Res Publica (2023) 29:177–183
https://doi.org/10.1007/s11158-023-09587-2

Discrimination, Fairness, and the Use of Algorithms

Sune Hannibal Holm [1] · **Kasper Lippert-Rasmussen**[2]

Accepted: 7 March 2023 / Published online: 28 March 2023

Computer programs trained on vast datasets to produce predictions of people are becoming ever more integrated in decision-making processes affecting people's lives in important ways. Predictive algorithms are deployed to make diagnostic and therapeutic decisions in health, to select candidates for jobs, and to decide whether prisoners should be offered parole, to mention but a few.

On the face of it, there seem to be good reasons to look to predictive algorithms for decision support. First, when making decisions under risk algorithmic predictions may improve our decision-making. A manager considering whether to offer an applicant a job may wonder whether the applicant will quit within a year. A physician may wonder whether a treatment will improve the condition of a patient. And a judge considering whether to release a defendant on bail may wonder whether the defendant will commit new crimes if released. In these cases a popular approach is to consider the little we do know about the individual decision-subject and make a best guess on that basis about whether she will default, benefit from treatment, or reoffend. Given their ability to uncover patterns in data that humans are unable to detect, machine learning algorithms have been shown to be able to provide much more accurate prediction than human judgements and hence their input may improve the quality of decision-making.

Second, it is well documented that human decision-makers suffer from a variety of cognitive biases and may be moved by irrelevant considerations (Kahneman et al. 1982). Decision-makers may unknowingly tend to treat decision-subjects more harshly depending on their race, gender, or, surprisingly, on whether the decision-

✉ Sune Hannibal Holm
suneh@ifro.ku.dk

Kasper Lippert-Rasmussen
Lippert@ps.au.dk

[1] Dept. Food and Resource Economics, University of Copenhagen, Rolighedsvej 23, 1958 Copenhagen, Frederiksberg, Denmark

[2] Dept. of Political Science, Aarhus University, Bartholins Allé 7, 8000 Aarhus C, Denmark

maker is hungry or tired. Relying on algorithms trained on hard data would seem to ensure uniform and unbiased decision-making based on the facts.

However, as documented by a series of important publications this preconception of algorithms and algorithmic decision-making is unfounded. There is always a risk that algorithms discriminate (Barocas and Selbst 2016). Thus the topic of discrimination and algorithmic fairness was kick-started by an article in which ProPublica, a group of investigative journalists, claimed that an algorithm called COMPAS, widely used in American courts to make bail decisions, was biased against Blacks (Angwin et al. 2016). ProPublica's criticism was based on an analysis of COMPAS which showed that the algorithm's error rates for Black and White defendants were very different. Blacks were almost twice as likely as Whites to be classified as being at high risk of reoffending when they were not. Whites were almost twice as likely as Blacks to be classified as being at low risk of reoffending when they were not.

In another important study Buolamwini and Gebru (2018) showed that widely used commercial gender classification algorithms, from amongst others IBM and Microsoft, displayed significant differences in accuracy for people of different gender and skin colour. The algorithms examined were all much better at classifying male than female faces, and they all did better on light-skinned than dark-skinned images. For females with dark skin IBM's classifier turned out to have an error-rate of 34.7%. For males with light skin the error-rate was 0.3% (Buolamwini and Gebru 2018, p. 9).

A third example of how using algorithms as the basis for decision-making may disadvantage one group relative to another is presented by Obermeyer et al. (2019). An algorithm used to allocate healthcare to approximately 200 million people in the US every year was shown to be significantly less likely to refer Black patients to personalized healthcare programmes than equally sick White patients.

An important insight to come out of these and other studies documenting algorithmic bias is that there is a lack of diversity in the data used to train algorithms as well as in the computer programers designing the algorithms. This problem has been presented with great force by Criado-Perez (2019) in which she provides multiple examples of how women have been underrepresented in datasets and the design of many kinds of societal infrastructure from traffic planning to algorithmic systems. With algorithms becoming widespread and used to make decisions about allocation of resources and opportunities to millions of people these concerns become ever more pertinent to consider.

