Law and Philosophy
https://doi.org/10.1007/s10982-024-09505-4

KASPER LIPPERT-RASMUSSEN

ALGORITHMIC AND NON-ALGORITHMIC FAIRNESS: SHOULD WE REVISE OUR VIEW OF THE LATTER GIVEN OUR VIEW OF THE FORMER?

(Accepted 19 April 2024)

ABSTRACT. In the 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 for white offenders. In response, some have argued that algorithmic fairness concerns, either also or only, 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–and that, for mathematical reasons, it is virtually impossible to equalize false positive rates without impairing the calibration. 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 in this respect. Hence, we should reject the view that calibration is necessary for fairness in an algorithmic context.

I. INTRODUCTION

In a US context, critics of courts' use of risk prediction algorithms such as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) have argued that black offenders are victims of machine bias. This is because recidivism risk prediction algorithms such as COMPAS burden black offenders with a higher rate of false positives (essentially: inaccurate predictions that an offender will reoffend) than white offenders face.[1] In response, some have argued that algorithmic fairness only concerns calibration across groups. Roughly, calibration across groups means that a score assigned to

[1] 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.

Published online: 24 October 2024

different individuals by the algorithm involves the same probability of the individual having the target property across different groups of individuals. By way of illustration: It is not as if offenders from one racial group assigned the risk score 8–i.e., a high risk–have the same probability of recidivating as offenders from another racial group assigned a risk score of only 6, especially not when a higher risk score translates into a harsher punishment. However, I argue that in standard non-algorithmic contexts, such as hirings, we do not think that lack of calibration entails unfair bias. Moreover, it is difficult to see why algorithmic contexts, as it were, should differ fairness-wise from non-algorithmic ones. Hence, despite appearances, we should reject the view that calibration is necessary for fairness in an algorithmic context.

I begin, in Section 2, by describing the well-known controversy over COMPAS. Section 3 briefly explores the implications both of a commonly held view about unfair bias on the job market considering audit studies and of 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. Specifically, for a given qualification score there is a greater chance of a male applicant being deemed qualified. 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. 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.

In a nutshell, I argue, *first*, that we should 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 is one I am quite confident of. I also think it is significant, since much discussion of algorithmic fairness fails to connect with discussions of fairness in other and more well-explored contexts. 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, we should allow for violating calibration.[2] This is not to say that, e.g., equal false positive/negative rates (henceforth: parity) is the correct criterion for algorithmic fairness. Perhaps neither calibration nor parity defines algorithmic fairness.

II. COMPAS AND CALIBRATION

I start, then, with a thumbnail sketch of the COMPAS debate. COMPAS uses information about an offender's employment and housing status, personality traits, criminal record, etc. to arrive at a risk of recidivism score. Basically, that score is a number from 1 (least likely) to 10 (most likely), indicating how likely it is that an offender will recidivate relative to other offenders. COMPAS does not use information about race. Presented with higher scores, the court will generally be less inclined to grant bail or parole, and more inclined to sentence an offender to longer periods of incarceration, than it would be if the scores were lower.[3] Hence, for the offender, a false positive is a bad thing and a false negative is a good thing.[4]

In a renowned article entitled "Machine Bias" in *ProPublica*, Angwin and co-authors suggested that COMPAS is unfair because it is racially biased.[5] Like other ways of assessing the risk of recidivism,

[2] Thus, in analogy with theorists who deny that differential false positive rates do not constitute algorithmic unfairness (e.g., Brian Hedden, "On Statistical Criteria of Algorithmic Fairness", *Philosophy & Public Affairs* 49 (2021): pp. 209–231; Robert Long, "Fairness in Machine Learning" (2020), https://arxiv.org/abs/2007.02890), I am not arguing that lack of calibration for reasons other than differential base rates might not amount to algorithmic unfairness. Hence, my argument is consistent with lack of calibration being a good indicator of algorithmic unfairness. For instance, in the context where, generally, male members of a racial minority group are stereotyped as dangerous, lack of calibration to the effect that male racial minority offenders are less likely to recidivate than racial majority offenders with the same risk score probably indicates that racial bias influences the risk assessment. In an algorithmic context, this might reflect a biased set of data. In a non-algorithmic, psychological-assessment-based risk assessment, this might reflect implicit biases against racial-minority men.

[3] 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, though noting that in most jurisdictions assessments of an offender's dangerousness can play a lawful role in sentencing.

[4] For a useful and insightful description and analysis of the case, see Deborah Hellman, "Measuring Algorithmic Fairness", *Virginia Law Review* 106 (2020): pp. 811–866.

