Discrimination in Algorithmic Decision Making

Isabel Valera
Department of Computer Science, Saarland University
Max Planck Institute for Intelligent Systems

ABSTRACT
As automated data analysis supplements, and even replaces, human supervision in decision making, there are growing societal concerns about potential unfairness of these systems. This article summarizes recent advances and challenges of fair machine learning. First, we discuss the design of automatic decision systems that incorporate a fairness definition in their training step to avoid discrimination towards particular groups of people sharing certain sensitive attributes, such as gender or race, while providing clear mechanisms to trade off fairness and accuracy. Then, we discuss some of the main limitations of existing approaches within the Fair ML community, stressing the need for interdisciplinary collaborations.

1. INTRODUCTION

Algorithmic decision-making processes are increasingly becoming automated and data-driven in both online (e.g., spam filtering, product personalization) as well as offline (e.g., pre-trial risk assessment, mortgage approvals) settings. However, as automated data analysis increasingly supplements, and even replaces, human supervision in decision making, there are growing concerns from civil organizations, governments, and researchers about potential unfairness of these algorithmic decision systems towards people from certain demographic groups (e.g., gender or ethnic groups).

To alleviate these concerns, a number of recent studies in the emerging field of fair machine learning (ML) have proposed and analyzed mechanisms to ensure that algorithmic decision systems do not lead to unfair outcomes, or perpetuate historic biases and harmful stereotypes. However, by simply taking a snapshot of media reports regarding algorithmic bias and discrimination, we can clearly observe that these concerns are still far from being alleviated.

As an example of algorithmic gender bias, we can easily find news articles discussing Amazon's AI recruitment tool, which was shown to be biased against women.[1] The recruitment tool was trained using the resumes submitted to Amazon over a 10-year period, where most of the applicants were white males. Using such data, the algorithm learnt to recognize word patterns in the resumes, which resulted in penalizations to resumes containing the word "women's", and thus, a bias against women.

Other examples of algorithmic gender bias include biased word associations. Researchers from Princeton University found that the words "women" and "girl" were more likely to be associated by ML systems with the arts instead of science and math, which were more likely connected to males.[2] Machine translation is one of the main applications of word associations, where gender biases have also been pointed out. In December 2018, Google claimed to have resolved gender bias in its neural machine translation tools.[3] However, the implemented solution has been recently revisited as it suffered from issues when scaling to a greater number of languages.[4] Unfortunately, machine translation is not the only application of word associations where algorithmic bias remains unresolved. If one performs a Google image search of words such as "nurse" or "secretary", most of the top images belong to women, while word searches such as "director" or "professor" result mostly in male images.

Importantly, algorithms have shown not only gender biases but other discriminative biases. For example, the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, which assists judges to make their decisions about pre-trial bail, was found to be biased against African-Americans, according to a report from ProPublica.[5] Similar issues have motivated the recent Facebook initiative to launch a new investigation into potential racial biases in its systems.[6]

In this article, we provide a summary of the main advances made by the research community on fair machine learning to address the above issues. Subsequently, we discuss some of the main technical challenges that explain why, despite the enormous effort made by both researchers and industry partners, algorithmic bias and discrimination remains a challenge.

---

1 *Hamilton*, 2018.
2 *Hadhazy*, 2017.
3 *Bond*, 2020.
4 *Bond*, 2020.
5 *Larson/Mattu/Kirchner/Angwin*, 2020.
6 Facebook to Launch New Investigation to Study Potential Algorithmic Bias In Its Systems, 2016.

Through the article, we focus on the widely studied problem of algorithmic consequential decision-making, where decisions may have long-term effects on the lives of individuals.[7] Examples of consequential decisions include pre-trial release decisions, loan approvals and fraud detection.

2. TECHNICAL ADVANCES IN FAIR MACHINE LEARNING

Literature on fair algorithmic decision making assume historical data that collect both non-sensitive information (e.g., education level, income, etc.) and sensitive information (e.g., gender or race) of individuals, as well as access to the values of the target quality the algorithm is trying to predict (e.g., whether they repaid the loan or not, or they reoffended, usually known in the ML literature as labels). Such information is often presented as tabular data, where each individual corresponds to a row of data, and the columns collect the features summarizing the individual's information. Additionally, each individual's label is represented as a binary output variable, which is only available during the training step of the ML algorithm. The goal of the algorithm is thus to provide an accurate prediction of the label, which will then be used by the decision maker (bank or law court) to inform its decision. Hence, the first question that arises is: How is it possible for predictions of machine learning algorithms to be biased?

