What is the Point of Algorithmic Fairness?: Condensed Version

Antonella D. Basso

Department of Philosophy, Agnes Scott College

Dr. Harald Thorsrud

December 1, 2020

Abstract

In recent decades, the growing use of predictive, data-driven algorithms has posed numerous ethical challenges for society, calling for urgent reassessments of our models and the outcomes they produce. In light of the grave social injustices caused by algorithmic decision-making, the notion of algorithmic fairness has become a more pressing topic and an emerging area of research. However, despite the tremendous progress that has been made in this space, most existing accounts of algorithmic fairness see it and define it as a mere statistical means to produce fair outcomes. Thus, making them subject to synonymizing distinct notions of fairness and vulnerable to the same problems they aim to solve. For this reason, we ought to understand algorithmic fairness as an ends-driven definition that appeals to acceptable standards of justice and is instrumental to generating and employing adequate statistical methods. Only such an account will guarantee the kind of outcomes we can reasonably accept as fair and capture the true point and motivation for algorithmic fairness. The goal of this paper is to distinguish between procedural and substantive fairness in the context of algorithms; demonstrate how statistical definitions miss the point of algorithmic fairness; and provide a more compelling, ends-driven account that is grounded on American anti-discrimination laws and supported by Rawlsian principles of distributive justice.

Unfairness and Algorithmic Bias: With our growing dependence on algorithmic decision-making, the growing instances of algorithmic injustice have been overshadowed by the field's practical advantages. For this reason, the ethical use of predictive ML algorithms is an area that deserves much attention, and hence why research in the space of algorithmic fairness has grown significantly. Yet, despite the massive efforts and significant progress made in this area, there continues to be much controversy over what constitutes algorithmic fairness and establishing a single definition remains a major challenge.[1] On the other hand, identifying sources of unfairness, from which to draw conclusions about how to reduce it, has proven to be a much simpler task. As research in this space reveals, causes of unfairness are directly related to bias. Panch et al, defines algorithmic bias as "the instances when the application of an algorithm compounds existing inequities in socioeconomic status, race, ethnic background, religion, gender, disability or sexual orientation to amplify them and adversely impact inequities" in society.[2] The most common among these is bias which is inherently part of the data used to train predictive algorithms and is oftentimes the product of "biased device measurements", or "historically biased human decisions".[3] Since algorithms themselves are incapable of making moral judgements about the data they take as input, they merely reflect the bias within it. Aside from the more obvious form of algorithmic bias that is intentional and stems directly from the objectives of an algorithm, another common form of bias is that which is caused by missing data (in regards to a particular group), which creates a dataset that inaccurately represents a target population.[3] In lacking this necessarily inclusive data, the algorithm inevitably becomes biased towards one group and can make inaccurate and unjust predictions about the other. One final type of algorithmic bias worth mentioning is bias caused by "proxy" attributes used in place of sensitive attributes that differentiate a privileged group from an underprivileged group (e.g., race, gender, economic class).[3] As these "proxies" often do not truly mask sensitive attributes, but instead stand as place holders, they can easily be exploited to reveal them. All forms of algorithmic bias, can and often lead to discrimination, which contributes to "self-fulfilling prophecies and stigmatisation in targeted groups, undermining their autonomy and participation in society".[4] Thus, the effort to understand and promote algorithmic fairness is critical and those who design and employ predictive algorithms have a legal and moral obligation to ensure their technologies are not only accurate, but substantively fair.

