December 13, 2019

Algorithmic Fairness: *How to Combine Classification Parity and Predictive Calibration*

1. Motivation and Overview

At its simplest, an algorithm is a series of well-specified instructions written in computer code with the purpose of automatically carrying out a task of interest. Algorithms are increasingly used by public and private sector entities to streamline decisions about healthcare, welfare benefits, child abuse, public housing, neighborhoods to police, bail and sentencing. There is considerable scholarly debate about the technical and ethical challenges that these algorithms pose for individuals and societies. For example, a pressing concern about algorithms is that they exacerbate existing inequities and structural biases in society (see, e.g., O'Neil, 2016; Barocas et al., 2016; Eubanks, 2018).

This project focuses on the ongoing debate among computer scientists, legal scholars and moral philosophers about the fairness of algorithms used in the criminal justice system. In this context, fairness is typically understood as predictive fairness or classification fairness. The received view in the literature, partly based on a number of impossibility theorems, is that no algorithm can satisfy fairness along both dimensions. The literature has thus often been concerned with tradeoffs between achieving one or the other form of fairness. But I believe that this framing of the debate is mistaken.

The project will first demonstrate a "possibility result" according to which classification fairness and predictive fairness, when properly understood, can be concurrently achieved. The second contribution of the project is to show that the specific notions of predictive fairness and classification fairness deployed in this possibility result provide the right normative framework for thinking about what algorithmic fairness consists in. The two contributions of the project will yield two scholarly articles to be submitted to peer-reviewed journals in philosophy. I am requesting release time and travel funding to support this work.

2. Background: The COMPAS Controversy

Algorithmic fairness in criminal justice became a prominent topic of discussion in 2016 when ProPublica published the piece 'Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks' (Angwin et al., 2016). ProPublica analyzed the software COMPAS to determine whether it was racially biased. This software, whose full name is Correctional Offender Management Profiling for Alternative Sanctions, relies on an algorithm that calculates a person's risk score of recidivism based on different factors such as age and prior arrests. The risk scores range from one to ten, where higher scores correspond to higher likelihoods of recidivism. Although the factors used by COMPAS did not include race, ProPublica showed that the false positive rate was higher for blacks than for whites, and the false negative rate was higher for whites than for blacks. In numbers, 45% of blacks who did not reoffend were classified as being at a high risk of reoffending, but only 23% of whites were. The racial disparities in false negatives were equally stark. Although 28% of blacks who actually reoffended were classified as low risk, as many as 48% of similarly situated whites were. COMPAS would seem to prefer to label blacks 'high risk' even when they did not reoffend, and prefer to label whites 'low risk' even when they did reoffend.

I should note that, in order to calculate false positive and false negative rates, ProPublica compiled a dataset of over 15,000 people in a Florida county. Each person's COMPAS risk score was recorded in the dataset alongside with their subsequent criminal history, including whether the person was rearrested for a criminal offense within a period of two years. The false positive and false negative rates, then, were calculated using 'rearrest' as a proxy for 'recidivism'. This is, of course, questionable, because rearrest is a fallible and itself potentially biased indicator of recidivism. But this need not undermine ProPublica's claim. If rearrests are themselves racially biased against blacks—for example, if a larger percentage of black non-reoffenders are mistakenly rearrested compared to similarly situated whites—then the racial disparities in the algorithm's false positive rates would be even starker than what the ProPublica data indicate. In addition, if predictive algorithms learn from data which are themselves racially biased, feedback loops would cause errors to compound diachronically (see, e.g., Ensign et al., 2018). For the purpose of this project, however, I shall set these discussions aside. This is not because I think they are secondary, but because bracketing them will help to keep the project focused on algorithmic fairness.

On their face, the ProPublica data would seem to show that COMPAS is biased against blacks. But a group of researchers at Northpointe, the company that designed the algorithm, disagreed. They showed that the *predictive* error rate is the same across racial groups (Dieterich et al., 2016). As it turns out, among those who are classified as 'high risk' by COMPAS, the percentage of reoffenders is roughy equal for whites (41%) and blacks (37%), and among those who are classified as 'low risk', the percentage of reoffenders is again roughly equal for whites (29%) and blacks (35%). There are minor racial differences in these numbers, but they have been shown not to be statistically significant (Flores et al., 2016).