The recognition that algorithms may be biased in the sense that they unjustifiably perform better on some groups than others has given rise to a whole area of research in machine learning focusing on *algorithmic fairness* (Mitchell et al. 2021). The central aim of algorithmic fairness researchers in machine learning is to 'uncover and rectify' unjustified performance disparities with respect to salient groups such as groups defined by race and gender (Mitchell et al. 2021).

In recent years several formal definitions of algorithmic fairness have been proposed (Verma and Rubin 2018). However, despite the proliferation of formal fairness definitions, it has also been remarked that little advance has been made concerning the question of what it *means* for an algorithm to be fair (Corbett-Davies et al. 2016). A key question thus concerns how to use these formal definitions (Loi and Christen 2021). In fact, philosophers have been called upon to engage in the debate about

algorithmic fairness and apply their expertise to make progress with respect to concerns about fairness, disparate impact, and discrimination (Narayanan 2018).

The articles published in this topical issue aim to contribute to the debate about the responsible use of algorithmic systems by offering analyses of key questions and arguments concerning fairness and discrimination in the context of algorithmic systems, which may inform the design and application of such systems for practical purposes. As such the issue should also be of interest to a wide range of stakeholders, including policy-makers, public administrators, businesses, and NGOs, many of which have developed principles and guidelines for the design and use of AI.

The Articles

The philosophical literature on discrimination is much concerned with the question *why* discrimination is wrong, when it is wrong. This approach is understandable given that history seems to provide an abundance of cases where there is consensus that a discriminatory practice is in place.[1] However, when it comes to the deployment of algorithmic decision-making it turns out that there is no comparable consensus. Many theorists, for instance, question whether the unequal false positive and false negative rates in the COMPAS case constitute an algorithmic unfairness. Hence, there is a more urgent need for careful consideration of what constitutes a discriminatory practice in an algorithmic setting: what makes (the use of) an algorithm discriminatory? To answer this question an important task is to consider how established definitions of discrimination relate to the specific case of algorithmic discrimination. The first two articles in the issue analyse cases of alleged algorithmic discrimination against the backdrop of recent theories of discrimination and consider how analyses of algorithmic discrimination may require revisions of established accounts of discrimination.

In their contribution Nappo et al. argue that in order for an algorithm to discriminate a certain counterfactual condition must be met. The idea is that an individual is only justified in claiming that discrimination occurs if the probability of receiving the same algorithmic prediction would be different had the individual belonged to a different socially salient group. This condition captures the idea that there is something discriminatory about an algorithm which is such that a woman would have a higher chance of being classified as a non-defaulter if we changed her gender to male and kept all other features the same. Nappo et al. then argue that if their account is sound then it presents a challenge to some of the most influential accounts of discrimination in recent years including those proposed by Hellman (2008) and Lippert-Rasmussen (2014). Another important implication of Nappo et al.'s account is that there is a *conceptual* constraint on how far claims about discrimination can be supported by

[1] This is not to say that there are no cases of differential treatment in a non-algorithmic context where it is controversial whether they amount to discrimination. It is simply to say that there is a set of paradigmatic cases, where almost anyone would agree that, say, differential treatment amounts to discrimination, e.g., gendered voting rights and apartheid laws.

reference to the actual consequences of deploying an algorithmic decision-making system. Counterfactual considerations are required too.

