[14] Or to put this requirement differently: TP/Predicted Positives=TP/FP + TP is the same across different relevant groups (compare footnote 6). There is a further requirement often labelled a requirement of calibration, i.e., that, for each group, the risk score is equal to the percentage of individuals who are assigned this risk score and reoffend. Since my focus here is on fairness to individuals across different groups, this aspect plays no role in my argument.

Several theorists have offered at least partial support for this response. For instance, Brian Hedden (2021, 227) writes: “none of the statistical criteria considered in the literature are necessary conditions for algorithmic fairness, except Calibration Within Groups”. Similarly, Robert Long (2020, 4, 17) submits that “when appropriate decision thresholds have been set, calibration is a necessary condition for procedural fairness … false positive [KLR: and negative] rate inequality is not, in itself, a measure of unfairness”.

One interesting point emerging from the burgeoning literature on algorithmic fairness of recent years is that, other than in special circumstances,[15] when two groups differ in terms of their base rates – as they do in the present case, since the frequency of recidivism is, as it happens, higher for black American offenders than it is for white American offenders – it is mathematically impossible for a predictive algorithm to be *both* well-calibrated across groups *and* have equal false negative and false positive rates across groups.[16] This insight has given rise to a substantial debate, involving computer scientists, philosophers and others, over the right criteria of algorithmic fairness.

Another important point is the following. In effect, if we accept the criticism levelled by Angwin and colleagues, we are committed to the view that there is at least a pro tanto reason in favor of a post hoc intervention to prevent the “machine bias” of COMPAS from resulting in unfair, unequal positive rates across white and black offenders. For mathematical reasons, such an intervention would involve giving up on calibration by adjusting the way COMPAS risk scores are assigned such that, for a given high risk score, it takes more predictors of recidivism for a black offender than for a white offender to be assigned this risk score (and the reverse for low risk scores), the result being that a higher proportion of black than white offenders who are assigned a low risk score will recidivate, i.e., the reverse situation of what was the case in 2016.[17] To explore whether such a post hoc intervention involving violating of

[15] For example, those where the predictive algorithm is perfect.

[16] For an excellent overview of the debate, and of various impossibility results, that is accessible to mathematically less sophisticated readers, see Hedden (2021; see also Eva 2022).

[17] As many contributors to the literature emphasize, the unequal base rate claim is problematic in various ways. What is known is the rate at which offenders are charged or convicted, not the rate at which they reoffend, and biases boosting charging or conviction rates in the case of black offenders might explain why those offenders face a higher risk of being convicted of further offenses in the future even if recidivism base rates are identical across white and black offenders. To the extent that such biases shape the base rates of black and white offenders, the relevant post hoc intervention would still qualify as a post hoc intervention, albeit arguably not one that

calibration is desirable in the present case in principle at least, I want to consider one that is similar but raised in the different and, it would seem, well-examined non-algorithmic context of *discrimination in hiring*.

3. Post Hoc Interventions in the Job Market

There is a well-established literature on bias in hiring. In this, so-called audit studies[18] present survey experiments in which one independent variable, such as race or gender, is altered to reveal the effect of doing that. For instance, the experimenters might send out a large number of job applications with accompanying CVs. These will be identical except for the applicant's name, which in half of the applications strongly suggests the applicant is a man and in the other half strongly suggests the applicant is a woman. If the subsequent call-back rates vary, with, say, male-looking applicants getting more calls than female-looking ones, then, other things being equal, the audit study will conclude that female applicants, in the sector being examined, are subjected to (unfair) bias. If there is no difference in call-back rates, it will conclude that there is no (unfair) gender bias in the call-back phase of hiring (which, of course, is not to say that there might be no unfair gender bias in later phases. Whether there is can also be studied through audit studies).[19] What I now want to consider is:

> *Job Market*: There are 500 male and 500 female applicants for a certain position. As a result of past sexist discrimination preventing female applicants from acquiring the much-needed work experience, 180 of the male applicants

counteracts machine bias as opposed to (explicit or implicit) psychological biases exhibited by people (e.g., police officers who are more inclined to charge black people than white people).

