Intersectional AI: Hacking Biases - Journal Club Philosophy and Ethics of Design and Technology

WS 20/21

Article: Fairness in Machine Learning: Lessons from Political Philosophy
Author: Reuben Binns
Publication Date: 2018

Prepared by:
Andreas Greiler Basaldua
Zeliha Zeynep Güçlükol
23.02.2021

Fairness in Machine Learning: Lessons from Political Philosophy
by Reuben Binns

This paper analyses the article "Fairness in Machine Learning: Lessons from Political Philosophy" written by Reuben Binns. Reuben Binns is an Associate Professor at the University of Oxford in the Department of Computer Science. The article's thesis is that lessons from political philosophy, especially with regards to discrimination and egalitarianism, can help inform fairness measures in machine learning models. To this end, the author carefully outlines two normative considerations from the realm of political philosophy - discrimination and egalitarianism - and mapps challenges in fair machine learning onto the rich history of related debates in political philosophy.

Introduction

In the introduction, the author explains common biases in machine learning and outlines some statistical metrics that are used to achieve fairer models. Formalizing and prioritizing definitions of fairness is crucial considering that satisfying multiple fairness metrics at the same time can be mathematically unfeasible. The author concludes the introduction by emphasizing that even some of the more nuanced statistical measures are often insufficient to significantly improve the fairness of ML models and thus continues into the realm of political philosophy, starting with one common definition of unfairness: discrimination.

2. What is discrimination, and what makes it wrong?

2.1 Mental State Accounts

Some argue that the cause of dicrimination in the real world is the mental state of the decision-maker. For example, it can be based on a bias regarding a group of people, such as in hiring a male candidate over a female candidate, which reflects some degree of systematic animosity or preference for or against one group by the decision-maker.

From the point of the mental state account, “the decision maker’s intent is key to discrimination” (p. 3). Yet, if we believe that the decision maker’s intent is critical for discrimination, can we call algorithms discriminatory even though they do not have human mental states such as contempt or animosity? There are different remarks from several researchers regarding this question. Some argue that data scientists, project managers, etc., should be responsible for any machines' discriminative decisions. Others support the concept of collective judgment, believing that the aggregated judgments of multiple individuals can result in moral responsibility and thus, be considered discriminatory.

The author mentions different arguments from different researchers. Still, the critical outcome of this section is explained by directly quoting from the text: “... mental state accounts of discrimination do not naturally transfer to the context of algorithmic decision-making” (p. 4) because algorithms do not have mental states.

2.2 Failing to treat people as individuals

Failing to treat people as individuals refers to discriminating against individuals based on generalizations made about groups of which they are members—for instance, hiring a non-smoker candidate over a smoker when non-smokers are considered more productive.

The author outlines some considerations regarding this sort of generalization. One argument is that "[s]tatistical discrimination is wrong, even if the generalizations involved have some truth to them because the generalization fails to treat the decision-subject as an individual" (p. 4). According to this

argument, however, a machine learning model is, by its very nature, discriminatory, since it relies on making generalizations based on group characteristics.

The second argument focuses not on the idea of generalization itself, but on the notion that generalization is only discriminatory if it relates to certain protected groups, such as gender, race or religion. The third argument is that the very ideal of "treating someone as an individual" is misconceived, since, every decision considering an individual is actually a disguised form of generalization. For example, let's consider the smoker - non-smoker example, again: if the employers want to have an alternative predictor for being productive, they could give some tests to the candidates. But even the decisions made based on those test results are likely to be generalizations based on scores.

In fact, criticisms of generalization are often rather criticisms of accuracy. There often are some more accurate decision data points or techniques available to decision-making, but using them is more expensive. Based on this tradeoff, the author posits that generalization can be morally acceptable when the number of people affected by the discrimination is justified by the high cost of a more accurate prediction technique. At this point, the author asks: “If the wrongness of algorithmic discrimination does not consist in the morally suspect intentions of decision-makers, or in fails to treat people as individuals, then what might it consist in?” (p. 5). The next section might offer some answers.