In order to answer such a question, we should first bear in mind that the information that the algorithm "sees" about individuals is a set of features, which may be less informative or not as representative for individuals belonging to minority groups. Moreover, the amount of data collected from minority groups is often significantly smaller than for majority groups. For instance, in the example above of Amazon's recruiting tool, most of the resumes belonged to males (majority group), while female applicants (minority group) where not representative. As a consequence, a prediction algorithm solely trained to maximize expected accuracy (or to minimize expected loss) of the training data, will lead to higher prediction errors for the minority group, as the prediction error decreases as more data is collected. The algorithm makes its prediction by finding patterns in the data; for example, common feature patterns of individuals that repay the loan or do not reoffend. Unfortunately, when the collected data contains historical human biases or stereotypes (e.g., most Amazon engineers are white males), the algorithm may learn such biases and exploit them to

7 *Kilbertus/Gomez-Rodriguez/Schölkopf/Muandet/Valera*, 2020.

make predictions. Even if sensitive information is not used as an input to the algorithm, the algorithm is often able to find proxies for the sensitive information. For example, the zip code may act as a proxy for race, and the individual's height and weight act as proxies for gender.[8]

A great deal of work on fair ML has been done to avoid discrimination in predictive machine learning algorithms that have been trained using potentially biased data. Such work has focused on introducing: i) definitions of fairness that apply to different contexts where decisions are informed by algorithms; ii) measures that translate a definition of fairness into a quantifiable metric that can be evaluated using data; and, iii) mechanisms to enforce a given fairness measure in the predictions outputed by the algorithm.

2.2. *DEFINITIONS, MEASURES AND MECHANISMS FOR FAIR MACHINE LEARNING*

As previously mentioned, most of the work on fair machine learning has focused on a particular definition of fairness—or rather unfairness—namely, discrimination. According to the definition by Altman,[9] discrimination is to "wrongfully impose a relative disadvantage on persons based on their membership in a salient social group". Of course, the next question that arises here is: What does "wrongfully" mean in each application domain of algorithmic decision-making?

In this context, and as we discussed in an earlier paper,[10] three of the most popular definitions of discrimination used in the machine learning literature are as follows:

- *Disparate treatment* (or direct discrimination), which occurs if individuals are treated differently according to their sensitive attributes (with the rest of the non-sensitive attributes being shared). To avoid disparate treatment, one should not inquire about individuals' sensitive data ("fairness by unawareness"). While this approach is intuitively appealing, the sensitive features, as previously mentioned, may often be accurately predicted (reconstructed) from non-sensitive features (proxies).
- *Disparate impact* (or indirect discrimination), which occurs when the outcomes of decisions disproportionately benefit or hurt individuals

8 *Zarsky*, 2014.
9 *Altman*, 2015.
10 *Zafar/Valera/Gomez-Rodriguez/Gummadi*, 2019.

from subgroups sharing a particular sensitive feature value. Much of the recent work done on fair learning has focused on approaches to avoid various notions of disparate impact. Specifically, *demographic parity* demands that the proportion of people in each sensitive group receiving the positive prediction must be equal.

- *Disparate mistreatment*, which occurs when the algorithm achieves different classification accuracy (or conversely, error rate) for different social groups. A decision-making system suffers from disparate mistreatment if individual misclassification rates (e.g., false positive rate, false negative rate) are different for groups of people sharing different values of a sensitive feature. This notion has also been referred to as "equality of opportunity",[11] which equalizes the true positive rates across groups.

The above definitions have been translated into statistical measures of fairness, which depend only on the joint distribution of predictor outcome, protected attribute, and label, and can be quantified from the observed data. Other definitions and measures of fairness exist in the context of algorithmic decision making, which often focus on fairness at the individual level and rely on causal reasoning to better understand and mitigate unfairness in algorithmic decision-making.[12]

Most of the articles on fair machine learning have focused on providing mechanisms to ensure that the predictor is fair according to a given definition of fairness. The majority of them fall into three categories: pre-processing,[13] in-processing,[14] and post-processing.[15] Pre-processing approaches are often limited to disparate impact, as they do not account for the predictor when removing unfairness. In comparison, both in-processing and post-processing can handle disparate impact and disparate mistreatment notions. However, Woodworth et al. have shown that in-processing of a potentially unfair predictor often produces better results than post-processing.[16]

11 *Hardt/Price/Srebro*, 2016.
12 *Barocas/Hardt/Narayanan*, 2019.
13 *Feldman/Sorelle/Friedler/Moeller/Scheidegger/Venkatasubramanian*, 2015.
14 *Zafar/Valera/Gomez-Rodriguez/Gummadi*, 2019.
15 *Hardt/Price/Srebro*, 2016.
16 *Woodworth/Gunasekar/Ohannessian/Srebro*, 2017.