[18] For some prominent examples, see Neumark (1996), Banerjee et. al. (2009), Widner and Chicoine (2011), Gaddis (2014), Pager and Quillian (2005).

[19] Or, more precisely, the audit study will conclude that there is no (unfair) *direct* bias in hiring. An audit study does not speak to the question of whether the requirements of the job are unfairly, indirectly discriminatory. Note also that the two inferences in question are not as straightforward as one might think, because the information provided in identical texts with differently gendered names might be different. For instance, in a sexist society information about a 9-month parental leave period will be interpreted differently depending on whether the applicant is male or female and thus differential responses might be informed by factors other than the mere gender of the applicant (see Hu forthcoming).

> are qualified, while only 20 female applicants are.[20] The hiring procedure is such that an audit study will conclude that it makes no difference whether the applicant is male or female and, thus, that there is no unfair gender bias in the hiring procedure – all other things being equal, for any hired and any non-hired applicant exactly the same outcome would have occurred had this applicant had a different gender. Hiring is conducted in a non-algorithmic way: I shall say more on this later, but briefly, it means that the members of the hiring committee look at the applications using their judgment and informal deliberation to form an opinion about who is, and who is not, qualified. As the audit study informs us, the hiring committee is unbiased, gender-wise, in its assessments. Finally, the hiring committee's assessments are quite accurate, but not perfect. If an applicant, whether male or female, is qualified, there is a 90% chance the committee will deem them to be qualified. If the applicant is unqualified, there is a 90% chance the committee will deem them unqualified.

Job Market, as described, gives:

Table 1: Confusion table

	In fact: qualified	**In fact: not-qualified**	
Prediction: **qualified**	162 (men)/18 (women) True Positives (TP)	32/48 False Positives (FP)	194/66 (260)
Prediction: **not-qualified**	18/2 False Negatives (FN)	288/432 True Negatives (TN)	306/434 (740)
	180/20 (200)	320/480 (800)	500/500

Since my aim is to compare fairness judgments in ordinary hiring contexts with fairness judgments in relation to machine bias, let me describe this situation in the language of COMPAS. Basically, it is a situation where the assessment of the applicants is not well-calibrated despite the fact that an audit study will conclude that the procedure involves no unfair bias. That is, the ascribing of the values "qualified" and "not-qualified" to the applicants does not, as it were,

[20] The assumption that the difference in base rates reflects past unjust discrimination is not essential to my argument, but it has certain presentational advantages, one being that, for some readers, it might make such a difference (see the discussion of compounding injustice below).

have the same meaning across gender.[21] If the hiring committee finds that a particular applicant is qualified, that implies that there is a greater chance that the applicant is qualified if the applicant is male (162/194) than there is if she is female (18/66).[22] However, the hiring procedure will involve equal false-positive and false-negative rates across gender, reflecting the fact that if an applicant is (un)qualified, then in 90% of those cases the committee will deem the applicant to be (un)qualified. Take, first, false positive rates. In the case of male applicants, 32 men are falsely predicted to be qualified (False Positives) relative to 320 who are unqualified (Actual Negatives). In the case of female applicants, 48 are falsely predicted to be qualified (False Positives) relative to 480 who are unqualified (Actual Negatives). So, the false positive rate is 10% for both male and female applicants.[23] Now take false negative rates. In the case of male applicants, 18 are falsely predicted to be unqualified (False Negatives) relative to the 180 who are qualified (Actual Positives). In the case of female applicants, 2 are falsely predicted to be qualified (False Negatives) and 20 are in fact qualified (Actual Positives). The false negative rate is therefore again 10% for both male and female applicants.

In the light of COMPAS, the interesting feature of Job Market is this. According to standard audit studies, there is no unfair bias in the Job Market hiring process.[24] Yet the hiring procedure is miscalibrated and involves equal false positive and false negative rates. On the face of it, a post hoc intervention to reduce miscalibration – e.g., by hiring a greater proportion of the men deemed qualified than of the female applicants deemed qualified – would not be a way of counteracting unfair bias. The message seems to be that in ordinary non-algorithmic hiring contexts with different base rates across different groups of applicants we should not worry about lack of calibration as

[21] The sense of "meaning" used here, and which is commonly used in the algorithmic fairness literature, is different and much more practically oriented than that involved when philosophers discuss the meaning of a term. In that sense, the fact that the same criteria are used across men and women to determine whether an individual applicant is (un)qualified implies that "(un)qualified" means the same whether it qualifies a male or a female candidate, e.g., "(un)qualified" applied to men and women has the same sense (in Frege's sense).