ProPublica and Northpointe did not disagree about the numbers, but they invoked different conceptions of algorithmic fairness. According to ProPublica, fairness requires that similarly situated individuals, regardless of their race, be equally subject to classification errors such as false positives and false negatives. Call this *classification parity*. COMPAS does not satisfy classification parity since blacks who do not reoffend are mistakenly classified as being 'high risk' at a much higher rate than similarly situated whites. Yet, according to Northpointe, this racial disparity is irrelevant. Fairness instead requires that among individuals who are classified 'high risk' or 'low risk', the proportion of those who actually reoffend be the same across groups. Call this *predictive parity*. On this interpretation, COMPAS exhibits no racial bias against blacks.

So is the algorithm racially biased or not?

3. Existing Literature

The literature in philosophy and legal theory on the topic is scant, but it is growing rapidly. The principal question here is whether, and if so to what extent, violations of one or the other conception of fairness constitute a moral wrong or violate legal principles of fair treatment and antidiscrimination (see e.g. Barocas et al., 2016; Castro, 2019; Hellman, forthcoming). Besides the philosophical and legal literature, there exists a significant body of research in computer science on algorithmic fairness (for an overview, see Berk et al., 2018; Corbett-Davies and Goel, 2018). This literature has focused on the technical question of whether algorithms can satisfy more than one conception of fairness at the same time or whether there are inevitable tradeoffs between achieving one form of fairness at the cost of the other.

The received view is that, under realistic conditions, different conceptions of fairness are incompatible with one another. For example, in its response to ProPublica, Northpointe argued that when the prevalence or base rate of recidivism is different across groups, it is unrealistic to expect equality in false positive and false negative rates and also maintain predictive parity. This is, in fact, a mathematical impossibility that was demonstrated, as an application of Bayes's theorem, by Alexandra Chouldechova (2017). When this impossibility is paired with the fact that the prevalence of recidivism is higher among blacks, it would seem inevitable to have a higher false positive rate for blacks than for whites. Jon Kleinberg et al. (2017) established a similar impossibility result. They demonstrated the incompatibility of two slightly different conceptions of fairness, call therm *predictive calibration* and *classification balance*. Predictive calibration requires that risk scores ranging from low to high values predict recidivism equally well for individuals belonging to different racial groups. Classification balance requires that the average risk score of similarly situated individuals—that is, individuals who are going to reoffend and individuals who are not—be the same regardless of the racial group they belong to. Predictive calibration and classification balance are similar to, albeit also slightly different from, predictive parity and classification parity.[1]

The impossibility results by Chouldechova and Kleinberg et al., combined, have often been interpreted to mean that different base rates of recidivism make it impossible for any algorithm to achieve predictive

[1] Predictive parity and classification parity both rely on a cutoff or threshold above which an individual is regarded 'high risk' and below which is regarded 'low risk'. Since calibration is about the accuracy of the predictions for each score and balance is about the average score for different groups, no cutoff is needed.

fairness (understood broadly as predictive parity or calibration) together with classification fairness (again, understood broadly as classification parity or balance). The literature has thus often been concerned with tradeoffs between achieving one conceptions of fairness at the cost of the other (see, e.g., Berk et al., 2018; Hellman, forthcoming). But this framing of the debate—I hold—is overly restrictive.

4. Contributions and Methods

(i) ***A Possibility Result*****:** Contrary to much of the existing literature, I will argue that classification fairness and predictive fairness are compatible. The key is to understand them, respectively, as classification parity and predictive calibration. This claim does not contradict any of the existing impossibility results. Chouldechova demonstrates that classification parity and predictive parity are incompatible, while Kleinberg et al. show that classification balance and predictive calibration are incompatible. Instead, I will show that classification parity and predictive calibration are compatible even when the base rates of recidivism differ across groups. I will demonstrate this compatibility by relying on analytic methods in formal epistemology as well as computer simulations.

The argument will be in three parts. First, I will offer a Bayesian analysis of the evidential value of factors such as age and prior offenses that are used as algorithmic predictors of recidivism. Next, I will show that the more independent predictors an algorithm takes into account, the closer the risk scores will be to the true values. In other words, as more independent predictors are used, the risk scores of reoffenders will approach higher values and the risk scores of non-reoffenders will approach lower values. Finally, I will show that when the simulated algorithm relies on a sufficiently large number of predictors, classification parity can be achieved together with predictive calibration, even when the base rates differ across groups. A complication that needs addressing is that when more predictors are taken into account, the set of individuals who bear the relevant characteristics shrinks and thus the uncertainty associated with the predictions increases (see, e.g., Meng, 2014).

