Optimizing Algorithm Fairness

Charles Zhang, Chloe Zheng

PHIL 225/COMP 154 Digital Ethics

December 6, 2019

Introduction

In this paper, we are inspired by Pak-Hang Wong's article *Democratizing Algorithm Fairness*, and discuss how the algorithm fairness can be delivered. In answering the question "What counts as a fair algorithm", we believe that it can be divided into mainly two parts: 1) what is the definition of fairness in algorithms; 2) how can we make sure the algorithm fairness is fairly delivered. Agreeing with Wong, we recognize the need for different definitions of fairness for different algorithms. Thus there requires a discussion on deciding the fairness. Based on Wong's proposal of a framework to provide discussions, we elaborate on the necessary conditions to maintain for the framework to function fairly. Other from the political views, we also point out that solving technical obstacles is still crucial in developing algorithm fairness, and that opposite from Wong's view, there is optimistic future in improving technical issues in algorithm. Lastly, we bring out an additional aspect of accountability in the discussion, and briefly address the problem.

Background

Machine Learning Algorithms, also known as MLAs, are programs that adjust themselves to perform better as they are exposed to more data. MLAs work by taking historical instances, also called the training data, of a decision problem as input and produces a decision rule or classifier that is then used on future instances of the problem (Hardt 2014). A learning algorithm is designed to reflect on the training data, and tries to optimize the solution with the pattern it recognizes from the environment.

Algorithms are expected to be true, objective, and fair. Nevertheless, there are three factors that might cause its unfairness. The algorithm biases we currently face are caused by two main reasons: the mismatch between the design of the algorithm and its results, and the implied unfairness intrinsic to the design itself. We believe that for a fair design of the

algorithm, we not only need procedural justice, but also substantial justice. Algorithm designers need to ensure that both the design and the results of the algorithm are able to undertake the merits they believe to be fair. Despite that designers have made considerable effort into making the design of the algorithm fair, there is still controversy in the design. This is due to the fact that definition of fairness varies within different contexts and different uses. In Pak-Hang Wong's paper, he argues that the controversial debate on whether the risk assessment algorithm COMPAS is fair, is led by the disparate understandings of fairness between the corporation, the designer of the algorithm, and the research team, the representative of people who are assessed by the algorithm. The last, also the most underestimated factor of the cause of algorithm bias, is the social environment itself. According to the fundamental principle of how algorithm works, it is constantly adjusting itself with more expose to training data, which means it often replicates the pattern within the training data. The difference between assessing by algorithms and by human beings ourselves is that algorithms have full trust in the environment they are put in. They learns and reflects the environment without doubt, while we human beings would criticize the environment and try the change the part where we don't agree with our values. Thus one essential bias within the algorithm, is the bias we have not yet eliminated from our society. It will be an ultimate goal for both the development of MLAs, as well as for our own society.

Wong mentioned about the Impossibility Theorem in his paper: that it is impossible for an algorithm to maintain more than one specific definition of fairness. Thus as we recognize the diversity of definitions for fairness, it is impossible to develop a universal algorithm. However, to coordinate with the diversity of understandings of fairness, we suggest a diversity in the design of the MLAs as well. Algorithms are now applied for different uses: university admission, hiring, insurance, credit rating, criminal risk

assessments, etc. The fairness we pursue in algorithms is not definite: the calling for algorithm fairness is only reasonable when we tolerate the diversity of fairness definitions.

Framework for Algorithm Fairness Discussion

One of Wong's arguments is that it is not sufficient to see algorithm fairness as merely a technical problem. He points out the political aspect of improving algorithm fairness: to democratize the discussion of fairness definition, as well as the supervision of the process of running the algorithm in the system. We agree with Wong that the important political dimension of algorithm fairness requires debate, both public and within the industry, to address the competing values for the justice of algorithm. Wong proposes a framework, under which people are able to have conversations about values to put into the algorithm that will ultimately lead to the style of the design for the algorithm. In Wong's framework, he had four main points: 1) the decision-making process must be accessible to the public; 2) the design of algorithm should be explained in untechnical language and include a broad range of stakeholders affected by these decisions; 3) the algorithm must be open to revision and improvement in light of new evidence or arguments; and 4), there should be public regulation of the process to ensure the implementation.