Neutrality: Despite how effortless it has been to identify instances of algorithmic bias and unfairness, it continues to be surprisingly difficult to establish what algorithmic fairness should look like and how we ought to define it. In legal terms, American anti-discrimination laws "prohibit unfair treatment of people based on sensitive attributes".[5] These laws appeal to two concepts when evaluating the "fairness" of a decision-making process: disparate treatment and disparate impact. "A decision-making process suffers from disparate treatment if its decisions are (partly) based on the subject's sensitive attribute".[5] On the other hand, it suffers from disparate impact if its outcomes disproportionately affect people with certain sensitive attributes. Disparate treatment takes shape in algorithmic decision-making as "unawareness", "blindness" or "neutrality"— the omission of individuals' sensitive attributes from training data (“anti-classification”), which on surface level appears “fair”.[6] How could algorithms possibly be biased towards a particular group or individual if they lack access to the basis on which to form such biases? The fact of the matter is that they can, and there is plenty of evidence to suggest that they often do. As is the case in healthcare, many “neutral” predictive algorithms have been shown to favor whites when making decisions about who’s health care needs are more urgent. These “risk-prediction models”, are trained purely on insurance claims data to anticipate “which patients would benefit from extra medical care, based on how much they are likely to cost the healthcare system in the future”.[7] One particular instance revealed that although black patients demonstrated more chronic illnesses than white patients, they were not “flagged” by the algorithm as needing extra care. These injustices, which are not obvious at first glance, are a direct cause of the failure to account for the systemic oppression that plagues our healthcare system and is reflected in insurance claim records. In using health cost as a proxy for health, software developers and users fail to acknowledge the fact that health care spending on black Americans is generally much lower than on white Americans with similar health conditions for reasons unrelated to actual health. What this means for black folks whose data is fed into these algorithms is that they generally have to be a lot sicker than white folks to be given the same priority.[7] Thus, not only are health costs unable to determine how sick people are, but using them as a basis for health care allocation perpetuates racial inequality. This is a prime example of how using historically biased data to make future decisions about the distribution of social goods produces more of the same biased data, which when fed back into the algorithm, creates an adverse feedback loop that gives rise to discrimination among other injustices.[8]

Statistically Defined Algorithmic Fairness: Surfeced evidence of algorithmic neutrality’s inability to prevent bias and promote fairness has motivated even more drastic methods to combat algorithmic injustice. Particularly, this has become a massive focus within the statistical community, where attempts continue to be made to quantify fairness in a way that places constraints on algorithmic outcomes. Within the various statistically defined notions of algorithmic fairness, the simplest and most recognized are variations of “statistical parity”. Namely, the idea of “equalizing outcomes across the protected and non-protected groups” within a target population.[1] While some forms of statistical parity do this by measuring the difference between positive prediction rates, others do so by measuring and restricting the ratio of positive prediction rates across groups. “Disparate impact” for example, mathematizes this legal principle, requiring that there be a high ratio between an algorithm’s positive prediction rates for both groups in a target population. So, if a “positive prediction” indicates acceptance for a job for example, this condition ensures that the proportion of accepted applicants is similar across groups. Similar variations of these “aware” statistical notions of fairness play on the idea of equalizing false positive and false negative rates among groups. That is, requiring that an algorithm’s error (inaccurate predictions) be roughly the same across groups, which ensures an equalized rate of individuals harmed by the algorithm.[3] In such cases, algorithms are justified and deemed fair on the grounds that sensitive attributes, despite being present, don’t play a role in advantaging or disadvantaging individuals. Despite these and other statistical definitions of fairness having significantly contributed to the fight against algorithmic injustice, like algorithmic neutrality, they have failed to guarantee a substantive level of fairness on both a group and an individual level. Their integration in loan lending prediction models for example, oftentimes don't produce outcomes we can reasonably accept as fair. To visualize how this plays out, suppose that a bank is using one such algorithm to decide which individuals to grant loans to based on creditworthiness.[9] Imagine that members of two groups are applying for loans: group A and group B. Letting group A represent the privileged group, and group B, the underrepresented minority group, suppose that group A has 100 applicants, 58 of which are deemed qualified, while group B also has 100 applicants, only 2 of which are deemed qualified.[5] If the bank decides to accept 30 applicants and satisfies a statistical parity metric, then their algorithm would predict that the bank ought to confer roughly 29 offers to group A, while only 1 offer to group B (50% of qualified applicants from each group). Given that these loans can provide more opportunities and better living conditions, it follows that group A will continue to thrive, while members of group B will continue to face hardships through no fault of their own, which inevitably contributes to enlarging the gap between group A and group B over time.[5]

Argument: Evidently, neutrality and statistical accounts of fairness do not live up to their apparent aims to produce fair outcomes as required by law and demanded by justice. I shall argue that this is not due to a lack of precision or quantitative complexity within such models, but rather to the power that they're given as definitions of algorithmic fairness. Defining algorithmic fairness in statistical terms presupposes that algorithmic fairness necessarily results from the successful application of a statistical method, letting it serve as sufficient justification for outcomes and hence, determine whether we've achieved the kind of fairness we ought to accept. Moreover, accepting statistical definitions implies that algorithmic fairness is merely procedural— a statistical means of obtaining "equal" and "fair" quantitative values, rather than the end-state of obtaining substantively fair outcomes. For this reason, I shall argue further that finding a solution to this problem is analogous to establishing a distinction between procedural and substantive fairness (or procedural and distributive justice). In a judicial context, "procedural justice concerns whether the processes used to arrive at some outcome are fair, whereas distributive justice concerns whether the outcome itself is fair".[10] In the examples mentioned previously, we've evidently only achieved the former, when in fact we seek and ought to produce the latter. This confusion between procedural and substantive notions of fairness not only leads society to misinterpret algorithmic decisions, but leads those who are harmed by them to accept unfavorable outcomes "when procedural fairness is high".[10] If the goal of algorithmic fairness is to promote substantively fair and equitable outcomes for individuals in this context, then we ought to look beyond current limited statistical procedures and accept an ends-driven definition that places a necessary constraint on algorithmic accountability. Under a more robust view, it would follow that algorithmic fairness has been achieved iff the outputs an algorithm produces yield the kind of fair outcomes that are required by law and socially acceptable principles of justice. Only such an account will guarantee the kind of outcomes we can reasonably accept as fair and capture the true point and motivation for algorithmic fairness.