[22] In short: TP/FP + TP is higher for male and female applicants.

[23] In short: FP/TN + FP is the same for male and female applicants.

[24] According to Brian Hedden (2021, 225-226): "<Lack of calibration> seems to amount to treating individuals differently in virtue of their differing group membership". In Job Market, lack of calibration amounts to exactly the opposite, i.e., to not treating applicants based on their differing group membership; indeed, achieving calibration requires doing just that.

explained, but we should, possibly, worry about unequal false positive/negative ratios as such.

Assuming these claims reflect a correct assessment of the case at hand, this suggests that Northpointe's defense of COMPAS is mistaken, and that fairness might require a post hoc intervention of the sort entertained above. That is, there is no algorithmic fairness objection to white offenders with a risk score equal to that of black offenders having a lower risk of reoffending, because calibration is not a necessary condition of algorithmic fairness.

In light of reflections like these, the following claims seem plausible:

> (1) Lack of calibration does not amount to unfair bias in job markets (the *Standard View*).
>
> (2) Job markets and sentencing do not differ as regards whether lack of calibration amounts to unfair (direct) discrimination (the *Equivalence Claim*).
>
> (3) Lack of calibration amounts to unfair (direct) discrimination in sentencing (the *Northpointe View*).

Admittedly, though I say the Equivalence Claim is plausible, I have so far said nothing to justify it. I will do so shortly. What we can see already, however, is that *if* we embrace it, we are obliged to abandon one of the other two claims: we must *either* stop assuming – as audit study encourages us to do, and as I think that many people do, in effect, unreflectively – that lack of calibration reflecting differential base rate qualifications does not render ordinary hiring procedures unfairly biased *or* reject the Northpointe View that lack of calibration in sentencing amounts to unfair bias. This obligation arises from the fact that claims (1)–(3) are trilemmatic: from any pair of them we can derive the negation of the third. So the wider question is: Which of the three claims should be dropped? With this question in mind, I will assess the three claims in turn over the next three sections.

4. Rejecting the Standard View

Should we reject the Standard View of unfair bias? A response to this question that I have heard on several occasions is that audit studies, at any rate, appear to present no obstacle to doing so. The thinking here is that audit studies usually include a *ceteris paribus* clause implying that information about, say, gender or race has no probative value. However, in Job Market information about gender does have such value, so the *ceteris paribus* clause would be unsatisfied in this case.

I have two thoughts about this response. First, we can simply stipulate that the employer in Job Market has no information about the relevant baseline differences, in their qualifications, between male and female applicants. This would mean that gender has no probative value, and that the *ceteris paribus* clause is satisfied. Yet our assessment of the case – no unfair bias – would remain, I submit, the same. Second, the fact that audit studies often apply an "other things being equal" clause favors the retention of the Standard View. The clause is meant to accommodate cases in which the employer believes that information about identity has probative value, not cases where such differences exist. Indeed, these clauses cover cases where, in fact, there are no base rate differences, in their qualifications, between, say, male and female applicants (same mean, same distribution etc.), but where the employers reasonably, but incorrectly, believe that such base rate differences obtain. In principle, once that is factored into an audit study, it might still conclude that there is no unfair discrimination despite lack of calibration (or, for that matter, lack of false positive rates).

What about the positive case for retaining the Standard View? One way to build that case is by pointing out that rejection of the view has implausible implications. Imagine that we tweak the hiring procedure in Job Market in favor of male applicants – e.g., applying the rubric "Give an extra five points for male gender" – so that in the case of equally qualified male and female applicants the male applicant is more likely to be deemed qualified. Even so, on the present view male applicants can have a complaint about unfair bias against them, because while the extra points mitigate miscalibration, they do not rule out the possibility that a male applicant deemed qualified is more likely to be qualified than a female applicant deemed qualified. However, it is quite unappealing to think that male applicants in these circumstances – circumstances, that is, involving a hiring procedure boosting their qualification score on grounds of their gender – can complain about unfair gender bias *against* them. If anything, intuitively, they benefit from unfair bias.