(ii) ***Philosophical Analysis of Algorithmic Fairness*****:** The possibility result I have just described asserts the compatibility of classification parity and predictive calibration, but leaves out predictive parity and classification balance. This needs to be philosophically motivated. I will do so in two steps.

First, I will argue that algorithmic fairness, at its core, requires classification parity. Here is the sketch of the argument I will develop. At its simplest, fairness is the requirement that similarly situated individuals be treated the same in some relevant sense. In hiring decisions, for example, fairness does not require that candidates with different qualifications be treated the same. It does require, however, that equally qualified candidates, regardless of race, gender and other irrelevant characteristics, be treated the same in some relevant sense. A hiring process would be racially biased, and thus unfair, if equally qualified candidates were considered more or less favorably depending on their race. In the context of criminal justice, similarly situated individuals should be treated the same by the algorithm, regardless of their race. If blacks who do not reoffend are more often misclassified as 'high risk' compared to similarly situated whites, that would be a violation of classification parity. It would be a failure to treat equally situated individuals the same way in the relevant sense. This is why, in a nutshell, classification parity is essential for algorithmic fairness. At the same time, fairness does not require classification balance. It does not require that similarly situated individuals, regardless of their race, be assigned the same *average* score. As will become clear from the computer simulation, it is possible for two groups to have different average scores and yet be treated the same in the relevant sense so long as the decision threshold is appropriately located.

Next, I will argue that algorithmic fairness does not require predictive parity. If fairness requires that similarly situated individuals be treated the same, individuals who are classified 'high risk' or 'low risk' by the algorithm do not have the right to be treated the same because in no relevant sense are they similarly situated. They only exist as groups created *ex post* by the algorithmic classification. The same argument applies to predictive calibration. It is not a requirement of fairness that those who are assigned *ex post* the same risk score be treated the same. This does not mean that achieving predictive calibration is unimportant. Algorithmic scores should be consistent. It would undermine their intelligibility if scores had a different meaning—that is, if they corresponded to a different likelihood of recidivism—depending on group membership. Thus, even though predictive calibration is not a requirement of algorithmic fairness *per se*, I will argue that it constitutes a consistency requirement that any algorithm should satisfy.

The two claims that I have described—first, that algorithmic fairness requires classification parity, but not classification balance, and second, that any algorithm should satisfy, as a matter of consistency, predictive calibration—will be defended using philosophical conceptual analysis. This is a method employed in philosophy in the analytic tradition for testing the boundaries of a given concept or differentiating between two concepts. The method uses examples, counterexamples and stylized scenarios to test whether or not a proposed definition, analysis or characterization of a concept is adequate. The goal is to arrive at a plausible characterization of algorithmic fairness that combines existing proposals and different pre-theoretical intuitions about what fairness requires.

To articulate my arguments, I will draw from the philosophical literature on statistical evidence (see, e.g., Di Bello and O'Neil, forthcoming) as well as the literature on equality of opportunity (for an overview, see Arneson, 2015). I will also rely on the conceptual connection between fairness in algorithmic classification and fairness of opportunity in hiring and other contexts (see, e.g., Heidari et al., 2019). If I am correct, predicting algorithms used in criminal justice should satisfy classification parity (as a requirement of fairness) and also predictive calibration (as a consistency requirement), but they need not satisfy classification balance or predictive parity. This underscores the importance of the possibility result that this project puts forward. The notions of predictive calibration and classification parity deployed in the possibility result provide a normative framework for thinking about fair and consistent algorithms.