One of the valuable outcomes of analysing discrimination in the context of algorithmic decision-making is that it might bring out important insights into the ethics of discrimination more generally. In his article Thomsen examines what the debate about algorithmic discrimination can learn from recent philosophical work on the ethics of discrimination. Thomsen's contribution directs attention to three issues concerning discrimination which benefit from consideration of algorithmic discrimination: the distinction between direct and indirect discrimination, the moral significance of disadvantageous treatment, and the question of when discrimination justifies not using algorithms to make decisions. Thomsen argues that in the context of algorithmic discrimination some central defining characteristics of direct and indirect discrimination do not apply and, thus, the case of algorithmic discrimination raises doubts about standard distinctions between direct and indirect discrimination. Second, arguing that algorithmic discrimination is mainly *between* and not *against* groups, Thomsen suggests that the case of algorithmic discrimination helps to show that indirect discrimination is more important than the attention devoted to it indicates. And thirdly, Thomsen finds, partly for reasons such as those mentioned above, that it is not clear that the *use* of an algorithm guilty of morally bad discrimination is morally bad. Sometimes non-algorithmic decision-making processes are even worse.

A natural way to think about what it requires for an algorithm such as COMPAS to be fair is to take it to be fair if it works equally well for different salient groups. However, as machine learners were quick to show, it is impossible for an algorithm to work equally well for different groups because the base rate of the feature that the algorithm is designed to predict will typically differ for these groups. And, after all, even the most accurate algorithms will make mistakes. This means that when designing an algorithm designers face important tradeoffs regarding the distribution across groups of different kinds of mistakes that the algorithm may produce (Chouldechova 2017; Kleinberg 2016; Verma and Rubin 2018). While the inequality in false positive and false negative predictions about Black and White defendants was criticized in the COMPAS case, the makers of COMPAS responded that this was simply a foreseeable consequence of ensuring that the risk scores assigned to defendants were calibrated. What would be unfair, they claimed, would be for risk scores not to be equally predictive for defendants regardless of their race, since this would mean that equally risky defendants might not be treated the same.

The second theme of the collection concerns the debate about how to understand and evaluate algorithmic fairness criteria. Given that not all plausible criteria of fairness can be met simultaneously for different salient groups, are there some of the plausible criteria that we should accept as necessary for algorithmic fairness? Or should we consider the fairness of an algorithm to be a matter of context? Does the apparent conflict between different fairness criteria reveal that the very notion of algorithmic fairness is incoherent? These are some of the questions considered in the articles by Castro et al. and Holm.

Castro et al. find that the debate about algorithmic fairness in the machine learning community is premised on a particular definition of when a predictive algorithm is fair. It is fair if and only if it is not wrongfully discriminatory. However, besides

general agreement that wrongful discrimination involves inequality in some respect, there is no consensus about what measure to use to determine whether an algorithm engages in wrongful discrimination in a given context. Assuming pluralism about fairness measures, Castro et al. focus their investigation on the question of how to choose a fairness measure for assessing an algorithm in a given context. Considering three widely discussed fairness measures (fairness through unawareness, equalized odds, and counterfactual fairness) and a wide range of cases, Castro et al. support their claim that which fairness measure to apply depends on context.

The performance measures discussed by Castro et al. are also the topic of Holm's article. Holm considers the view that algorithmic fairness is impossible to achieve because of the incompatibility of achieving equality across groups with respect to popular algorithmic performance measures. Holm's article targets the view that the mathematical incompatibility of classification parity criteria of fairness shows that such formal criteria express different views of what it means for an algorithm to be fair. In response to this claim Holm argues that these different criteria can all be understood as applications of a single fairness principle put forward by John Broome. The real disagreement is about the basis on which individuals have a claim to the good being distributed by the algorithmic decision procedure. Moreover, Holm argues that the Broomean interpretation favours some of the incompatible fairness criteria over others.

The last two contributions consider what might be called *algorithm-external* issues, in that they are both about the way in which the data used to train algorithms may themselves be the result of unjust practices and misleading social categorizations.