We might also ask: Who can have a fairness complaint about lack of calibration in Job Market?[25] Arguably, the answer to this question will depend

[25] I assume that only individuals have morally relevant complaints. This assumption is consistent with the view that individuals have complaints about how they are treated qua members of specific groups. It is also consistent with the view that, in a derivative sense, groups can have complaints, i.e., those deriving from the complaints of their members.

on what the alternative hiring procedure is. If the alternative is a procedure in which calibration is secured, then presumably those men who are presently deemed unqualified but would be deemed qualified with calibration might have a complaint.[26] How much moral weight this complaint would have will depend on how much weight we should attach to the fact that most of these men are not qualified. It may seem problematic to suppose that one is being subjected to unfair bias when one is not deemed qualified if, in fact, one is not qualified – especially, if one even enjoys a better chance of being deemed qualified than equally, or even better, qualified female applicants.[27] In any case, a complaint of this sort will have to be weighed against the complaint of those qualified women who, because of calibration, have a lower chance of being hired.[28] I recognize that these consideration are inconclusive, but in view of how we normally think of fairness – at least in the form of procedural justice – in cases of the kind I have been looking at, I fail to see that the complaints of the men in question are decisive.

5. Rejecting the Equivalence Claim

Are the COMPAS and Job Market cases different in that in the former the consideration of fairness gives us reason to be concerned about whether calibration is satisfied, whereas in the latter that same consideration gives us no reason to be concerned about lack of calibration? I take it the burden of proof here is on those who think the cases differ.[29] Hence, in defending the

[26] For simplicity, let us assume that who is deemed qualified does not change in surprising ways – e.g., a female applicant who is deemed unqualified with lack of calibration is deemed qualified in the presence of calibration.

[27] One option here is to reject the meritocratic view that fairness requires people to be hired on the basis of their qualifications. There is a real debate here (Lippert-Rasmussen 2020, 230-252). But for present purposes it is not especially interesting, because rejecting it would seem to undermine the case not only for equal false positives/negatives ratios but also (qualification-based) calibration.

[28] If only a subset of the applicants is deemed qualified, the male and female applicants who are deemed qualified and are so also have a complaint against calibration, since calibration will reduce their risk of not being hired as a result of the greater number of unqualified males being deemed qualified.

[29] Unlike jobs, the number of years of incarceration one is being sentenced to is not a positional good. Positional goods are special in the sense that if one gets the good, others are excluded from it and have a lower chance of enjoying a good of this kind. Plausibly, fairness

Equivalence Claim I shall merely rebut some suggestions as to why they are different. This will amount, I realize, to an inconclusive argument in favor of the equivalence claim. Still, if my sense of where the burden of proof lies is correct, we will be entitled for the time being to continue to affirm the Equivalence Claim.

One obvious difference between the two cases is that whereas in Job Market hiring decisions are not made algorithmically, in court cases relying on COMPAS the verdicts are partly so made.[30] It could be argued, then, that what is crucial is whether a decision is made algorithmically, or at least in an algorithmically assisted way, thereby introducing the risk of machine bias.

I do not think this suggestion works. Let us distinguish between algorithms in a narrow and in a broad sense. In a narrow sense, an algorithm involves a precise mathematical formula that is applied – either by software or in manual calculations – to a certain dataset. In a broad sense, an algorithm is a process or procedure that "extracts patterns from data" (Lee and Floridi 2021, 170). If in the present context "algorithm" is intended in the broad sense, both cases – COMPAS and Job Market – involve algorithmic decisions and thus there is no difference of the proposed kind between the two cases. Members of the hiring committee in Job Market are not self-consciously applying a mathematical formula to process the information they receive. However, they do apply a procedure involving the extraction of "patterns from data", and it may even be that unselfconsciously their brains are operating along the lines of articulable mathematical formulae.