3. TECHNICAL CHALLENGES OF FAIR MACHINE LEARNING

The vast amount of work done on creating fair predictive models could suggest that the problem of fair ML learning is solved. While one may think that training the ML algorithm may resolve any unfairness coming from the bias in the data, this is unfortunately not the case. As discussed in the introduction, there are plenty of examples showing that achieving fair machine learning is not as easy as it may look like. In the following section, we discuss some of the technical aspects that explain why training fair ML algorithms is not so easy.

The first and most obvious reason is that, in general, there is a trade-off between the algorithm's accuracy and fairness.[17] Such a trade-off depends on the definition of fairness that applies to the considered scenario, the training data, and the ML model to be trained. Importantly, it is not possible to know a priori what this trade-off looks like. Only when the model is trained (with or without fairness constraints), can one evaluate both its accuracy and fairness. Moreover, up to the best of my knowledge, there are no general approaches to estimating the Pareto frontier between accuracy and fairness for a given family of ML models. As a consequence, most of the existing approaches aim to maximize accuracy, subject to fairness constraints that enforce, for example, demographic parity. Unfortunately, enforcing such constraints may lead to underwhelming performance (accuracy) and thus be unacceptable in terms of business objectives.[18] As a result, the algorithm unfairness level is often compromised in favor of accuracy, as an inaccurate algorithm may be disadvantageous for all the stakeholders involved. This opens up the question of what accuracy and fairness trade-off is socially, or even legally, acceptable.

Second, there is, in general, little agreement on what fairness definition is appropriate for each scenario.[19] Thus, the practitioner may either opt to select a unique definition of fairness to be enforced or, alternatively, choose to enforce several fairness definitions. However, there are several pieces of work, summarized in Barocas et al.,[20] which prove that it is impossible to simultaneously enforce two different definitions of fairness, and to avoid both disparate impact and mistreatment.[21] Such an impossibility shows the inherent trade-off between different fairness criteria, and

17 *Zafar/Valera/Gomez-Rodriguez/Gummadi*, 2019.
18 *Zafar/Valera/Gomez-Rodriguez/Gummadi*, 2019.
19 *McMahon*, 2010.
20 *Barocas/Hardt/Narayanan*, 2019.
21 *Barocas/Hardt/Narayanan*, 2019.

thus should discourage practitioners to impose a single fairness definition without additional checks on other fairness criteria. For example, it might be undesirable to equalize error rates across groups at the cost of amplifying the demographic disparity (i.e., the differences between positive rates) in the data.

Finally, it is important to check the validity of the assumptions made through the overall design of the ML algorithm; keeping in mind that such assumptions start with the data collection process. Observed features are indirect, noisy and potentially biased measurements of the "state of the world". Thus, it is vital to assess which features used as input for the algorithms may trigger unfairness issues,[22] as well as to consider the underlying causal structure in the data to incorporate causal pathways into fair training procedures.[23]

Additionally, one should check if there are feedback loops during the data collection process. Most of the existing work on fair ML assumes that the available data is a representative sample of the population, that is, it is an independent and identically distributed (i.i.d.) sample of the population. However, as we discussed in our recent work,[24] such an assumption does not hold true for many of the algorithmic consequential decision-making scenarios, where the individual's label is collected only when a favorable decision is made. For example, only if bail is approved, can we observe if the individual reoffends; similarly, only if a loan is approved, can we observe whether they repay the loan. As a consequence, the labeled data used to train predictive models often depend on the decisions taken, leading not only to suboptimal performance (accuracy) for the justice system (or bank), but also potentially amplifying the unfairness/biases that exist in the training data. As a consequence, the fair ML problem in such cases should be reformulated as a learning-to-decide task rather than as a learning-to-predict task. Unfortunately, the latter is usually an easier task than the former.

