Complex Equality and Algorithmic Fairness: A Social Goods Approach to Make Statistical Fairness Metrics Less Abstract

Bauke Wielinga [1]

[1] *TU Delft, Jaffalaan 5, Delft, 2628BX, Netherlands*

Abstract

I argue that the theory of complex equality, through its focus on social goods, can help adress certain problems caused by the abstractness of statistical algorithmic fairness measures.

Keywords

Fairness, Distributive Justice, Algorithmic Fairness, Complex Equality

1. Introduction

One of the main worries in both public and academic (see for example [1]) discourse regarding algorithmic decision making systems is about fairness. Many cases, for example in healthcare [2], have been documented where decisions of algorithms have had disproportionate negative impact on marginalized groups. In order to prevent such unfair outcomes, statistical fairness metrics have been developed which aim to ensure fair treatment between groups based on the rates at which an algorithm (accurately) categorizes each group [3]. Statistical fairness metrics have been widely critiqued [4, 5, 6, 7] primarily for oversimplifying the notion of fairness in the attempt to reduce it to equal probabilities of one kind or another. In doing so, it is argued, these metrics miss much of the context of processes being described [4] and their simplifications can inflict serious harm [8].

Such criticisms have been nicely captured by [9] in a number of "abstraction pitfalls": flaws of statistical fairness metrics caused by the way they abstract from algorithmic fairness situations. This paper will deal with three of these pitfalls. Firstly, statistical fairness metrics presuppose a framing of situations which can obscure fairness-relevant contextual considerations (the framing trap). Secondly, this lack of contextual information can wrongly lead to the conclusion that an algorithm that was fair in one context, will also be fair in another (the portability trap). Finally, the statistical fairness metrics represent an abstract mathematical conception of fairness, which ignores aspects of fairness which are unquantifiable or outcome-independent (the formalism trap).

This paper will work towards addressing these problems by construing fairness through the lens of Michael Walzer's notion of *complex equality* [10]. According to Walzer, different social goods, such as money, health care and political power, are each associated with their own distributive sphere, and their own criteria for how they ought to be distributed. Insofar as these criteria ask that each good *not* be distributed based on how much one has of another good, justice then requires the autonomy of different spheres of justice. I argue that applying complex equality to algorithmic fairness situations involves answering certain questions about the social goods involved in that situation, which lead to a less abstract understanding of the situation, and can help make a more informed choice of statistical fairness metrics.

Section 2. explains the theory of complex equality, and how to apply it to algorithmic fairness situations. Section 3. then argues that doing so can help to avoid the abstraction pitfalls of framing, portability and formalism. Section 4. summarizes these findings and reflects on the limitations of this approach to algorithmic fairness.

EWAF'23: European Workshop on Algorithmic Fairness, June 07–09, 2023, Winterthur, Switzerland
EMAIL: b.wielinga@tudelft.nl
ORCID: 0009-0008-5128-808X (A. 1)

2. Complex Equality and Spheres of Justice

Complex equality is an approach to justice which takes as a starting point the idea that there are many different social goods, and that just distribution can mean different things for different kinds of goods. A social good is something which is valued, not by one person, but socially, by a community that has a shared understanding of what that good is [10]. According to Walzer, what the just way to distribute any social good is depends on how the community understands that good, on what that good means for them. For example, the reason that health care should in a given community (say, the Netherlands) be distributed according to need would be that health care is understood in that community as something that is needed.

This relies of course on the assumption that there is a truly common understanding of that social good in the community in question. As Walzer takes the political community to be the place where distributive justice mainly takes place, it is certainly doubtful whether such shared understandings exist for many goods in large, diverse societies. I will not address this issue here, but will note that the approach as I outline it is compatible with other accounts of what appropriate distributive criteria are. Thus, I will focus on complex equality's emphasis on social goods, and aim to show how it can contribute to choosing a fairness measure. I leave the question of how to choose distributive criteria aside for other work.