We agree that the decision-making process should be publicly supervised. The underlying core of this condition is to guarantee transparency. A research institute from NYU finds that too many of the scoring systems that determine important events in life like granting bail, sentencing, enforcing and prioritizing services are opaque to the citizens they hold power over. These systems are referred to as "black box", which implies the lack of transparency. The research team points out that it can be possible to disclose information about systems and their performance without disclosing their code, their protected intellectual property. The transparency is crucial in the sense of fulfilling citizens' right to know the

systems they are in, as well as to ensure the fairness throughout the decision-making progress. Transparency also improves the fairness in the design in the algorithm itself as the public provides a different perspective to view the system. In 2016, Northpointe's tool for assessing criminal defendants' likelihood of becoming a recidivist, COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), is accused of posing biased effect on the risk assessment scores. Research team from ProPublica argues that the scoring system was biased against black people by taking races as one of the factors into the assessment. Governments are increasing rely on mathematical formulas to inform decisions on criminal justice, child welfare, education and other arenas. Yet it is impossible for citizens who are largely affected by these algorithms to see how the system works and are being used. Transparency is in demand to a great extent, for the fairness to decide the definition used for fairness, which will lead to a fair algorithm as well.

In the relevance condition, Wong argues that algorithm developers should provide a reasonable explanation of why priorities are selected under certain contexts, and the deliberation it should include a broad range of stakeholders affected by these decisions. The explanation for the algorithm will contribute to the transparency of the algorithm fairness, however, the people involved in the discussion of values to put into the algorithm should not be only the stakeholders. We see that addressing different demands from different stakeholders is one of the major reasons that lead to the need different interpretations of fairness for systems of different uses. Still, it will be unfair if people who are not contributing financially or intellectually but are affected by the algorithms are neglected and not invited to the conversation. For instance, most governmental used algorithms are inaccessible to the public: not that the public has no knowledge of the system at all, but they are unable to see or decide how they will be valued. In the COMPAS incident, if ProPublica never starts the research, the black victims who were misjudged by the system will never know they were

unfairly treated simply because the algorithm picked up the bias within its training data. Wong points out in his paper that Northpointe were having conflicts with ProPublica in arguing whether COMPAS was impartial because they were having different understandings for the definition of fairness. Although Northpointe refutes that they had justice, but just in a different kind, we believe that their algorithm as well as the results are still unfair because they failed in addressing a mutual fairness among themselves, relevant stakeholders, and most importantly, people who were assessed by the system. Therefore, we want to emphasize that only reaching to a consensus on the fairness definition among stakeholders is not sufficient: people who are affected, either users in the platform, or the ones who are being scored, should have a say in the decision making process, rather than just having information about what has been decided.

In consideration of the hallmark of the MLA, the training data should be constantly updated to the latest values as well as the social environment, to ensure the algorithm is able to function with the most advanced values. Despite that algorithm designers and cooperations aim to develop fair algorithms, we still doubt if the algorithm will ever reach to a perfect state of fairness, because the society is not completely unbiased itself. We as human beings, as designers of the algorithms, are still on our way to try to understand the definitions of fairness as well as the adequate measures to deliver our interpretations for fairness into the practice. The flexibility of the algorithms will help with its ability to be fair: this means the bias will be eliminated as we improve fairness within our society.

Back to the Technical Dimension

Although Wong promotes to solve the algorithm fairness problem with a shift of focus from the technical aspect to the political perspective, we believe that there is still considerable merits in improving the fairness through technical measures. Wong concludes

the inherent trade-off between fairness and accuracy in algorithms. For example, for those who value public safety, fairness measures that significantly reduce public safety will not be acceptable. In consideration of the balance between fairness and accuracy, different fairness measures can be understood as representing the interests of different stakeholders affected by an algorithm (Narayanan 2018). The trade-off between the accuracy and the fairness of the algorithm outcome is essential to the nature of algorithms, thus it is an unavoidable issue that AI experts will have to encounter.