The above examples summarize some of the technical challenges of fair ML in the context of algorithmic decision-making systems. However, one may expect that similar issues arise when considering other application domains, such as the word association example discussed in the introduction. Remarkably, as we discuss in the next section, the challenges of a fair de-

22 *Grgic-Hlaca/Zafar/Gummadi/Weller*, 2016.
23 *Barocas/Hardt/Narayanan*, 2019.
24 *Kilbertus/Gomez-Rodriguez/Schölkopf/Muandet/Valera*, 2020.

ployment of ML algorithms in the real world go way beyond these technical aspects.

4. DISCUSSION

So far, we have seen that although significant advances have been made in the context of fair ML, we have only scratched the surface of the problem. In particular, we have discussed fairness in consequential algorithmic decision-making problems, and seen that even in such a simplistic scenario, there are significant technical issues that need to be addressed. For example, although we have used the terms "unfairness" and "discrimination" interchangeably in this article, it may be important to consider fairness concepts that go beyond discrimination. For example, China's Social Credit Scoring System penalizes video gamers,[25] which, as far as I am aware, are not recognized as a salient social group according to discrimination laws. However, one may still consider it unfair to penalize individuals based on how they spend their spare time. Other real-world scenarios—for example, word association problems such as translation and information recovery—would also require careful analysis in order to define and mitigate unfairness.

However, the analysis of every scenario where ML plays a role should not be performed by ML engineers and practitioners in isolation. As ML becomes ubiquitous in society, there should be societal agreement and legal policies that determine the scope in which algorithms can be applied, either independently or under human supervision. Similar to other fields, such as medicine, the ML community may need to consider making it obligatory to perform an ethical assessment during the overall design of a new ML-based system. However, in order for practitioners to be able to properly assess the ethical aspects of their systems we should first be able to provide them with guidelines and protocols to question and justify any assumption made during the process. Perhaps that would prevent issues like the one recently under debate in an open letter to the Springer Editorial Committee, where several AI researchers urged Springer not to publish a paper on algorithmic criminal risk prediction based on face images.[26] The fact that something seems technically possible to do does not mean that it is the right thing to do.

---

25 *Nittle*, 2018.
26 *Coalition for Critical Technology*, 2020.

REFERENCES

Altman, Andrew. (2015). Discrimination. In Edward Zalta (Ed.), *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*. Retrieved July 30, 2020, from: https://plato.stanford.edu/archives/win2016/entries/discrimination/.

Barocas, Solon, Moritz Hardt, and Arvind Narayanan. (2019). *Fairness and Machine Learning*. fairmlbook.org. Retrieved July 30, 2020, from: http://www.fairmlbook.org.

Bond, Esther. (2020, April 29). Google Fixes Gender Bias in Google Translate (Again). *slator*. Retrieved July 30, 2020, from: https://slator.com/machine-translation/google-fixes-gender-bias-in-google-translate-again/.

Coalition for Critical Technology. (2020, June 23). Abolish the #TechToPrisonPipeline. *Medium.* Retrieved July 30, 2020, from: https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16.

Facebook to Launch New Investigation to Study Potential Algorithmic Bias In Its Systems. (2020, July 23). *News 18*. Retrieved July 30, 2020, from: https://www.news18.com/news/tech/facebook-to-launch-new-investigation-to-study-potential-algorithmic-bias-in-its-systems-2729487.html.

Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. Sydney, Australia (259–268). Retrieved July 30, 2020, from: https://doi.org/10.1145/2783258.2783311.

Grgic-Hlaca, Nina, Muhammad Bilal Zafar, Krishna P. Gummadi, and Adrian Weller. (2016). The case for process fairness in learning: Feature selection for fair decision making. *NIPS Symposium on Machine Learning and the Law*. Vol. 1.

Hadhazy, Adam. (2017, April 18). Biased bots: Artificial-intelligence systems echo human prejudices. *Princeton University*. Retrieved July 30, 2020, from: https://www.princeton.edu/news/2017/04/18/biased-bots-artificial-intelligence-systems-echo-human-prejudices.

Hamilton, Isobel A. (2018, October 13). Why It's totally unsurprising that Amazon's recruitment AI was biased against women. *Business. Insider*. Retrieved July 30, 2020, from: https://www.businessinsider.com/amazon-ai-biased-against-women-no-surprise-sandra-wachter-2018-10.