In the narrow sense of "algorithm", things are different: the hiring case does not involve an algorithmic decision in this sense. However, the problem with appealing to this narrow notion is that, with it in place, it is unclear why it should make any difference, from the point of fairness, whether one makes an algorithmic decision or not. Suppose there are two different openings. The first is filled by the hiring committee. The second is filled using a computer running a particular algorithm to determine which applicants are qualified and which are not. Suppose the same applicants apply for the two positions, and that, for every applicant, the hiring committee and the algorithm reaches the same

considerations have greater weight when it comes to positional goods than when it comes to non-positional goods. Hence, if calibration is a fairness concern, one would expect calibration to be even more important in the hiring case and this means that there is a particularly heavy burden of proof on those who think we should only be concerned with calibration in the sentencing case.

[30] "Partly" because judges are free to disregard COMPAS's predictions.

verdict. It seems incredible to suppose that some applicants can complain about unfair bias in one of these cases, but not in the other. Where fairness is concerned, the machine bias is surely no worse than the hiring committee's "non-machine" bias.

A second suggestion is that punishment involves harming whereas hiring involves benefiting, and that this difference somehow explains why concern about fairness has rather different implications in the two cases. The simple response to this is to note that the good in the penal context could be described as the benefit of avoiding long incarceration, in which case the two cases would no longer differ in the respect appealed to. But even granting the harm/benefit asymmetry, I fail to see how it would justify different fairness-based concerns about calibration. Suppose the job in question turns out to be a bad job. The successful applicant would have been better off with a different job. (Or suppose the punishment turns out to be better than the alternative.) To my mind, these suppositions would not oblige us to revise what we think matters, from the point of view of fairness, in each of the two cases.[31]

Third, it might be suggested that the two cases differ because in Job Market people are (primarily, at least) assessed on the basis of individualized evidence freely offered by the applicant, whereas in the COMPAS scenario the merits of different offenders are assessed using non-individualized evidence that is not freely offered by the offender and was obtained from criminal registers available to the court, etc. Setting aside the question whether this allegedly factual difference is as stark as this suggestion would require in order to go through, I think the difference fails to do the necessary explanatory work.[32] Suppose, in a job-market, that applicants simply indicate an interest in their preferred position, and that the employer then assesses their qualifications by collecting information about the applicants using statistical data on various reference groups to which the applicants belong. Similarly, suppose that offenders can decide, voluntarily, to have their risk of recidivism assessed by

[31] It might be suggested instead that the two cases differ morally because the harm of unjustifiably long incarceration imposed by an uncalibrated risk prediction instrument are morally wrong, while the harm of not being hired imposed by an uncalibrated hiring procedure are not. However, whether correct or not this suggestion is unhelpful in the present context, which in effect involves searching for an, and not just begging the answer to the question of what makes harms in the COMPAS case morally wrong (because unfair) and does not make harms in the ordinary hiring case morally wrong (because unfair).

[32] Depending on what, exactly, is meant by individualized evidence, COMPAS does in part use individualized information, e.g., information about prior convictions. In part, it also uses information offered – though perhaps not freely so – by offenders.

COMPAS (rather than a psychiatrist), and that the algorithm is adjusted in such a way that it is fed only individualized information. My conjecture is that, again, this would not result in our caring about lack of calibration in Job Market, or our ceasing to care about calibration in COMPAS.

A final suggestion: the key difference between the COMPAS scenario and Job Market is that in the former it is the state that makes the decisions (through the courts), whereas in the latter the decisions are made by a private employer. This could be held to be significant for various reasons. Thus, it might be said that it makes a difference because, arguably, the state cannot say "Nothing to do with me" in response to the different recidivism base rates across black and white offenders. Arguably, this difference reflects, in part at least, unjust political policies. By contrast, a private company will often be able to disclaim responsibility for the fact that fewer women have the necessary job experience than men. Again, I do not think these differences are significant in the way that is being imagined. Suppose all black offenders in the US are recent immigrants whose criminal dispositions are the result of injustices in their country of origin. My guess is that people who care about calibration would still care about it across white and black offenders in this scenario. Also, in the hiring context it makes no difference whether the employer is the state or a private employer.