One of the central criticisms levelled against algorithmic decision-making procedures has been that they are unfair *because* they *compound* injustice (Hellman 2020). In his contribution, Lippert-Rasmussen presents and discusses the objection that algorithmic procedures are objectionable because they disadvantage certain groups by relying on information that is itself a result of unjust disadvantages suffered by members of these groups. While Lippert-Rasmussen finds that actions that compound injustice are often wrong, the reason why they are wrong, when they are, is not that they violate a duty not to compound injustices. When it is wrongful that an algorithm such as COMPAS compounds injustices to Black Americans, this is, according to Lippert-Rasmussen, because it is disrespectful or because it violates a duty not to disadvantage an already unjustly disadvantaged group disproportionally.

As the case of recidivism prediction shows, there is great interest in using social data to produce predictive algorithms. In her article Greene discusses the pitfalls of taking social science concepts to represent accurate categories that can be used for training predictive algorithms. Greene focuses on algorithms for recidivism prediction. The algorithms are trained on social data acquired using questionnaires. Greene points out that using answers to questionnaires as data suggests a kind of precision and uniformity that she argues is unjustified. When it comes to social variables and concepts such as 'has criminal associates' those who respond the same may have very little in common. Some who answer in the positive might not be associating in the relevant sense of having companions that encourage criminality. While this might suggest that social science concepts are unfit for the task of making predictive

algorithms, Greene emphasizes that such variables can indeed be used, but it requires careful attention to the way in which questions are formulated.

This topical issue grew out of an online workshop in November 2020 as part of the Responsible AI initiative funded by the Faculty of Science, University of Copenhagen in collaboration with the Center for the Empirical and Philosophical Study of Discrimination at the University of Aarhus. The editors, Sune Holm and Kasper Lippert-Rasmussen, would like to thank all those who attended the workshop and provided comments and feedback on the papers. We would like to thank the work done by anonymous reviewers whose comments have been of great value for the contributions to the issue. Kasper Lippert-Rasmussen would like to thank DRNF144 for financial support. Finally, we would like to thank the editors of *Res Publica*, Clare Chambers and Sune Lægaard, for their support and for giving us the opportunity of publishing the articles as a topical issue.

References

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*. Accessed 1 February 2023. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

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

Buolamwini, J., and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research* 81: 1–15.

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

Corbett-Davies, S., Avi Emma Pierson, Feller, and Sharad Goel. 2016. A computer program used for bail and sentencing decisions was labeled biased against Blacks. It's actually not that clear. *Washington Post*, October 17. Accessed 1 February 2023. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/.

Hellman, D. 2008. *When is discrimination wrong?* Harvard University Press.

Hellman, D. 2020. Measuring algorithmic fairness. *Virginia Law Review* 106: 811–866.

Kahneman, D., Paul Slovic, and Amos Tversky, eds. 1982. *Judgment under uncertainty: Heuristics and biases*. Cambridge University Press.

Kleinberg, J., S. Mullainathan, and M. Raghavan. 2016. Inherent trade-offs in the fair. Determination of risk scores. *arXiv e-prints*.

Lippert-Rasmussen, K. 2014. *Born free and equal? A philosophical inquiry into the nature of discrimination*. Oxford University Press.

Loi, M., and M. Christen. 2021. Choosing how to discriminate: Navigating ethical trade-offs in fair algorithmic design for the insurance sector. *Philosopshy and Technology* 34: 967–992. https://doi.org/10.1007/s13347-021-00444-9.

Mitchell, S., E. Potash, S. Barocas, S, A. D'Amour, and K. Lum. 2021. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application* 8: 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902.

Narayanan, A. 2018. Tutorial presented at the 'Conference on fairness, accountability, and transparency' on February 23 2018. https://fairmlbook.org/tutorial2.html. Accessed 1 February 2023.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (6464): 447–453. https://doi.org/10.1126/science.aax2342.

Perez, C. C. 2019. *Invisible women: Data bias in a world designed for men*. New York, NY: Abrams Press.

Verma, Sahil, and J. Rubin. 2018. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/3194770.3194776.