Walzer's claim, then, is that it is often unfair to distribute one social good on the basis of another, e.g. to give me better healthcare because I happen to be very wealthy. Wealth is not part of the distributive criterion of healthcare, which, I have assumed, is need, and so this transaction is not permitted. Such disregard for distributive criteria is what Walzer calls *tyranny*: when money determines how much healthcare you get, this is tyranny of money over healthcare. When one good tyrannically determines the distribution of *many* different social goods, it becomes what Walzer calls a *dominant good*.

Walzer expresses his principle against tyranny formally as follows: "Complex equality means that no citizen's standing in one sphere or with regard to one social good can be undercut by his standing in some other sphere, with regard to some other social good". Walzer uses the term spheres here because, associated with each social good, there is a distributive sphere, a part of society in which the distribution of that good takes place. For example, associated with the social good of political power is the sphere of politics, which encompasses the government, elections, ministries and all other processes through which political power is distributed. Tyranny can then also be conceived as unwarranted influence between spheres.

It is not the case that complex equality forbids all distributive influence among different social goods however. Insofar as the social meanings of goods are not entirely distinct, or there is some logical connection between distributions of certain goods, dependence between them can be permissible. An example might be the distribution of jobs based on education. The autonomy of different distributive spheres is therefore only *relative autonomy*. This relative autonomy can inform our choices with respect to algorithmic fairness, as I show next.

2.1. Complex Equality in Algorithmic Fairness Situations

In order to apply complex equality to algorithmic fairness situations one has to answer the following questions:

1. What social goods are being distributed?
2. What are the distributive criteria of the involved social goods?
3. What distributive spheres do (a) the distributors and (b) the potential recipients represent?

Answering these questions gives a set of preconditions for fairness, which can guide the selection of a fairness metric for an algorithm that has not yet been deployed. When judging the fairness of an existing algorithm, question 4 should also be answered, to compare the optimal situation stemming from questions 1-3 with the actual selection criteria implied by the operation of the algorithm:

4. What distributive criterion does the algorithm implement for the relevant social goods?

The first two questions are included, because in order to assess whether any impermissible transactions are taking place we need to know what social goods are being distributed, and how their

distribution ought to go. Note that often multiple social goods are being distributed at once. For example, a fraud detection algorithm might distribute not only punishment, but also emotional harm, suspicion and money. We need some idea of how those goods ought to be distributed, before we can say whether they are being distributed fairly.

I include the third question for two reasons: First, complex equality can evoke intuitions regarding the sorts of goods that ought and ought not to be distributed by parties representing specific social spheres. Second, it is necessary to assess what sorts and quantities of goods the recipients in general have. The algorithm might select people based on appropriate seeming variables, while being applied to a population that is predominantly composed of a particular ethnic group or social class, as in cases where fraud detecting algorithms are applied only in specific poor neighborhoods [11]. Hence having an idea of the social goods of the potential recipients is important to assessing the fairness of an algorithm.

3. Complex Equality and Statistical Fairness

To illustrate how these questions relate to and improve upon statistical fairness, I will take as examples the demographic parity and equalized odds criteria, as discussed for example in [3], and asses what aspects of an algorithmic fairness situation their formulas capture compared to the questions of complex equality above. As an example situation, I will take an algorithm allocating unemployment benefits, of which we are solely considering whether it unfairly disadvantages people with a non-university educational background rather than a university education, by flagging them more often for fraud inspections.

Demographic parity demands that members of a marginalized and non-marginalized group have equal probability of being assigned to a positive predictive class:

$$P(R = + \mid A = a) = P(R = + \mid A = b)\ \forall a, b \in A$$

Where R is the prediction, which can either be + (positive, do not flag) or – (negative, do flag), A is the population, divided into groups a and b and where P is the probability. Thus the probability of a positive prediction, given membership of group a, must equal the probability of a positive prediction given membership of group b. Demographic parity looks only at the distribution of the algorithm's *outputs* between the different groups. Equalized odds on the other hand, also takes accuracy into account. It demands that members of a marginalized and non-marginalized group have equal rates of being truly as well as falsely classified:

$$P(R = + \mid Y = y,\ A = a) = P(R = + \mid Y = y,\ A = b)\quad y \in \{+, -\}\quad \forall a, b \in A$$

Where Y represents true class membership (of which y is a specific option), which R predicts. Thus the probability of an accurate positive prediction should be equal among different groups, according to equalized odds. Both metrics define two probabilities P, which, they state, should be equal.