Policymakers have demanded that high-stakes decision systems be designed and audited to ensure outcomes are equitable. The research community has responded to this challenge: they claim that with three mathematical definitions of fairness: anti-classification, classification parity, and calibration, algorithms are able to be improved in its accuracy to deliver fairness (Corbett-Davies, Sam and Sharad Goel 2018). It is recognized that enforcing anti-classification and classification parity can often harm the very groups that these measures were designed to protect. In mind of the fact that mathematical procedures often fail to directly address fairness, designers point out such issue can be improved by assessing the potential effects of the algorithm. Also, they believe that a more explicit focus on consequences is necessary to make progress. In response of the technical and political aspects of the algorithm fairness, researchers recommend decoupling the statistical problem from the policy problem of designing interventions. They point out that outcome from algorithms might be interpreted in different meanings between designers and judges, thus cause a misunderstanding in the accuracy of algorithms. Researchers also point out, in dividing different scenarios for the algorithms to apply, policymakers tend to have false decisions due to their lack of understanding towards algorithms.

From this research we conclude that: 1) algorithms designers are confident that there is considerable progress to make in addressing the accuracy of algorithms tackled with the

adequate concept of fairness; and 2) better communication between policymakers and designers should be developed. Accurate understanding towards what the outcome means in risk assessment will help with the practical use algorithms, as well as the elimination of misunderstanding and false accusations towards the inefficacy of algorithms.

Another Aspect: Accountability

On the base of Wong's argument, we would like to add a new perspective: the accountability for the fairness of the algorithm. There had been debates going on about who should be responsible for the bias of the algorithms. It seemed unreasonable to burden the designer with full responsibility of developing and implementing a fair algorithm, but attributing the responsibility towards the public will be meaningless. With a deeper understanding from the algorithmic fairness discussions, we find a middle ground to allocate the responsibilities.

In deciding the definition of fairness, algorithm designers or corporations who own and develop the algorithm should guarantee all members from the society who will be in part of the algorithms are involved in the conversation, as well as have the right to vote for their expectations for fairness. Once the definition of fairness is determined, algorithm designers should guarantee the values are accurately translated into the algorithms, and a presentation or explanation for how the final algorithm will work should also be in demand. Finally, owners of the algorithm should report the information and the recent outcome of the algorithms to guarantee that the algorithm is functioning in the way they claimed to be. As Wong suggested in his fourth condition, there should be public regulations to ensure the conditions are met. Thus our discussion of accountability will be constructive for setting up these regulations.

Conclusion

In our paper, we discussed how to develop a fair algorithm. We believe that the fairness in algorithms should be mainly in two aspects: the procedural justice for the design of the algorithm which determines how well it adapts the mutual value for fairness, and the substantial justice that improves the equivalence between the accuracy and the fairness for the outcome. On the basis of Wong's proposal of the framework for discussions on interpretations for fairness, we elaborate on why the conditions are important to deliver fairness into the procedure. We also point out that although the political dimension should be added into the implementation of algorithmic fairness, the technical aspect is just as important as it was before. Algorithm experts are working on the substantial justice for the algorithm fairness, and showed that it is possible for the accuracy to accurately reflect fairness as well. Last but not least, we propose an additional branch for the discussion: the accountability for the algorithmic fairness. We argue that with deeper understanding towards algorithm fairness, we are able to determine who are responsible in delivering algorithm fairness. We divided the responsibilities into several parts and distributed to different groups in the delivering process.

References

Corbett-Davies, Sam and Sharad Goel. “The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning.” *ArXiv* abs/1808.00023 (2018): n. pag.

Hardt, Moritz. "How big data is unfair: Understanding sources of unfairness in data driven decision making." Unpublished paper. (medium.com/@mrtz/how-big-datais-unfair-9aa544d739de) (2014).

Narayanan, Arvind. "fairness definitions and their politics.(Feb. 23 2018)." In Tutorial presented at the Conference on Fairness, Accountability, and Transparency. 21.