What is Fairness?: As mentioned previously, a legal notion of fairness, as it pertains to decision-making or distributing resources, is roughly defined in terms of “disparate treatment” and “disparate impact”. Renowned political philosopher, John Rawls, expands on these principles and provides a more exhaustive conception of fairness that is grounded within his theory of distributive justice. In his famous work, *Justice as Fairness*, he articulates “the central liberal ideas that cooperation should be fair to all citizens regarded as free and as equals”.[11] Ultimately, he believes in reforming the structure of society by reshaping the way in which political policies and principles are made. Through his thought experiment, “the original position” or “the veil of ignorance”, Rawls emphasizes the importance of changing the method by which principles of justice have come to be, which in the past have been alienating and created vast inequalities. In this way, he believes in promoting equality and creating a fair and just society through establishing an ethical and unbiased political system and a just method of distribution. Negatively, he asserts that citizens are not entitled to more of the benefits of social cooperation simply because of their sensitive attributes.[12] That is, in accordance with the disparate impact law, “the fact that a citizen was born rich, white, and male provides no reason in itself for this citizen to be favored by social institutions”.[11] Positively, Rawls's distributive thesis surrounds the idea of “equality-based reciprocity”.[12] Namely, social goods should be “distributed equally, unless an unequal distribution would be to everyone's advantage”.[11] Rawls stresses that equality should be seen as the outcome of fostering a fair distribution of goods for the sake of justice and valued in the sense that when citizens regard one another as equals, they are committed to preserve the ethical norms of this relation. One of his main reasons for valuing equality in this way is the fact that “it is morally wrong for a fraction of society to be amply provided for while the rest are forced to live in scarcity and hardship through no fault of their own”.[11] Under Rawls’s view, inequality in itself is bad because it not only creates a sense of inferiority for those who are “worse off”, forcing them to accept unjust treatment and inadequate ways of living which diminishes what he calls “the social bases self-respect”, but it allows the “well-off” to view them as inferior and undeserving, encouraging a will to dominate, which violates the moral principles associated with being human.[11] Moreover, Rawls stresses the importance of publicity within justice as fairness, stating that in a “well-ordered society”, all citizens ought to be aware of and mutually agree upon the basic structure and fundamental principles of justice.[12] The idea behind this being that all free and reasonable citizens have the right to know and accept the principles that will be enforced on them, so as to prevent an underlying social hierarchy of power. As Rawls puts it, fairness requires that, in “public political life, nothing need be hidden... there is no need for the illusions and delusions of ideology for society to work properly and for citizens to accept it willingly”.[11]

Substantively Defined Algorithmic Fairness: Given the supplemental appeal of Rawlsian principles of fairness and distributive justice and their conformity to our nation’s anti-discrimination laws, I shall argue that a promising definition of algorithmic fairness can be obtained by appealing to these philosophical ideologies in tandem with established legal standards. Thus, I shall define algorithmic fairness in two parts as follows:

1. Algorithmic fairness is an end-state of an algorithmic decision-making process that abides by socially acceptable standards of substantive fairness and is instrumental to generating technical (be them statistical) procedures.
2. In line with American anti-discrimination laws and Rawlsian principles of distributive justice, algorithmic fairness has been met iff the outcomes an algorithm produces on an individual or group level do not suffer from disparate treatment or disparate impact; are distributed equally, unless an unequal distribution would be to everyone's advantage; are obtained on grounds acceptable to all reasonable people; and do not in any way diminish the social bases of self-respect or the status of individuals as free and equal members of a democratic society.