Question 1: What social goods are distributed? In our example, the answer is only "social security", since unemployment benefits are a kind of social security. Question 1 asks what R recommends (the positive recommendation may itself be a social good) *and* what the distributive consequences of this prediction will be. Where R suggests a binary classification (flag, do not flag) as well as a binary evaluation (positive, negative), the answer to question 1 can involve multiple social goods, and suggests assessing outcome desirability in terms social goods distribution. The answer will depend heavily on the socio-technical and societal context in which the system is embedded, which R abstracts away from: R may flag someone for inspection, but the distributive consequences of the flagging may differ for different social goods depending on what procedures are followed in the institution around the algorithm. Hence, question 1 is able to capture more fairness-relevant considerations by considering multiple goods (counteracting the framing trap), yields more context-dependent answers, and gives a principled way to think about outcome desirability.

Question 2: What are the distributive criteria of the involved social goods? Question 2 enquires into the distributive criteria of the distributed social goods. For the distribution of social security (welfare in [10]) this is, I assume, need. It is the kind of social good which should be given to those who need it. For other social goods, the distributive criteria may be less determinate. And, indeed, it would be more difficult still to spell out when someone is in need, and what goods count as social security. These distributive criteria are requirements for algorithmic fairness, independent of

statistical fairness metrics, and can function as a limitation on the choice a statistical fairness measure. If the distributive criterion of social security is need, then tyranny and all kinds of discriminatory distribution are ruled out. Need also demands a minimization of false positives (falsely flagging for inspection), since wrongfully withholding a needed good is a very serious harm. This implies that fairness in the distribution of needed goods depends not only on equal distributions of predictions or accuracy, but also demands accuracy with regard to predicted negative predictions regardless of group membership. Insofar as far as not all fairness-demands can be met, need prioritizes minimization of false positives. As an approximation, false positive (or negative) error rate balance, as discussed in [3] then seems the most appropriate statistical fairness metric for needed goods.

Because the demands of the distributive criteria of multiple social goods at once can lead to situations where assessing fairness is truly difficult and unclear, I argue that complex equality does not treat fairness as a mere formalism, thus also countering the formalism trap.

Question 3: What distributive spheres do (a) the distributors and (b) the potential recipients represent? In our example: the distributor represents the sphere of political power, and the recipients represent the sphere of education. Question 3 corresponds most closely to *A*, which represents the population, which is divided into groups *a* (university background) or *b* (non-university background). It assumes a cleanly divisible population, and a known marginalization attribute (i.e. type of education). Question 3b offers a limited way to think about the choice of marginalization attribute, namely in terms of social goods. It also invites a critical examination of the social goods of the target population as a whole: is the algorithm being applied to a population composed predominantly of the least or most well off of certain distributive spheres? 3a invites a similar examination of the owners of the algorithm: do they represent a sphere that ought not to distribute this type of good among this kind of population? For example, a government-run algorithm assigning people possible romantic partners might trigger such considerations. In this way, question 3 takes a broader scope (contra the framing trap), and encourages a more detailed description of the situation than statistical fairness metrics.

4. Conclusion and Limitations

I have argued that assessing algorithmic fairness situations through the lens of complex equality can help to alleviate some of the pitfalls of statistical fairness, argued by [9] to be caused by its abstract nature. To assess situations using complex equality, I have argued that it is necessary to identify what social goods are distributed, what distributive criteria are appropriate for those social goods, and what distributive spheres the distributors and recipients in the situation represent. Thoroughly answering these questions gives a concrete understanding of algorithmic fairness situation, draws out many considerations that statistical fairness metrics alone do not, and sets limitations on the choice of an appropriate fairness metric. It can help to prevent instances of the formalism, portability and framing traps for algorithmic fairness.