Hardt, Moritz, Eric Price, and Nati Srebro. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*. Retrieved July 30, 2020, from: http://papers.neurips.cc/paper/6374-equality-of-opportunity-in-supervised-learning.pdf.

Kilbertus, Niki, Manuel Gomez Rodriguez, Bernhard Schölkopf, Krikamol Muandet, and Isabel Valera. (2020). Fair decisions despite imperfect predictions. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, in PMLR 108:277-287*. Retrieved July 30, 2020, from: http://proceedings.mlr.press/v108/kilbertus20a.html.

Larson, Jeff, Surya Mattu, Lauren Kirchner, and Julia Angwin. (2020, May 23). How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica*. Retrieved July 30, 2020, from: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

McMahon, Christopher. (2010). Disagreement about Fairness. *Philosophical Topics, 38*(2), 91–110.

Nittle, Nadra. (2018, November 2). Spend "frivolously" and be penalized under China's new social credit system. *Vox*. Retrieved July 30, 2020, from: https://www.vox.com/the-goods/2018/11/2/18057450/china-social-credit-score-spend-frivolously-video-games.

Woodworth, Blake, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. (2017, July 7–10). Learning non-discriminatory predictors. *Proceedings of the 2017 Conference on Learning Theory, Volume 65 of Proceedings of Machine Learning Research, Amsterdam, Netherlands (1920–1953). PMLR.*

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. (2019). Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research,* 20:75. Retrieved July 30, 2020, from: http://jmlr.org/papers/volume20/18-262/18-262.pdf.

Zarsky, Tal. (2015). Understanding Discrimination in the Scored Society. *Washington Law Review*, *89*(4), 2014. Retrieved July 30, 2020, from: https://papers.ssrn.com/abstract=2550248.

ABOUT THE AUTHOR

Professor Isabel Valera

Isabel Valera is a full Professor at the Department of Computer Science of Saarland University in Saarbrücken, and an independent group leader at the Max Planck Institute for Intelligent Systems in Tübingen. Prior to this, she held a German Humboldt Post-Doctoral Fellowship, and a Minerva Fast Track Fellowship from the Max Planck Society. She obtained her PhD in 2014 and MSc degree in 2012 from the University Carlos III in Madrid (Spain), and worked as a postdoctoral researcher at the Max Planck Institute for Software Systems (Germany) and at the University of Cambridge (UK). She is an active researcher of the European Laboratory for Learning and Intelligent Systems (ELLIS – https://ellis.eu/en), involved in the ELLIS program on Robust Machine Learning, and a faculty member of the ELLIS institutes in Tübingen and Saarbrücken. She is area chair of the main conferences in machine learning (NeurIPS, ICML, AISTATS, AAAI and ICLR), and has been program chair of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases 2020 (ECML-PKDD 2020). Isabel's research focuses on developing machine learning methods that are flexible, robust, interpretable and

fair. Flexible means they are capable of modeling complex real-world data, which are often heterogeneous in nature and collected over time. The group's research also aims to improve the robustness of algorithms. An algorithm is considered robust when it is able to point out "what it does not know". Finally, the group is researching ways to make algorithms interpretable and fair; if they are part of important decision-making processes, the outcomes should be explainable and fair.

ABOUT THE INSTITUTE

Max Planck Institute for Intelligent Systems, CPTS/Probabilistic Learning Group

The research presented in this article is on the topic of fair machine learning (ML), which aims to ensure that the outcomes of algorithmic decision-making system do not discriminate individuals based on their membership of a salient social group; for example, based on their gender. The discussion presented in this article summarized the research on fair ML carried out by the Probabilistic Learning Group led by Isabel Valera at the Max Planck Institute for Intelligent Systems (MPI-IS), in Tübingen. The MPI-IS covers AI related topics such as computer vision, robotics, control, the theory of intelligence, and machine learning, which is the main field related to this research. In addition, the Probabilistic Learning Group closely collaborates with the Social Computing Department at the Max Planck Institute for Software Systems, led by Krishna Gummadi, and researchers from the Cluster of Excellence of Tübingen on "Machine Learning for Science", in order to access expertise in the social and ethical aspects of artificial intelligence. In terms of potential future collaborations, this line of research would benefit greatly from collaborations with other Institutes, such as the Max Planck Institute for the Study of Societies and the Max Plank Institute for Innovation and Competition, as they complement the societal and legal aspects necessary for developing artificial intelligence that is aligned with the goals and values of our society.