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

e.g., simply relying on the judge's impression of the offender and a statement from a psychiatrist, COMPAS is far from perfectly accurate.[6] In some cases, it predicts it to be highly likely that an offender will reoffend and in fact they do not (false positives).[7] 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 while 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.[8] 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 to black offenders, because it seems that COMPAS imposes a greater risk of unduly long incarcerations etc. on them.[9] At any rate, this was the intuitively forceful complaint set out in the "Machine Bias" paper.[10]

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.[11] Essentially, in the case at hand this means that, for any given risk score, the probability that the offender

[6] 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.

[7] Strictly speaking, COMPAS's 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 also indicate that). Additionally, the risk score system is not binary (low vs. high risk) but numerical (1–10), and the requirement of equal false positive rates only applies to binary classification algorithms. However, by setting an arbitrary threshold somewhere between 1 and 10–say, at 5 (as, in fact, authors of the *Pro Publica* article did) – we construct a binary score system based on the numerical one.

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

[9] 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.

[10] Some might respond that if the alternative to using COMPAS is using even more racially biased (expert) human predictions, e.g., the judge's or a psychiatrist's impression of the defendant, COMPAS might be an improvement relative to its non-use, unfair discrimination-wise (cp. Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Cass R. Sunstein, 2019. "Discrimination in the Age of Algorithms", *Journal of Legal Analysis* 10 (2019): pp. 113–174). I will set aside this point, which, whatever its implications are, is irrelevant for present purposes.

[11] Aziz Z. Huq "Racial Equity in Algorithmic Criminal Justice", *Duke Law Journal* 68 (2019): pp. 1043–1134, p. 1048.

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 risk score, the percentage of individuals assigned this score who are positive is the same for each relevant group.[12] Calibration across groups, Northpointe submitted, is necessary and sufficient for algorithmic fairness.

Several theorists have offered at least partial support for this response. As Corbett-Davies et. al. point out: "The dominant fairness criterion in … [the case of "*risk scores*, like those produced by COMPAS"] is calibration."[13] Brian Hedden writes: "none of the statistical criteria considered in the literature are necessary conditions for algorithmic fairness, except Calibration Within Groups."[14] Similarly, Robert Long submits that "when appropriate decision thresholds have been set, calibration is a necessary condition for procedural fairness … false positive [author: and negative] rate inequality is not, in itself, a measure of unfairness."[15] It is also worth noting that in much of the algorithmic fairness literature, when base rate probabilities across groups differ and it is therefore virtually impossible to achieve both calibration and equal false positive/negative rates across groups, this is presented as a fairness dilemma or,

[12] Or to put this requirement differently: TP/Predicted Positives=TP/FP + TP is the same across different relevant groups (compare footnote 1). There is a further requirement, often labelled a requirement of calibration, that, for each group, the risk score is equal to the percentage of individuals who are assigned this risk score and reoffend; seeBenjamin Eva, "Algorithmic Fairness and Base Rate Tracking", *Philosophy & Public Affairs* 50 (2022): pp. 239–266, pp. 247–248) on strong calibration. Since my focus here is on fairness to individuals across different groups, this aspect plays no role in my argument.

[13] Samuel Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq, "Algorithmic Decision Making and the Cost of Fairness" (2017), p. 3 https://arxiv.org/abs/1701.08230. Samuel Corbett-Davies and Sharad Goel ("The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning" (2018), http://arxiv.org/abs/1808.00023: pp. 1–117, pp. 1–2) point to calibration as one of three formal criteria of fairness that has gained "prominence… over the last several years." The other two prominent criteria are anti-classification, which was not violated in COMPAS regarding race, and classification parity, which they illustrate with the condition of "false positive and false negative rates." Like me, Corbett-Davies and Goel express doubts about calibration. Unlike my doubts, however, theirs concern whether calibration is a sufficient, not a necessary, condition of fairness: "… it is often straightforward to satisfy calibration while strategically misclassifying individuals in order to discriminate" (Corbett-Davies and Goel "Measure and Mismeasure", p. 2, p. 16; see also Fabian Beigang, "Reconciling Algorithmic Fairness Criteria", *Philosophy & Public Affairs* 51 (2023): pp. 166–190, pp. 174–175; Michele Loi and Christoph Heitz, "Is calibration a fairness requirement?" ACM Conference on Fairness, Accountability, and Transparency (FAccT'22): pp. 2026–2032, p. 2026: https://doi.org/10.1145/3411764.3445195). Sandra G. Mayson, "Bias In, Bias Out", *The Yale Law Journal* 128 (2019): pp. 2218–2300, p. 2294) thinks that "it is a corollary of the very concept of statistical prediction that the relationship between a risk score and risk itself be constant across racial groups" (Mayson "Bias", p. 2294; see Beigang, "Reconciling", for a discussion of this claim).