Specifically, I argue that, through questions 1 and 2, complex equality offers an approach to fairness that is not a mere formalism, but is rooted in the social goods that are being distributed. The context-sensitivity of all three questions means that it is difficult to step into the portability trap when the considerations of complex equality are held in mind. Finally, questions 1 and 3 provide a principled way to think about the framing of the problem in any algorithmic fairness situation, and so help to alleviate the framing trap.

A major limitation of this approach is that the ontology of spheres of justice does not represent all the marginalized groups that an account of fairness should forbid discrimination against. While the emphasis on social goods works well to identify class-based discrimination (tyrannies by the spheres of money, education, political power, etc.), it does not seem to make sense to conceptualize race, gender, ability, sexual orientation or ethnicity as social goods. The analogy with tyranny among social goods is easy to make: just as the possession of money should not be a criterion for the distribution of social security, neither should race, gender, ability, sexual orientation or ethnicity (except insofar as they determine need). Certainly such discrimination is against any plausible distributive criterion, and so ruled out by complex equality, but if these categories cannot be conceived as social goods, complex equality does not appropriately center them in its ontology. Walzer does not speak of a

sphere of gender for example, only one of family relations, in his analysis of the oppression of women.

If this last consideration implied that complex equality is too limited in scope, then the opposite is also a problem: the list of social goods is open-ended, and it may be difficult in practice to determine which social goods are distributed, and which are relevant in a given situation, especially with regard to question 3b: "What social goods do the recipients possess?". Everyone possesses all social goods to a degree; even if a group does not posses a social good at all, this may be a relevant consideration. Hence, it would be helpful to supplement complex equality with a theory that offers additional guidance in choosing through what social goods a situation should be analysed.

5. Acknowledgements

This research was funded through a collaboration between TU Delft and the Dutch Employee Insurance Agency UWV.

My deepest thanks go to my supervisors, Dr. Stefan Buijsman and Prof. Dr. Jeroen van den Hoven, without whose guidance, help and feedback I would not have been able to write this paper. I am also thankful to many individuals who gave feedback on earlier versions of this paper, at the Delft Digital Ethics Center, my faculty peer group and many other colleagues.

6. References

[1] A. Khan *et al.*, "Ethics of AI: A Systematic Literature Review of Principles and Challenges." arXiv, Sep. 12, 2021. doi: 10.48550/arXiv.2109.07906.

[2] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019, doi: https://doi-org.tudelft.idm.oclc.org/10.1126/science.aax2342.

[3] A. N. Carey and X. Wu, "The statistical fairness field guide: perspectives from social and formal sciences," *AI Ethics*, Jun. 2022, doi: 10.1007/s43681-022-00183-3.

[4] O. Keyes, J. Hutson, and M. Durbin, "A Mulching Proposal." arXiv, Aug. 10, 2019. doi: 10.48550/arXiv.1908.06166.

[5] L. Hu and I. Kohler-Hausmann, "What's Sex Got To Do With Fair Machine Learning?," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Jan. 2020, pp. 513–513. doi: 10.1145/3351095.3375674.

[6] A. Birhane, "The Limits of Fairness," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, in AIES '22. New York, NY, USA: Association for Computing Machinery, Jul. 2022, p. 2. doi: 10.1145/3514094.3539568.

[7] S. Benthall and B. D. Haynes, "Racial categories in machine learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Jan. 2019, pp. 289–298. doi: 10.1145/3287560.3287575.

[8] A. Hanna, E. Denton, A. Smart, and J. Smith-Loud, "Towards a critical race methodology in algorithmic fairness," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, in FAT* '20. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 501–512. doi: 10.1145/3351095.3372826.

[9] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, in FAT* '19. New York, NY, USA: Association for Computing Machinery, Jan. 2019, pp. 59–68. doi: 10.1145/3287560.3287598.

[10] M. Walzer, Spheres Of Justice: A Defense Of Pluralism And Equality. Basic Books, 2008.

[11] D. Davidson and S. Adriaens, "In arme wijken voorspelt de overheid nog altijd fraude," *VPRO*, Dec. 20, 2022. https://www.vpro.nl/argos/lees/onderwerpen/artikelen/2022/in-arme-wijken-voorspelt-de-overheid-nog-altijd-fraude.html (accessed Feb. 22, 2023).