At this point I shall move on. I have not demonstrated that the Equivalence Claim is true, but I have, I hope, shown that we have good reason to be skeptical about several (and in my view, the most obvious) suggestions as to why it is best regarded as false.

6. Rejecting the Northpointe View

Perhaps in the light of the above we should reject the Northpointe View – and this is indeed what I propose to do now. I shall propose a somewhat roundabout argument for this option that starts from Long's no preference argument against equal false positive/negative rates being necessary for algorithmic fairness:

> (4) *No preference:* When there is group-wise inequality of false positive rate, a higher false positive rate does not give members of a group reason to prefer that they had belonged to a group with a lower false positive rate.

> (5) *No preference, no complaint:* If inequality of some metric Y does not give members of some group a reason to prefer that they belonged to another group, then members of this group do not have a procedural fairness complaint grounded in the inequality of metric Y.
>
> (6) *No complaint, no unfairness:* If no member of a group has a procedural fairness complaint grounded in the inequality of metric Y, then group-wise inequality of metric Y is not sufficient for procedural unfairness towards members of this group.
>
> (7) *Conclusion*: Group-wise inequality of false positive rate is not sufficient for group-wise procedural unfairness.

I think this argument is forceful. For argument's sake, let us grant (5) and (6) and focus on (4). In defense of this premise, Long offers an analysis of the following complaint from a black offender whose conviction was based in part on input from COMPAS:

> I am a black defendant who was not rearrested, but I was detained. False positive rate inequality shows that I was unfairly more at risk of this false classification than a non-rearrested white defendant. After all, a greater share of non-rearrested blacks are false positives. (Long 2020, 13)

According to Long, this complaint involves a fallacy. The complaint goes subtly wrong because it incorrectly links "'risk of error' to the false positive rate. While miscalibration or inappropriately differential thresholds *are* evidence of systematically unequal risk of error, false positive rate inequality is not" (Long 2020, 13). To see this, suppose that the black defendant in a COMPAS setting is white instead, and that all other things are equal.[33] Here

[33] Why is this the relevant counterfactual to consider, one might ask? This question is particularly relevant because, in the US context, race is causally tied to many of the other properties that are used as data input in COMPAS. In the closest possible world in which the black defendant is white, plausibly, the defendant would also have been better educated, lived in an area with lower crime-rates, had a better job situation, and so on. So why is the question to ask (for purposes of assessing premise (1)) not: Would the black offender have received a high-risk score if all those things, and not just the offender's race, had been different? I take it that at this point Long could plausibly respond that the no preference argument pertains to procedural fairness complaints – see (5) – and not, say, some broader notion of social justice. For the former and narrow purpose, i.e., Long's own purpose, the indicated narrow counterfactual is relevant (both in the case of COMPAS and audit studies). That, of course, is not to deny that, in a broader social justice assessment, other counterfactuals may (also) be relevant. ("Also" because on many views social justice in a broad sense would include procedural fairness.) I thank Jenny Magnusson for pressing me on this issue.

COMPAS would have generated the same prediction, and accordingly the defendant would have faced the very same risk of ending up being a false positive, since the same information would have been feed into the algorithm. Hence, *No preference* applies in this case.

Suppose we accept this argument. It seems we can then construct a similar argument against calibration. Consider the following complaint – one mirroring that of Long's black defendant in the COMPAS setting – from an unqualified male applicant over the female-friendly calibration of the hiring procedure in Job Market:

> I am an unqualified man, who was not deemed qualified. Unequal calibration shows that I was unfairly denied a greater chance of this false classification than a non-qualified female applicant. After all, a greater share of women deemed qualified are false positives.