[14] Hedden, "On Statistical Criteria", p. 227.

[15] Long, "Fairness", p. 4, p. 17.

as Berk et al. put it, as something that presents decisionmakers with ''challenging tradeoffs'' (see next paragraph).[16] Such formulations strongly suggest that these theorists think lack of calibration is a source of unfairness, whether or not they also think it is the only, or the most important, fairness condition.[17]

One interesting point emerging from the burgeoning literature on algorithmic fairness in recent years is that, other than in special circumstances,[18] when two groups differ in terms of their base rates – as they do in the COMPAS 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.[19] 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 that 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 adjusting risk scores to prevent the ''machine bias'' of COMPAS from resulting in allegedly unfair, unequal positive rates across white and black offenders. For

[16] Richard Berk et al., ''Fairness in Criminal Justice Risk Assessments: The State of the Art'', *Soc. Methods & Research* OnlineFirst 1, 23 (2018): pp. 1–42, https://journals.sagepub.com/doi/-10.1177/0049124118782533

[17] Admittedly, not everyone takes this impossibility result to generate a dilemma. Huq, ''Racial Equity'', p. 1055, p. 1111, thinks that fairness requires avoiding imposing a net burden on communities of color. Hellman, ''Measuring'', argues that calibration is relevant to what we should believe, but not to fairness in the treatment of people, which is what, in her view, equality of false positives/negatives (and their impacts) bears on. Others focus on extra-algorithmic fairness problems. Thus, Mayson, ''Bias'', focuses on the idea that how the data algorithms are fed reflects background unfairness such as racially biased tendencies to arrest or differential exposure to criminogenic factors, and on how these features of social life should be addressed. Most recently, Beigang, ''Reconciling'', proposes modifying both the criteria of calibration and equal false negative/positive rates such that both criteria retain ''their intuitive appeal'' and become ''universally compatible.''

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

[19] For an excellent overview of the debate, and of various impossibility results, that is accessible to mathematically less sophisticated readers, see Hedden, ''On Statistical Criteria''; see also Eva, ''Algorithmic Fairness''). The mathematically more sophisticated might consult (Berk et. al. ''Fairness'', pp. 17–25; Alexandra Chouldechova, ''Fair Prediction with Disparate Impact'', *Big Data* 5 (2017): pp. 153–163; Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, ''Fairness in Machine Learning'' (2016), https://arxiv.org/abs/1609.05807;Jon Kleinberg et al., ''Inherent Trade-Offs in the Fair Determination of Risk Scores'', 67 LIPIcs 43 (2017): pp. 1–23, https://drops.dagstuhl.de/opus/volltexte/2017/8156/pdf/LIPIcs-ITCS-2017-43.pdf; Thomas Miconi, ''The Impossibility of 'Fairness''' (2017), https://arxiv.org/abs/1707.01195). Calibration and equal false positive/negative rates are far from the only algorithmic fairness criteria that have been proposed in the literature (see, e.g., Mayson, ''Bias'', esp. pp. 2238–2249; Berk et. al., ''Fairness'', pp. 12–15). However, they are the two criteria that, in the words of Hellman, ''Measuring'', p. 811, ''stand out.''

mathematical reasons, such an intervention would involve giving up on calibration. It would require adjusting the way COMPAS assigns risk scores such that, to be assigned a given high risk score, more predictors of recidivism would be required for a black offender than for a white offender—the result being that a higher proportion of black than white offenders who are assigned a low risk score will recidivate.[20] To explore whether such an adjustment involving a violation of calibration is desirable in the present case in principle at least, I want to consider deviation from calibration in a non-algorithmic context of *discrimination in hiring*.[21]

III. POST HOC INTERVENTIONS IN THE JOB MARKET

There is a well-established literature on bias in hiring. In this, so-called audit studies[22] present survey experiments in which one independent variable, such as race or gender, is altered in order to reveal the effect of so doing. 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, say, applicants with male-sounding names get more calls than those with female-sounding names, then, other things being equal, the audit study will conclude that female applicants, in the sector being

[20] 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. 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 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).

[21] By "desirable in principle at least," I mean that it is not the case that such an adjustment together with principles of algorithmic fairness entails that such an adjustment is undesirable algorithmic fairness-wise. Whether it is desirable all things considered is a different matter, which hangs on, inter alia, equal false positive rates as a requirement of algorithmic fairness as well as concerns other than algorithmic fairness, e.g., the desire to prevent crime and the concern for political legitimacy of the legal system and much else.