Here, (1) serves to articulate the form algorithmic fairness ought to take and lays the foundation for how we ought to understand it. On the other hand, (2) provides the conditions under which we can reasonably accept the outcomes of an algorithm as substantively fair. I shall argue that a viable definition of algorithmic fairness must be one in which (1) necessarily remains constant and (2) conforms to contemporary and socially acceptable standards of fairness and justice. While (2) is flexible, it nonetheless must protect our equal rights as human beings and citizens of a democratic state, and maintain a sufficient level of transparency and mutual agreement. I shall stress that the goal of such a definition is not to provide a formula that can be weaved into an algorithm to produce fair outcomes. Rather, it is meant to reveal the true form of algorithmic fairness and outline its demands. We may think of this new framework as a way of separating two distinct ideas that were previously synonymized, which in turn, provides a more robust form of algorithmic fairness that prevents us from accepting unfair outcomes in disguise. Demanding that outcomes under given statistical conditions satisfy this definition allows us to make definitive claims about the fairness of our outcomes and prompts appropriate reassessments of our procedures. By focusing on ends rather than means in this way, we are given a general idea of what our algorithmic outputs should look like, which then allows us to derive and/or enact particular statistical methods. If after exhaustively employing such methods we still do not satisfy the conditions for algorithmic fairness, we may infer that the model or project in question itself is incapable of producing justifiably fair outcomes, in which case we would be faced with a broader moral dilemma about whether or not we ought to use it.

Possible Objections: One may object to my claims about a procedural notion of fairness' inability to produce a substantive level of fairness under the belief that we have merely failed to derive the right mathematical formulation. Namely, one may endorse the idea that statistical definitions have not yet, but can in fact define algorithmic fairness. Under this view, and a seemingly popular one at that, the motivation behind research in algorithmic fairness is precisely to derive such a statistical definition. It would appear thus that statisticians and other researchers in this space have not only acknowledged and continue to acknowledge the shortcomings of existing statistical definitions, but they are driven to derive one such definition that will be successful in guaranteeing substantively fair outcomes. Moreover, one may object to the assertion that the point of algorithmic fairness is to produce a substantive level of fairness on the grounds that algorithms are procedural mathematical entities. Many are likely to side with the idea that achieving algorithmic fairness equates to achieving fairness in merely this procedural sense. Presumably, this line of thinking is backed by the idea that achieving a more robust form of fairness is a social and political matter, not one supported by science. Placing this level of emphasis on what algorithmic outcomes entail, provokes debates that statisticians, developers, and other such researchers are not equipped to handle. One may argue moreover that the kind of fairness people deserve is indeed one that is far beyond procedural, however it is not one that algorithms can or ought to provide. Those who work on behalf of algorithmic fairness have a duty, not to satisfy a political theory of justice, but rather a legal and perhaps moral responsibility to motivate an ethical production of algorithmic outcomes. Whether these outcomes are to be accepted as substantively fair is something that an algorithm can neither determine nor guarantee and must be up to society and policy-makers to decide.

Response to Objections: In response to the first objection, I shall firstly point out that in undermining the competence of procedural accounts of algorithmic fairness, I am in no way suggesting that it may never be possible for a statistical method to succeed in producing substantive fairness whenever used. Rather, I am arguing that even if such a metric did or could exist, it does not mean that we should let it define algorithmic fairness. Letting a statistical procedure define fairness in this context, despite how effective it may seem, as has been shown, leads us to accept algorithmic outcomes as fair without any knowledge of what fairness should look like other than the satisfaction of some complicated formula. The major problem with defining algorithmic fairness procedurally, as I've mentioned, is that even when our models don't produce substantive fairness, we continue to justify them because in satisfying these definitions, we trust that they always will. Moreover, if and when we do acknowledge instances where they've failed to produce substantively fair outcomes, we may not know whether these were the only times they've failed. A more robust definition of algorithmic fairness allows us to be proactive in the fight against injustice, as opposed to reactive. For this reason, as I have stressed, we need to distinguish between procedural and substantive fairness, and work to mathematize ways of achieving fairness, not ways of defining it. In response to the claim that algorithmic fairness is merely a procedural endeavor and does not pertain achieving a substantive level of fairness, I shall argue that if the goal of algorithmic fairness is to effectuate a statistical means to produce fair numerical outcomes, then it would appear that existing procedural notions of algorithmic fairness have already done their job. However, much evidence in the research space of algorithmic fairness suggests that statisticians and researchers in the field are working towards a much larger goal that has not yet been accomplished. As neither they nor society are satisfied with the outcomes such definitions have produced, it's safe to assume that the goal of algorithmic fairness is not merely to provide a means to produce fair numerical outcomes. It is their incapability to guarantee a substantive level of fairness that speaks to why we should not let statistical methods define algorithmic fairness. Again, the fact that they aid in the pursuit of fairness is not up for debate, but they cannot be used as definitions of algorithmic fairness when they serve no greater purpose than making numbers fair. Thus, they ought to be understood as means of producing fairness and not ways of defining it in this context. For the sake of the argument, let us suppose that the claim of the objector holds. That is, let us assume that the goal of algorithmic fairness is to produce mere procedural fairness. I shall then argue that this is not what we would want from algorithmic fairness. Precisely, because we would hence be choosing to ignore major societal issues that arise from the use of predictive ML algorithms. Given that injustices persist irrespective of the employment of such definitions, algorithmic fairness is, under this view, fundamentally pointless and we ought to reestablish its aims. To those who claim that substantive fairness is beyond the scope of algorithmic fairness, I ask; what then is the point of producing fairness in numbers if not to answer some greater call from justice?