This complaint against lack of calibration involves a misunderstanding analogous to the one involved in Long's black defendant's complaint. Suppose the unqualified man had instead been an unqualified woman. By stipulation, this person's prospect of being falsely deemed qualified would be the same as it is in the actual scenario where he is a man: 10%. Given this, we can replace (4) in Long's argument with a similar premise regarding calibration (4*) and tweak Long's argument so that it targets the view that lack of calibration is sufficient for unfairness:

> (4*) *No preference:* When there is base rate-based lack of calibration, the lack of calibration does not give (unqualified) members of a group reason to prefer that they had belonged to a group where the (expected) percentage of individuals assigned this score ("qualified") who are qualified is lower.
>
> (5) *No preference, no complaint:* If inequality of some metric Y does not give members of some group a reason to prefer that they belonged to another group, then members of this group do not have a procedural fairness complaint grounded in the inequality of metric Y.
>
> (6) *No complaint, no unfairness:* If no member of a group has a procedural fairness complaint grounded in the inequality of metric Y, then group-wise inequality of metric Y is not sufficient for procedural unfairness towards members of this group.
>
> (7*) *Conclusion*: When there is base rate-based lack of calibration, lack of calibration is not sufficient for group-wise procedural unfairness.

In the light of this, and given the strengths of the arguments I presented above in support of the two other horns of the trilemma, a possible lesson to draw is that we should replace the third horn in the trilemma – that is, (3) the Northpointe View – with:

> (3*) Lack of calibration amounts to unfair (direct) discrimination in a sentencing context unless it reflects differential base rate (the *Northpointe* View*).[34]

(1), (2), and (3*) do not form an inconsistent triad, and all three claims seem to be compatible with the arguments I have presented. Specifically, the assertion of (1), (2), and (3*) is compatible with the way in which (I have argued) Long's argument against the idea that unequal false positive rates are sufficient for unfair bias generalizes to calibration. Neither equal false positives, nor calibration, is necessary for fairness. Perhaps, on reflection, this is unsurprising on the assumption that fairness is about the chances facing each individual of harms and benefits and given that algorithmic parity requirements such as equal false positive rates and calibration are about group probabilities. Note, finally, that (1) and (3*) are also consistent with the notion that lack of calibration and differential positive rates are indicators of unfair bias. In a version of Job Market where, on average, male and female applicants are equally qualified, lack of calibration might strongly suggest a gender-biased assessment of the applicants' qualifications. Similarly, in a US court the setting

[34] If, alternatively, we insist that COMPAS and Job Market are different, we can replace the first horn of the trilemma with (1*) "Lack of calibration does not amount to unfair bias in a job market when it reflects differential base rates resulting from injustices against the group favored by calibration", and the third horn with (3**) "Differential false positive/negative ratios amount to unfair (direct) discrimination in sentencing unless they reflect differential base rates across the two groups resulting from injustices against the group favored by the differential false positive/negative ratios". The rationale for the latter view would be that COMPAS and Job Market are different, since in COMPAS the differential false positive/negative ratios favor a privileged group, whereas in Job Market the lack of calibration favors a group subjected to unfair treatment. One take on this is that in the former case calibration compounds injustice against women, whereas in the latter calibration compounds injustice against blacks – that is why the two cases differ. For reasons I have no space to explain here, I am skeptical about the idea that there is a non-derivative reason not to compound injustice, so I mention this possibility simply to flag it, not to signal my acceptance of it. I have, however, suggested an alternative way of capturing the intuition pertaining to compounding injustice that may be relevant here (Lippert-Rasmussen 2022). Note, finally, that if the form of fairness that we are concerned with here is procedural, it is less clear what the relevance of compounding injustice is, since procedural fairness can, on some occasions, stand in the way of social justice.

of white-offender friendly lack of calibration – something COMPAS avoids – might well, in part at least, be an indicator of a racially biased legal procedure.

7. Conclusion

In this article, I have shown why the Northpointe View of COMPAS introduces a way of thinking about unfair bias that diverges from the way we think about unfair bias in the job market, especially in the context of audit studies. This way of thinking, I have argued, lands us in a trilemma to which we should respond by rejecting the view that calibration is necessary for algorithmic unfairness. My arguments suggest that post hoc interventions to prevent bias in relation to false positives and false negatives might be commendable, fairness-wise, even if they clash with calibration across groups.