[22] For some prominent examples, see David Neumark, "Sex Discrimination in Restaurant Hiring", *Quarterly Journal of Economics* 111 (1996): pp. 915–941; Abhijit Banerjee, Marianne Bertrand, Saugato Datta, and Sendhil Mullainathan, "Labor Market Discrimination in Delhi", *Journal of Comparative Politics* 37 (2009): pp. 14–27; Daniel Widner and Stephen Chicoine, "It's All in the Name", *Sociological Forum* 26 (2011): pp. 806–822; S. Michael Gaddis, "Discrimination in the Credential Society", *Social Forces* 93 (2015): pp. 1451–1479; Devah Pager and Lincoln Quillian, "Walking the Talk?", *American Sociological Review* 70 (2005): pp. 355–380.

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. (Whether there is unfair gender bias in later phases of recruitment can also be studied through audit studies.)[23]

What sort of fairness notion underpins this pattern of inference? One powerful line of thought is that male and female applicants have a claim that their chances of success not be influenced by their gender (and, more generally, that applicants have a claim that their prospects do not depend on their social identities). Of course, applicants do not have a claim to be called in for an interview, let alone to be hired. After all, other applicants might be better qualified. However, they have a claim that their prospects of being invited for an interview etc. depend on their qualifications, and their qualifications alone. Indeed, an employer treating applicants fairly in this context is for the employer to respect this right to qualification-based-only prospects and respect it equally across applicants.[24] If the employer does so, the employer treats applicants who are alike in the relevant respect–qualifications–equally and applicants who are not alike in the relevant respect differently, i.e., the better-qualified

[23] 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 (see my discussion of the unfair disparate impact observation in Section 4). 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 nine-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 Lily Hu, "Interventionism in Theory and in Practice in the Social World" (forthcoming(a), on file with author); Lily Hu, "What is 'Race' in Algorithmic Discrimination on the Basis of Race?" *Journal of Moral Philosophy* (forthcoming (b)): pp. 1–23, esp. pp. 8–18; Lily Hu and Issa Kohler-Haussman, "What's Sex Got to Do with Fair Machine Learning?" (2020), https://arxiv.org/abs/2006.01770. https://doi.org/10.48550/arXiv.2006.01770: 1–11; Issa Kohler-Haussman, "Eddie Murphy and the Dangers of Counterfactual Causal Thinking About Detecting Racial Discrimination", *Northwestern University Law Review* 113 (2019): pp. 1163–1227, p. 1216; for a reply, see Naftali Weinberger, "Signal Manipulation and the Causal Analysis of Racial Discrimination", *Ergo* 9 (2023): pp. 1264–1287, p. 1273–1275, pp. 1280–1283. https://doi.org/10.3998/ergo.2915). Accordingly, Hu and Kohler-Hausmann argue that, on a constructivist understanding of race, audit studies fail to do what they purport to do, i.e., fail to identify racial discrimination understood as the causal effect of race. While I think their criticisms of audit studies are forceful, I am not sure that, even in principle, audit studies cannot be designed to soften or even accommodate them. In any case, constructivist challenges to audit studies can be generalized to assessments of whether algorithms unfairly discriminate in such a way that this challenge raises a problem that is somewhat different from the one I focus on in this article.

[24] Cases where there is a gap between what the employer is epistemically justified in believing the applicant's qualifications to be and what the applicant's qualifications in fact are bring out the question of whether the right in question is a fact-relative or an evidence/belief-relative right. I shall proceed as if the right in question is an evidence-relative right.

applicant is preferred over the less well-qualified applicant.[25] Thus, if an audit study shows that gender statistically influences call-back rates, then that suggests that the employer is not treating applicants fairly in this respect–the employer is not treating like applicants alike – whereas the opposite is the case if gender has no statistical influence on call-back rates.

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 are qualified, while only 20 female applicants are qualified.[26] 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, this means that the members of the hiring committee look at the applications and use their judgment and informal deliberation to form an opinion abou 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 to be unqualified.

Job Market, as described, gives:

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 even though 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, have the same meaning across gender.[27] If the hiring committee finds that a particular applicant is qualified, that implies that there is a greater

[25] Interestingly, Kate Vredenburgh, ''Fairness'', in Justin B. Bullock (ed.) *Oxford Handbook of AI Governance*, Oxford: Oxford University Press (2020): pp. 129–148, suggests a similar rationale for calibration rooted in formal equality. In effect, I am suggesting in light of audit studies and the case described below that this cannot be quite right.

[26] Cp. Kleinberg et al., ''Discrimination'', p. 145. 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, however, the discussion of compounding injustice below).

