The Fair Chances in Algorithmic Fairness: A Response to Holm

Clinton Castro (clinton.g.m.castro@gmail.com), Florida International University (corresponding author)
Michele Loi, Politecnico di Milano

Abstract. Sune Holm (2022) argues that a class of algorithmic fairness measures, that he refers to as the "performance parity criteria," can be understood as applications of John Broome's Fairness Principle. We argue that the performance parity criteria cannot be read this way. This is because in the relevant context, the Fairness Principle requires the equalization of actual individuals' *individual*-level *chances* of obtaining some good (such as an accurate prediction from a predictive system), but the performance parity criteria do not guarantee any such thing: the measures merely ensure that certain *population*-level *ratios* hold.

Introduction

Predictive systems can appear[1] to be unfair in a variety of ways.[2] To cite a well-trodden case, COMPAS—a predictive system used across the United States to inform a variety of high-stakes correctional decisions—was found to have differential false positive and negative rates across racial groups.[3] Further, the differentials were oriented such that the system seems to have systematically disadvantaged black defendants and advantaged white defendants.

COMPAS certainly seems unfair, but whether it is unfair in virtue of its unequal error rates is contested.[4] According to one way of looking at the matter, algorithmic fairness consists in having equal false positive and/or negative rates across certain groups, such as black and white defendants. That is, we might think algorithmic fairness consists in ensuring[5]

Equal False Positive Rate (FPR) $=_{\text{df.}} P(\hat{Y} = 1 \mid Y = 0, A = a) = P(\hat{Y} = 1 \mid Y = 0, A = b), \forall (a, b) \in \mathrm{A}$[6]

and/or

Equal False Negative Rate (FNR) $=_{\text{df.}} P(\hat{Y} = 0 \mid Y = 1, A = a) = P(\hat{Y} = 0 \mid Y = 1, A = b), \forall (a, b) \in \mathrm{A}$

Yet, these criteria cannot generally be satisfied by systems that satisfy another set of *prima facie* desirable constraints:[7]

Equal Positive Predictive Value (PPV) $=_{\text{df.}} P(Y = 1 \mid \hat{Y} = 1, A = a) = P(Y = 1 \mid \hat{Y} = 1, A = b), \forall (a, b) \in \mathrm{A}$

and

Equal Negative Predictive Value (NPV) $=_{\text{df.}} P(Y = 0 \mid \hat{Y} = 0, A = a) = P(Y = 0 \mid \hat{Y} = 0, A = b), \forall (a, b) \in \mathrm{A}$

[1] What algorithmic fairness consists in is contested. We will for the most part remain neutral on what, exactly, it consists in, hence our use of the word "appear" and its cognates here and throughout the paper when discussing the merits or accuracy of any account of algorithmic fairness.

[2] In the interest of brevity, our background on these issues will be compressed. For a more comprehensive introduction, see Fazelpour and Danks (2021).

[3] For the original reporting of the case, see Angwin et al. (2016) for a response to the reporting from the creators of COMPAS, see Northpointe Inc. (2016).

[4] For arguments against equal error rates as necessary for fairness, see Hedden (2021). For an argument in favor of parity in error rates as normatively significant, see Hellman (2020).

[5] Our formulations of these ideas are adaptations of the formulations presented in Barocas et al. (2019). The labels for them are taken from Holm (2022).

[6] Where Y is the target variable (e.g., actually being a reoffender), $\hat{Y}$ the prediction (e.g., predicting reoffense), and A sensitive characteristics (e.g., race).

[7] See Kleinberg et al., (2016) and Chouldechova (2017) for impossibility theorems demonstrating this.

As Sune Holm (2022) notes, the fact that these criteria—which he calls the “performance parity criteria”—cannot be mutually satisfied might be understood as showing that the notion of fairness is incoherent and that algorithmic fairness unachievable. However, Holm resists this interpretation. Instead, he argues, each of these measures can be understood as an application of John Broome’s

> **Fairness Principle**: “fairness requires that claims should be satisfied in proportion to their strength” (1990, p. 95)

If Holm is right, he can argue that the notion of fairness is coherent and algorithmic fairness achievable. On his account, we can see that the notion of fairness is coherent by attending to how the performance parity criteria can be interpreted as applications of the Fairness Principle. This, he argues, allows us better understand the key issues in the debates over the measures and allows us to appreciate new arguments for a subset of the measures, which, if successful, ensure the possibility of algorithmic fairness.

We argue that Holm’s interpretation of the performance parity criteria does not work. Our reason for thinking this is that in the relevant context, the Fairness Principle requires the equalization of actual individuals’ *individual*-level *chances* of obtaining some good (such as an accurate prediction from a predictive system), but the performance parity criteria do not guarantee any such thing. The measures merely ensure that certain *population*-level *ratios* hold.