Conclusion: As a society, we may not have yet come to a consensus on what substantive fairness looks like, but we can agree on the difference between procedural and substantive fairness and that the two are neither synonymous nor constitutive of the other. Likewise, we ought to recognize that mathematized notions of fairness are procedural, while ends-driven and justice-based accounts are substantive, and that numerical fairness does not entail substantive fairness. Only a definition that reflects the demands of justice and embodies a broader conception of fairness can guarantee an ethical distribution of social goods. It may not be an algorithm's job to provide this level of fairness, but those who develop and use them have a moral duty to distinguish the instrumental nature of statistical methods employed from a true characterization of fairness in an algorithmic context. A failure to do this, leads us to accept a standard of fairness that is merely procedural, making us vulnerable to discrimination among other injustices. A failure to do this, leads us to accept a system of distribution that is not in our best interest and does not fully acknowledge the equal moral worth of persons. It is not enough to remedy unfairness after it has occurred. We must redirect our attention to finding ways of preventing it. We must focus not on maximizing the equality of algorithmic outcomes, but rather maximizing the equality they produce. This is what people deserve from algorithmic fairness. This is what society expects from justice.

References

1. Yona, Gal. "A Gentle Introduction to the Discussion on Algorithmic Fairness." Medium, October 7, 2017. https://towardsdatascience.com/a-gentle-introduction-to-the-discussion-on-algorithmic-fairness-740bbb469b6.
2. Panch, Trishan, Heather Mattie, and Rifat Atun. "Artificial Intelligence and Algorithmic Bias: Implications for Health Systems." *Journal of Global Health* 9, no. 2. Accessed October 18, 2020. https://doi.org/10.7189/jogh.09.020318.
3. Pessach, Dana, and Erez Shmueli. "Algorithmic Fairness." *ArXiv:2001.09784 [Cs, Stat]*, January 21, 2020. http://arxiv.org/abs/2001.09784.
4. Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. "The Ethics of Algorithms: Mapping the Debate." *Big Data & Society* 3, no. 2 (December 2016): 205395171667967. https://doi.org/10.1177/2053951716679679.
5. Zhong, Ziyuan. "A Tutorial on Fairness in Machine Learning." Medium, June 19, 2020. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb.
6. Corbett-Davies, Sam, and Sharad Goel. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning," n.d., 25.
7. "The Case for Causal AI (SSIR)." Accessed November 16, 2020. https://ssir.org/articles/entry/the_case_for_causal_ai.
8. Annette Zimmermann, Elena Di Rosa. "Technology Can't Fix Algorithmic Injustice." Text. Boston Review, December 12, 2019. http://bostonreview.net/science-nature-politics/annette-zimmermann-elena-di-rosa-hochan-kim-technology-cant-fix-algorithmic.
9. Kun, Jeremy. "One Definition of Algorithmic Fairness: Statistical Parity." *Math ∩ Programming* (blog), October 19, 2015. https://jeremykun.com/2015/10/19/one-definition-of-algorithmic-fairness-statistical-parity.
10. Bornstein, Brian H, and Hannah Dietrich. "Fair Procedures, Yes. But We Dare Not Lose Sight of Fair Outcomes," n.d., 6.
11. Rawls, John. "Justice as Fairness." *The Philosophical Review* 67, no. 2 (1958): 164–94. https://doi.org/10.2307/2182612.
12. Wenar, Leif. "John Rawls." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Spring 2017. Metaphysics Research Lab, Stanford University, 2017. https://plato.stanford.edu/archives/spr2017/entries/rawls/.