[27] The sense of ''meaning'' used here, and which is commonly used in the algorithmic fairness literature, is 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).

chance that the applicant is qualified if the applicant is male (162/194) than there is if she is female (18/66).[28] However, the hiring procedure will involve equal false-positive and false-negative rates across gender. This reflects 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). Thus, the false positive rate is 10% for both male and female applicants.[29] 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. While I have stipulated that the hiring process is one in which gender has no causal influence on whether applicants are hired, equal false positive and equal false negative rates across gender are what one would expect if gender has no causal influence on hiring decisions and if the hiring process is equally good at determining applicants' level of qualification whatever their skill levels. The latter condition reflects the fact that if the assessment procedure were, say, better at determining the skill levels of highly skilled applicants, one would expect the false positive rate for women to be higher than that for men on the assumption that a higher proportion of male applicants than female applicants are highly skilled applicants.

In 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.[30] Yet the hiring procedure is miscalibrated and involves equal false positive and false negative rates. On the face

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

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

[30] According to Hedden, "On Statistical Criteria", pp. 225–226; cf. Vredenburgh, "Fairness"): "<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.

of it, adjusting the assignment of the scores ''qualified''/''unqualified'' to reduce miscalibration 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 so explained.

It might be objected that I am only able to pump this intuition because Job Market is a case where miscalibration works to the advantage of an already disadvantaged group that has been subjected to past injustice. If instead miscalibration worked to the ''advantage'' of an already privileged group, we would care about miscalibration.[31] To examine this challenge, imagine the following variant of Job Market:

> *Reversed Job Market*: This case is exactly like Job Market except for the following. As a result of past sexism, women had to be more qualified than male applicants to get hired. While the hiring process is no longer gender biased, female applicants are for that reason–following their compensatory extra efforts to acquire qualifications–much more qualified than male applicants. In fact, out of the 500 female applicants 180 are qualified, whereas only 20 out of the 500 male applicants are. Hence, if the hiring committee deems a female applicant to be qualified, she is much more likely to be qualified than if it deems a male applicant to qualified.

In this case, because the employer (again) is unaware of base rate differences across women and men, and because an audit study would conclude that the hiring process involves no bias, miscalibration would work to the ''advantage'' of men. Plausibly, in Reversed Job Market we might think that we should increase the proportion of female applicants hired and, thus, reduce miscalibration. However, the disparate impact-related reasons that might justify hiring more women are not reasons to care about calibration *as such*. Rather, they might be reasons to care, meritocracy-style, about, say, rewarding the efforts of the many women who have invested more than men in becoming highly qualified for the jobs for which they apply and to ensure that a higher proportion of those hired are qualified. Suppose that, for some injustice-unrelated reason, female applicants are much more qualified than male applicants–e.g., it is simply a statistical fluke that, in this case, female applicants are on average much more qualified than their male competitors. This being so, the disparate impact-related concerns about counteracting the effects of past sexism would no longer apply and, thus, would

[31] I thank an anonymous reviewer for this challenge. I put ''advantage'' in scare quotes for reasons that become apparent in my discussion of Robert Long's no preference argument in Section 6.

not speak in favor of reducing miscalibration. By contrast, the meritocratic concern of ensuring that a higher proportion of recruited applicants are best qualified still applies. Hence, if we believe the case for reducing miscalibration is no weaker in this modified version of Reversed Job Market than it is in Reversed Job Market, disparate impact-related concerns are not the driving concerns.

Second, even if we did care about calibration as such in Reversed Job Market, this still would not defeat my claim that calibration as such is not a fairness requirement. After all, the present challenge concedes that it is not a requirement in Job Market, where miscalibration works to the ''advantage'' of an unjustly disadvantaged group.

Finally, it might be felt that, unlike lack of calibration in Job Market, lack of calibration in Reversed Job Market is objectionable for expressive reasons, e.g., it is demeaning to women because of its signifying past sexism and society's present failure to come to terms with this history.[32] I accept the force of this complaint. However, one can accept it consistently with my main claim to the extent that complaints about how a certain practice is demeaning are a different complaint than complaints about unfairness. Arguably, a practice can demean members of a certain group without treating any member unfairly, in a comparative sense at least (out-group members are also being treated in a demeaning way), and a practice can be unfair but not demeaning in Hellman's sense, e.g., because the unfair practice is completely unacknowledged and, thus, has no objective cultural meaning.

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 is consistent with at least certain deviations from calibration across groups. Specifically, there is no algorithmic fairness objection in principle 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.

[32] Deborah Hellman, *When Is Discrimination Wrong?* (Cambridge, MA: Harvard University Press, 2008).