5. Outputs, Work Plan and Funding Requested

The outputs of the project are two scholarly articles to be completed by the end of Summer 2021. The first article—titled 'Is Algorithmic Fairness Possible?'—will put forward the impossibility result described in Section 4, part (i). I am currently working on this article. I reviewed the pertinent literature and wrote extensive notes about the main contentions. I presented some preliminary ideas at Rutgers University last November 2019. The next step is to address the complication I mentioned in Section 4 about the tradeoff between the number of predictors and the uncertainty of the predictions. I plan to complete a draft of this first article by June 2020. This will be made possible by a one-course release time provided by the Faculty Fellowship Publication Program (FFPP) for Spring 2020. This article will be written together with statistician Ruobin Gong at Rutgers University who will ensure that the statistical details are formally correct. Having completed the first article, I will then work on a second article titled 'What Algorithmic Fairness Is and Is Not.' This will contain the philosophical arguments described in Section 4, part (ii). I am requesting a one-course release time from PSC CUNY to work at an accelerated pace on this article during Fall 2020.

In addition to a one-course release time, I am also requesting travel funding to present my research on algorithmic fairness at the 2021 ACM Conference on Fairness, Accountability, and Transparency, abbreviated ACM FAT* (*https://fatconference.org/*). As mentioned in their website, this conference 'brings together researchers and practitioners interested in fairness, accountability, and transparency in socio-technical systems.' This is the ideal venue to present my research on algorithmic fairness and receive feedback. The ACM FAT* conference takes place every year in January. I plan to present my research in January 2021. The location of the conference for 2021 is not yet public. Previous venues have been New York City, Atlanta, and Barcelona. The venue will probably be a major US city. I have budgeted a total of USD 900 for travel, specifically, USD 350 for a domestic flight, USD 400 for a two-night accommodation, meals and local transportation, and USD 150 for conference registration. In case the airfare will exceed USD 350, I will seek additional internal funding from my own institution. Lehman College provides travel funding for faculty who are presenting a paper at an academic conference.

Finally, on the basis of the feedback received at the ACM FAT* conference, I will prepare revised versions of the two articles during Spring 2021. The articles should be ready for submission to peer-reviewed journals in philosophy by Summer 2021. The first article—which argues for the compatibility of classification parity and predictive calibration—is better suited for journals that publish formal results, such as *Philosophy of Science*, *Mind* or *Synthese*. The second article—which argues that classification parity lies at the core of algorithmic fairness, while predictive calibration is a consistency requirement that algorithms should satisfy—is better suited for journals with a focus on moral and political philosophy such as *Philosophy & Public Affairs* or *Journal of Political Philosophy*.

References

Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks, *ProPublica* **https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing**.

Arneson, R. (2015). Equality of opportunity, *Stanford Encyclopedia of Philospphy* .

Barocas, S., and Selbst, A. D. (2016). Big data's disparate impact, *California Law Review* **104**(3): 671–732.

Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art, *Sociological Methods & Research* **https://doi.org/10.1177/0049124118782533**.

Castro, C. (2019). What's wrong with machine bia?, *Ergo* **6**(15): 405–426.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *Big Data* **https://doi.org/10.1089/big.2016.0047**.

Corbett-Davies, S. and Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning, *Manuscript* **arXiv preprint arXiv:1808.00023**.

Di Bello, M. and O'Neil, C. (forthcoming). Profile evidence, fairness and the risk of mistaken convictions, *Ethics* .

Dieterich, W., Mendoza, C. and Brennan, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity performance of the COMPAS risk scales in Broward county, *Technical report*, Northpointe.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing, *Proceedings of Machine Learning Research*, Vol. 81, pp. 1–12.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*, St. Martin's Press.

Flores, A. W., Bechtel, K. and Lowenkamp, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks.", *Federal Probation Journal* **8**(2): 38–46.

Heidari, H., Loi, M., Gummadi, K. P. and Krause, A. (2019). A moral framework for understanding fair ML through economic models of equality of opportunity, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 181–190.

Hellman, D. (forthcoming). Measuring algorithmic fairness, *Virginia Law Review* .

Kleinberg, J., Mullainathan, S. and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores, *in* C. H. Papadimitrou (ed.), *8th Innovations in Theoretical Computer Science Conference*, Vol. 43, pp. 43:1–43:23.

Meng, X.-L. (2014). A trio of inference problems that could win you a Nobel Prize in statistics (if you help fund it), *in* X. Lin, C. Genest, D. L. Banks, G. M. abd David W. Scott and J.-L. Wang (eds), *Past, Present, and Future of Statistical Science*, Vol. 45, CRC Press, pp. 537–562.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, Crown.