In what follows, we explain the steps in the progression from the Fairness Principle to what we will call *Broomean Algorithmic Fairness*—Holm’s claim that the Fairness Principle is satisfied by applying performance parity criteria to the proper subset of candidates and goods. We then deliver our criticism of Holm’s account. We conclude with a brief summary of what we have done and a sketch of how an account might be developed that captures Holm’s key insights while being sensitive to our criticism of his approach.

Broomean Algorithmic Fairness

We will begin by explaining how, according to Holm, Broome’s Fairness Principle allows us to understand performance parity criteria as different applications of one coherent concept of fairness. Recall that the Fairness Principle states that fairness requires claims to be satisfied in proportion to their strength, and note that in some cases it is not possible to fully satisfy all claims. For instance, we might have fewer kidneys than patients who need them (Holm, 2022). In such a cases, we can, according to Broome,

> go some way toward treating the candidates equally: we can give them all an equal chance of getting the good by choosing between them randomly (Broome, 1984, p. 45).

It is worth clarifying here that in contexts where an agent cannot—for practical or epistemic reasons—satisfy claims in proportion to their strength, no one has a fairness-based complaint against the agent if she gives equal claimants equal chances of getting the goods. Applying this insight to algorithms, Holm claims that we can consider an algorithm fair if it treats the relevant claimants equally by giving them equal chances at the good in question, such as an equal chance at bail.

To get from this thought to Broomean Algorithmic Fairness, we need to combine it with another. It has two parts. One part is Holm’s claim that different fairness metrics identify different sub-populations (e.g., defendants that will not go on to reoffend) as having equal claims to goods (e.g., pretrial release). The other is the idea, implicit in Holm’s account, that the satisfaction of certain of parity criteria, e.g., Equal FPR, ensures that certain groups—e.g., (again) defendants that will not go on to reoffend—have equal chances of getting the goods—e.g., (again) pretrial release.

We now have all of the elements in place for an argument, implicit in Holm’s paper, for Broomean Algorithmic Fairness:

> **P1.** When an agent cannot—for practical or epistemic reasons—satisfy claims in proportion to their strength, fairness is satisfied if she gives equal claimants equal chances of getting the goods.

> **P2.** Performance parity criteria (applied to the proper subset of candidates and goods) ensure that candidates have equal chances of getting the goods.
> **C.** Thus, fairness is satisfied if the agent applies performance parity criteria to the proper subset of candidates and goods.

Let us briefly unpack what Broomean Algorithmic Fairness implies.

For Equal FPR, the sub-population of individuals with the same claim to a decision comprises individuals whose “true label”—their *Y* value—is 1. Consider testing an algorithm with historical data, where the outcome of the variable one tries to predict is known. For example, suppose that this value is equal to 1 for those individuals who will not reoffend when released. If the sub-population characterized by $Y = 1$ (say, those who will not reoffend) is the population of individuals who all have an (equal) claim to of being released, then Equal FPR is required by Broomean Algorithmic Fairness because equal FPR (allegedly) equalizes the chances of release for those who will not reoffend, independently of their membership to specific groups (e.g, men or women, black people or white people).

For other measures, the relevant sub-populations can be identified in a similar way. For example, for Equal PPV, the population of individuals with an equal claim to the decision (i.e., to be released) is not defined by the value of the “ground truth” (i.e., their *Y* value), but by whether they have been deemed by the system as having the feature it is looking for (i.e., their $\hat{Y}$ value).

Holm argues that on the Broomean interpretation, what grounds equally strong claims must be “outcomes,” i.e., *Y* values, and not predictions, i.e., $\hat{Y}$ values. This is initially intuitive: it is far easier to imagine that individuals have claims to concrete resources, circumstances, or actual decisions, on the basis of their actual features (needing a kidney to live, not being one will reoffend). Here, we would like to note that we are not convinced that this further claim is true, but we wish at this point to remain agnostic about it.[8] The objection we raise in the coming section is logically independent of this further position.

Chances vs. Ratios

The central problem with Broomean Algorithmic Fairness is that satisfying any of the performance parity criteria does not necessarily involve equalizing any actual individual's chances. That is, P2. in the above argument—the claim that the performance parity criteria (applied to the proper subset of candidates and good) ensure that candidates have equal chances of getting the good—is false.

There are a few ways to see this.

One way is to simply appreciate the fact that when calculating false negative rates, true positive rates, positive predictive value, and negative predictive value, one simply calculates a *group*-level *ratio.* In the case of the FPR, for example, the calculation runs as follows:

$$\frac{The\ number\ of\ false\ positives\ (FP)}{FP + the\ number\ of\ true\ negatives}$$

Equalizing this ratio across groups is consistent with individual members of those groups having individual-level chances that do not mirror the group-level ratios. To give an extreme example, we can imagine a fully deterministic system where each individual's chance at a false positive is either zero or one, yet the FPR for the group falls between zero and one.