In light of reflections like these, the following claims might 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 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. This trilemma arises even if one thinks that calibration is not the only, or the dominant, algorithmic fairness condition. It suffices that one shares the common view that calibration is a fairness condition such that, if it is not met, there is less than perfect algorithmic fairness (recall Section 2). The wider question is therefore: 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.

IV. 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 from the point of view of the employer. What this means is that information about the applicant's gender provides the em-

ployer with no evidence about the applicant's level of qualifications.[33] 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 qualifications between male and female applicants. This would mean that gender has no known probative value from the point of view of the employer, and that the *ceteris paribus* clause is satisfied.[34] Yet our assessment of the case–no unfair bias–would remain, I submit, the same.

Some will say that the mere fact that an employer is not unfairly biased does not prevent them from treating applicants unfairly. Most of those who write about this issue are willing to allow that employers who engage in disparate impact discrimination need not be biased against discriminatees and yet can, even so, treat them unfairly. So why, it might be argued, cannot Job Market similarly involve unfair treatment because base rate differences between male and female applicants are not considered? I accept the premise of this challenge–that unfair treatment is possible in the absence of bias. But in my view the concerns underpinning the accusation of unfair disparate impact do *not* imply that we should consider lack of calibration in Job Market unfair. After all, those concerns motivate a desire to mitigate the negative effects of past sexist discrimination on women. However, given the stated assumptions, an attempt to reduce miscalibration would boost the number of men deemed to be qualified while increasing the number of women deemed to be unqualified, thus further boosting the amount of disadvantage imposed on women. In short, perhaps Job Market does involve unfair treatment despite a lack of gender bias. But, if it does so, the ''female-friendly'' miscalibration is not what constitutes this unfairness.[35]

[33] In Job Market, the difference in base rates across men and women is so large that it would be unrealistic to assume that employers would remain ignorant of it over time. However, we can disregard this fact for present purposes. In any case, employers are typically legally prohibited from taking the probative value of gender and race into consideration (Kleinberg et. al., ''Discrimination'', pp. 122–124; cf. Corbett-Davies and Goel, ''Measure and Mismeasure'', on race discrimination.

[34] This is consistent with the fact that information about gender has probative value for an outsider observer who is aware of base rate differences. However, this is irrelevant to an assessment of whether *the employer* is unfairly biased.

[35] I thank Sander Beckers and an anonymous reviewer for pressing this challenge.

My second response to the ceteris paribus clause observation is this: the fact that audit studies often apply this clause favors 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 there are no base rate differences in 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.

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, which involve 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?[36] Arguably, the answer to this question will depend on what the alternative hiring procedure is. If the alternative is a procedure in which calibration is secured, then those men who are presently deemed unqualified but would be deemed qualified with calibration might have a complaint.[37] How much

[36] Only individuals can 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.

[37] 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.

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. This is so especially if one even enjoys a better chance of being deemed qualified than equally, or even better, qualified female applicants. Moreover, under the present alternative procedure being male actually causes one to enjoy a greater chance of being deemed qualified![38] In any case, a complaint of this sort will have to be weighed against the complaint of those qualified women who, because of calibration and on account of their gender, will then have a lower chance of being hired.[39]

The nature of the complaint from men who are presently deemed unqualified but would be deemed qualified under a different procedure satisfying calibration is also unclear. In a related context, Deborah Hellman distinguishes between a complaint based on a (putative) non-comparative claim "to be treated by the most accurate test available," on the one hand, and a complaint based on a comparative claim that members of one's group are treated no worse than members of other groups, on the other.[40] She thinks that the first claim is illusory, though sometimes invoked.[41] If she is right about this and if the impulse animating any comparative fairness-based complaint on behalf of male applicants is a concern for male applicants who are presently deemed unqualified but would be deemed qualified under an alternative, more accurate procedure satisfying calibration, then we should disregard this complaint. We should do so either because there is no non-comparative claim "to be treated by the most accurate test available," or because, although there is such a claim, given its non-comparative nature it is not a claim about unfairness between groups.[42] Or we should do so be-

[38] One option here is to reject the meritocratic view that fairness requires people to be hired based on their qualifications. There is a real debate here (Kasper Lippert-Rasmussen, *Making Sense of Affirmative Action* (New York: Oxford University Press, 2020): pp. 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 positive/negative ratios but also (qualification-based) calibration.

[39] 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 increase their risk of not being hired because of the greater number of unqualified males being deemed qualified.

[40] Hellman, "Measuring", p. 833.

[41] Hellman, *When is Discrimination?* chs. 4–5.