I should conclude by noting that this article does not argue that such post hoc interventions are justified. I am not arguing, for example, that judges in the US context should update risk assessments of white and black offenders in a way that generates miscalibration but equivalent false positive rates. Avoiding unfair bias – assuming for the moment that unequal false positive/negative rates manifest unfair machine bias – is one concern. But there are others, such as the concern to prevent crime and concern for political legitimacy, and nothing in this article has shown that post hoc interventions to eliminate differential false positive and false negative rates in the legal context are justified all things considered. However, given our views on job market discrimination, and given also the difficulty of explaining why the job-market and punishment contexts should be assessed differently, it is difficult to see how such interventions could fail to serve fairness well – in principle, at least.

Acknowledgements

Previous versions of this paper were presented at the European Workshop on Algorithmic Fairness (EWAF), June 9, 2022, University of Zürich, a workshop on statistical discrimination at University of Mainz, August 18, University of Mainz, online at the Philosophy Department at Wuhan University, September 26, 2022, the Post Hoc Interventions conference at Lund University, October 6, 2022; and at the Nordic Network for Political Theory workshop, November 3, 2022, UiT-The Arctic University of Norway. I thank the audiences – especially my assigned commentator on the Lund event, Jenny Magnusson, and Mattias Gunnemyr, Martin Jönsson, and Frej Klem Thomsen for helpful written comments and Lund University's Pufendorf Theme on post hoc interventions and the Danish National Research Foundation (DNRF144) for funding in relation to this article.

References

Angwin, Julia, Jeff Larson, Surya Mattu and Laure Kirchner (2016). Machine Bias. *ProPublica* May 26. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Banerjee, Abhijit & Marianne Bertrand, Saugato Datta, Sendhil Mullainathan (2009). Labor Market Discrimination in Delhi. *Journal of Comparative Politics, 37.1*, 14-27.

Beeghly, Erin & Alex Madva, (2020). *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*. New York: Routledge.

Brownstein, Michael (2019). Implicit bias. Stanford Encyclopedia of Philosophy: https://plato.stanford.edu/entries/implicit-bias/.

Brownstein, Michael & Jennifer Saul (eds.) (2016). *Implicit Bias & Philosophy vol. 1&2.* Oxford: Oxford University Press.

Eva, Benjamin (2022). Algorithmic Fairness and Base Rate Tracking. *Philosophy & Public Affairs, 50*(2), 239-266.

Gaddis, S. Michael (2015). Discrimination in the Credential Society. *Social Forces, 93.4*, 1451-1479.

Hedden, Brian (2021). On Statistical Criteria of Algorithmic Fairness. *Philosophy & Public Affairs, 49*(2), 209-231.

Hellman, Deborah (2020). Measuring Algorithmic Fairness. *Virginia Law Review, 106*(4), 811-866.

Hu, Lily (forthcoming). Interventionism in Theory and in Practice in the Social World. (On file with author).

Husfeldt, Thore (2023). Six Ways of Fairness. *This volume.*

Jönsson, Martin (2022). On the Prerequisites for Improving Prejudiced Ranking(s) with Individual and Post Hoc Interventions. *Erkenntnis*.

Jönsson, Martin, & Bergman, Jakob (2022). Improving Misrepresentations Amid Unwavering Misrepresenters. *Synthese*, *200*.

Jönsson, Martin and Sjödahl, Julia (2017). Increasing the veracity of implicitly biased rankings, *Episteme 14*(4), 499–517.

Lippert-Rasmussen, Kasper (2022). Is there a Duty not to Compound Injustice?. *Law and Philosophy*, online first: https://link.springer.com/article/10.1007/s10982-022-09460-y.

Lippert-Rasmussen, Kasper (2020). *Making Sense of Affirmative Action*. Oxford: Oxford University Press.

Long, Robert (2020). Fairness in Machine Learning. https://arxiv.org/pdf/2007.02890.pdf.

Neumark, David (1996). Sex Discrimination in Restaurant Hiring. *Quarterly Journal of Economics, 111.3*, 915-941.

Pager, Devah and Quillian, Lincoln (2005). Walking the Talk?. *American Sociological Review, 70.3*, 355-380.

Widner, Daniel and Chicoine, Stephen (2011). "It's All in the Name", *Sociological Forum, 26.4,* 806-822.