We can further appreciate this point by tending to the fact that many algorithmic systems are, for all practical intents and purposes, deterministic. Considering a case might help to make this vivid. Suppose we are making a website and want to predict on the basis of user information whether to present ad content in English or Spanish. For any given individual, whether they can speak Spanish is not going to change in the relevant time frame (i.e., the time between them first loading the page and the ad appearing). Further, we can imagine that the algorithm for determining whether

[8] For an exploration of this thought that we are sympathetic to, see Loi and Heitz (2022), especially section 4.2.

a given individual speaks Spanish is, like many such algorithms, deterministic in the following sense: the inputs it receives about a user fully determine its output, meaning that in many cases one's chances of receiving, e.g., a false positive, is, for all intents and purposes, either zero or one. Many of these systems are not perfectly reliable (or perfectly anti-reliable) so it is only in a very special subset of cases where one's individual chances will align with anything like, say, the FPR for their group.

This is not to say that the problem only arises in deterministic contexts. We can appreciate this by attending to the basic lesson of intersectionality.[9] Imagine a system that is probabilistic and awards goods to a population composed of men, women, black people, and white people. Such a system could achieve parity across men and women, as well as parity across black people and white people, but it is entirely compatible with these facts that black women in particular face diminished chances at getting the good. If in this case the group-level ratio maps perfectly onto the individual level chances when individuals are described by their race *and* gender, this is a good illustration of how equalizing certain group-level ratios (say, parity across gender and race, considered individually) does not translate into equal chances for, in this case, particular women or black people.

The major problem with Broomean Algorithmic Fairness is that there is a chasm between what the view says individuals are owed (i.e., equal individual chances) and the methods (i.e., performance parity criteria) it uses to deliver to individuals what they are owed. This chasm exists because satisfying performance parity criteria does not (at least not generally) involve equalizing actual individual's individual-level chances. For this reason, the argument for Broomean Algorithmic Fairness is fatally flawed.

Conclusion

In this paper, we have presented and criticized the case for what we have called Broomean Algorithmic Fairness on the grounds that it involves a mismatch between its goals and methods. We would like to note, however, that we do not want to conclude that there is no possibility that Holm is right in thinking that two concepts of fairness, fairness in relation to prediction-based decisions and fair chances, are connected in a significant way. Our claim is more modest: the bridge principle he used to connect the two—which consists in treating equal average probabilities for groups *as if* they entailed equal chances for all the individuals forming those groups—must be rejected. Perhaps, something in the vicinity of this interpretative framework may turn out to reveal some illuminating truth about what one means when one treats performance parity criteria as criteria for fairness. For example, one might start from the observation that parity criteria provide equal chances to *statistical individuals*, e.g., applicants who are men and will in fact repay a loan, and applicants who are women and will in fact repay a loan. Such "men" and "women" can be described in an individualistic language, using singular nouns in the sense in which they appear in national statistics (e.g., "the average American"). Perhaps, a viable interpretation can be found for the idea that a statistical man and a statistical woman have the same claims to a good and that this justifies (in certain circumstances) the attempt to equalize their chances to that good, but much work remains to be done to see if such a proposal can work.

Sources

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016, May 23). Machine Bias: There's Software Used across the Country to Predict Future Criminals and It's Biased against Blacks. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessmentsin-criminal-sentencing

Barocas, S., M. Hardt, and A. Narayanan. 2019. Fairness and machine learning. fairmlbook.org.
Accessed August 27, 2021.

Broome, J. 1984. Selecting people randomly. Ethics 95 (1): 38–55.

Broome, J. 1990. Fairness. Proceedings of the Aristotelian Society 91: 87–101.

[9] Namely, one of the main lessons of Crenshaw (1989).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153– 163.

Crenshaw, Kimberlé Williams. 1989. “Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory, and Antiracist Politics.” *University of Chicago Legal Forum* 1989: 139–167.

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, 16(8), e12760. https://doi.org/10.1111/phc3.12760

Hedden, Brian (2021). On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs* 49 (2):209-231.

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

Holm, S. The Fairness in Algorithmic Fairness. Res Publica (2022). https://doi.org/10.1007/s11158-022-09546-3

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv preprint arXiv:1609.05807.

Loi, Michele and Christoph Heitz (2022.) Is calibration a fairness requirement?: An argument from the point of view of moral philosophy and decision theory. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT ’22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3531146.3533245

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

Compliance with Ethical Standards

No conflict of interest to disclose.