[42] Hellman, "Measuring", p. 833.

cause miscalibration as such could not ground a comparative fairness-based complaint–it is only how the miscalibration ''operates'' that can do that.[43] While reducing miscalibration by deeming more men qualified would benefit unqualified men (who would then be deemed qualified), it would also harm qualified men who lose the ''ability to distinguish'' themselves from the now misclassified unqualified male applicants.[44] For this reason, this complaint is hard to construe as a complaint about being unfairly treated qua male applicant. However, surely, the former category is the one that is relevant for the purpose of fairness assessments.

I recognize that these considerations are inconclusive, but in view of how we normally think of fairness in cases of the kind I have been looking at, I fail to see that the complaints of the men in question have any force at all.

V. 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.[45] While one might think that it is morally more urgent that the state treats citizens fairly when it puts some of them in prison than it is that private employers treat applicants fairly, e.g., in light of what is at stake for the people involved in the two cases at hand, it is not immediately clear why the conditions of fairness would differ here. Similarly, if our shared lottery ticket wins $1 million it is more urgent that we divide the prize fairly than it would be if we were to win $10, but presumably there would be no difference in what constitutes a fair division in the two cases.[46] Ac-

[43] Hellman, ''Measuring'', p. 833.

[44] Hellman, ''Measuring'', p. 833.

[45] 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 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, typically 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.

[46] I thank Deborah Hellman for this way of putting the point.

counts of what makes treatment fair do not refer to the site of the relevant treatment. According to John Broome, for instance, fairness requires that people's equally strong claims are (un)satisfied to an equal degree.[47] On that account, fairness both in sentencing and job markets means the same. Hence, in defending the Equivalence Claim I shall merely rebut some suggestions as to why they are different.

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.[48] 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 to a certain dataset, e.g., using computer software. In a broad sense, an algorithm is a process or procedure that ''extracts patterns from data.''[49] If in the present context ''algorithm'' is intended in the broad sense, both COMPAS and Job Market involve algorithmic decisions and thus there is no difference of the proposed kind between the two cases, which means the present suggestion is a non-starter. Members of the hiring committee do apply a procedure involving the extraction of ''patterns from data.''

In the narrow sense of ''algorithm,'' the hiring case does not involve an algorithmic decision. However, it is unclear why it should make any difference, from the point of view of fairness, whether one makes an algorithmic decision in this narrow sense 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 reach the same verdict. It seems incredible to suppose that some applicants can complain about unfair bias in one of these cases but

[47] John Broome, ''Fairness'', *Proceedings of the Aristotelian Society* 91 (1991): pp. 87–102.

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

[49] Michelle Seng Ah Lee and Luciano Floridi, ''Algorithmic fairness in mortgage lending: From absolute conditions to relational trade-offs'', *Minds and Machines* 31 (2021): pp. 165–191.

not in the other. Where fairness is concerned, the machine bias is surely no worse than the hiring committee's "non-machine" bias.

Some might resist this analogy, submitting that whereas human decision-makers can be unfairly biased and therefore disrespectful, computers running an algorithm are constitutionally incapable of being disrespectful.[50] Only agents sensitive to moral reasons can disrespect. However, even if that is so, it does not help us to explain the difference between COMPAS and Job Market. By stipulation, the decision-makers in Job Market have no gender bias and, thus, are not treating any applicant disrespectfully on account of their gender.

A second suggestion is that punishment involves harming, whereas hiring involves benefiting, and that this difference somehow explains why a concern about fairness has rather different implications in the two cases. However, the good in the penal context could be described as the indirect benefit of avoiding a harm, e.g., a longer incarceration. If so, the two cases would no longer differ in the respect appealed to.[51] Moreover, it is generally assumed that fairness norms regulate the distribution of both harms and benefits.

Note, finally, that the Equivalence Claim merely states that job markets and sentencing do not differ regarding whether lack of calibration amounts to unfair (direct) discrimination. Hence, to defeat it, it is not enough to show that job market and sentencing contexts differ in non-algorithmic fairness-related contexts. For instance, it might be thought that harms to others as a result of miscalibration are greater in sentencing contexts than, say, job market contexts. Thus, a person's being incarcerated for a longer period of time because of a miscalibrated risk score does more harm to their spouse and child than the harm done to a spouse and child when an applicant is not given a job as the result of a miscalibrated qualification score prediction algorithm. It might be suggested that this justifies caring differentially about misclassification across the two contexts. However, although this is undoubtedly a morally significant concern, it is not a concern about the unfairness to those scored by the risk/qualification predictions in question.[52] I conclude that we

[50] I thank Lennart Ackermans for this challenge.

[51] We could also imagine a harm-focused version of Job Market where what is at stake is the question "Whom should we fire?"