3. Egalitarianism

In this section, the author turns to the concept of egalitarianism. To begin, he provides two ideas for how to describe egalitarianism: 1. the idea that people should be treated equally, and 2., the idea that certain valuable things should be equally distributed. He focuses on how egalitarian norms might help explain why and when algorithmic systems can be considered unfair.

3.1. The currency of egalitarianism and spheres of justice

Machine learning systems often map individuals onto different outcome classes such as loan denial/approval, insurance prices, or the number of years spent in prison. These classes can be seen as means or barriers to some fundamentally valuable object(s) which should be more equally distributed (e.g. in the prison example, the object in question might be “freedom”).

But what is the general definition of this object that should be more equally distributed? In other words, what is the “currency” of egalitarianism? Drawing from political science literature, some things that are considered such "currencies" are: welfare (preference satisfaction), resources, capability, or equal political & democratic status.

However, even if one were to have a definition for the “currency” of egalitarianism, there are various considerations that make it difficult to define and attach value to this “currency”: depending on the context, some “currencies” of egalitarianism can be considered more or less important. For example, the distribution of loans might, in some cases, be considered more important than the distribution of rights to participate in an online discussion. Another complication is that different people may value the same outcome or set of outcomes differently. This is often not reflected in machine learning, which often assumes uniform valuation of decision outcomes across different populations.

Further, there might be “spheres of justice” in which different logics of fairness apply, and between which distributions might not be appropriate. For example, when voting, we think that all should have equal access to casting a vote that counts. But when it comes to taking a test for a job, we instead tend

to believe that the more qualified candidate deserves the job and the related economic benefits. Simply put, different contexts might require balancing different approaches to equality such as equality of outcome (in voting) and equality of opportunity (in testing).

3.2. Luck and desert

In this section, the author discusses views relating to the circumstances under which and the extent to which people should be held responsible for the unequal status they find themselves in. He introduces the idea of "luck egalitarianism": the position that inequalities which are the result of free choice and informed risk-taking are permissible, but not the inequalities that result from luck.

However, how does one clearly differentiate between inequalities that one deserves due to own choices vs. inequalities that result from luck? The author brings up the example of the US-American criminal recidivism risk scoring system that uses variables such as family, social circle, and neighborhood crime rates. But how can one tell whether these variables are the result of personal choice or luck? After all, one can either be born into a dangerous neighborhood (due to luck) or one can actively choose to move into one (due to choice) – yet the variable doesn't record which of the two scenarios actually applies. In short, this section illustrates how "luck egalitarianism" might serve as a good starting point for defining fairness but ultimately remains insufficient.

3.3 Deontic justice

This section discusses the idea of deontic justice, which is primarily concerned with understanding how some groups came to be unfairly disadvantaged in the first place in order to meaningfully evaluate whether a disparity is unfair. This train of thought necessarily leads us to consider the historical and sociological contexts, which brings its own complications.

One of the questions raised is the question of attribution of responsibility: who caused the disparity and who should be responsible for fixing it? Another question raised is that of redistribution across time: should inequalities caused by an institution in the past justify redistribution today? Both of these dilemmas are common in the context of systemic injustices and require careful consideration.

A third question raised is whether an instance can be considered worse because it exists in a broader context. One example of this is that racial profiling, which, due to its history, might be considered worse than other sorts of profiling.

3.4. Distributive versus representative harms

In this section, the author introduces the idea that some aspects of egalitarian fairness may not be distributive in a direct way. Rather, they may be about fair representation of different identities, cultures, etc. Some examples include: 1. states with multiple languages may have a duty to display things in all languages, 2. representation in digital cultural artifacts: search results, natural language classifiers, word embeddings might unjustly over- or underrepresent certain groups.

4. Conclusion

Current approaches to fair machine learning are often based on narrow protected classes derived from law and devoid of the necessary information to properly account for idiosyncratic lives and varying social contexts. Considering the related philosophical accounts of discrimination and egalitarianism can prompt reflection on more fundamental questions, and might ultimately provide guidance for fairness in machine learning beyond what is covered by existing interventions at the data preparation, model-learning, or post-processing stages.