[52] Hellman, "Measuring", pp. 834–835.

have good reason to be skeptical about rejection of the Equivalence Claim.

VI. REJECTING THE NORTHPOINTE VIEW

Perhaps in 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 rates is not sufficient for group-wise procedural unfairness.

This argument is forceful and shows that unequal false positives do not entail the existence of a fairness-based complaint. 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.[53]

According to Long, this 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

[53] Long, "Fairness", p. 13.

not.''[54] To see this, suppose that the black defendant in a COMPAS setting is white instead, and that all other things are equal.[55] Here 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 fed 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:

---

[54] Long, ''Fairness'', p. 13.

[55] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva, ''Counterfactual fairness'', *Proceedings of the 31st International Conference on Neural Information Processing Systems* (2017): pp. 3–5, https://arxiv.org/abs/1703.06856v3: 4069–4079; Hamed Nilforoshan, Johann Gaebler, Ravi Shroff, Sharad Goel, ''Causal Conceptions of Fairness and their Consequences'', *Proceedings of the 39th International Conference on Machine Learning* (2022): p. 3, https://arxiv.org/abs/2207.05302: 1–40; cp. Ludwig Bothmann, Kristina Peters and Bernt Bischl, ''What Is Fairness? Philosophical Considerations and Implications for FairML'' (2022): p. 13, https://arxiv.org/abs/2205.09622. Why is this the relevant counterfactual to consider? 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 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.) See (Hu and Kohler-Hausmann, ''What Has Sex Got to Do''; Kohler-Hausmann, ''Eddie Murphy''; Alexandre Marcellesi, ''Is Race a Cause?'', *Philosophy of Science* 80 (2013): pp. 650–659; Weinberger, ''Signal Manipulation'').

(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 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 rates (the *Northpointe* View*).[56]

(1), (2), and (3*) do not form an inconsistent triad. Moreover, 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 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, upon 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

[56] 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. 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 (Kasper Lippert-Rasmussen, ''Is there a duty not to compound injustice?'', *Law and Philosophy* 42 (2023): pp. 93–113). 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.

Table 1. Confusion table

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

group probabilities.[57] Note, finally, that (1) and (3⋆) are also consistent with the notion that, under certain circumstances, lack of calibration and differential positive rates are indicators of, or even amount to, unfair bias.[58] In a modified version of Job Market where, on average, male and female applicants are equally qualified and, thus, lack of calibration does not reflect differential base rates across gender, lack of calibration might amount to defeasible evidence that the assessment of the applicants' qualifications is gender-biased. In this modified case, gender would be a causal factor affecting hiring decisions and, presumably, an audit study would not conclude lack of bias. Similarly, in a US court the setting of white-offender-friendly lack of calibration might well provide prima facie evidence of a racially biased legal procedure–e.g., it might be evidence of implicit racial bias among judges, leading them to assess, e.g., black offenders to be more prone to recidivism than similar white offenders.[59]

VII. 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

[57] Similarly, from the point of view of a causal conception of fairness, this result looks plausible since the applicant's prospects would have been no different had they had a different gender. I should add that I did not rely on these claims–the individual vs. group probabilities observation, nor the causal fairness one–as premises in the arguments I presented in Sections 4 and 6. The claims are merely suggestions about how to understand conclusions I have reached on different grounds. They are speculations loosely motivated by and distinct from the main critique of the paper, not a summary of that critique. I thank two anonymous reviewers for pointing out the need to clarify this.

[58] Loi and Heitz, ''Is Calibration?'', p. 2030.

[59] See, for instance, Jeffrey J. Rachlinski, Sheri L. Johnson, Andrew J. Wistrich, and Chris Guthrie, Chris, ''Does unconscious racial bias affect trial judges?'' *Notre Dame Law Review* 84 (2008): pp. 1195–1246; Vredenburgh, ''Fairness''.

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.

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 Posthoc 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; and at the Causality and Ethics for Society Conference, LMU, Münich, July 24, 2023. I thank the audiences–especially my assigned commentator on the Lund event, Jenny Magnusson, and Lennart Ackermans, Sander Beckers, Mattias Gunnemyr, Martin Jönsson, Michele Loi, Frej Klem Thomsen, and Naftali Weinberger for helpful comments. I also thank Deborah Hellman and an anonymous reviewer for excellent challenges and constructive suggestions. Finally, I am grateful to Lund University's Pufendorf Theme on post hoc interventions and the Danish National Research Foundation (DNRF144) for funding in relation to this article.

OPEN ACCESS

Department of Politics
University of Aarhus, Aarhus, Denmark
E-mail: lippert@ps.au.dk

Department of Philosophy
UiT-Arctic University of Norway, Tromsø, Norway