Methods to Prevent Unfairness from Emerging in Machine Learning Algorithms

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Callie Hartzog
Spring 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

I. Introduction: *The Rise of Machine Learning*

Machine learning is quickly spreading to all industries as a solution to many complex problems. A surprising outcome of the COVID-19 pandemic was the accelerated shift of businesses towards artificial intelligence to handle the increased influx of online interactions between businesses and consumers. More than 55% of businesses poured their resources into accelerating their artificial intelligence programs (McKendrick, 2021, Introduction). These are just the beginnings of a new level of artificial intelligence in society. The past has been spent programming computers to do what we want. The future lies in training these computers, whether they be household objects or sophisticated algorithms that can detect diseases (Heaven, 2021, Computer Knows Best). Machine learning is quickly becoming perceived as the more rational implementation of artificially intelligent systems. The most prevalent and recognizable use of machine learning is in the creation of image recognition algorithms. These algorithms are being used to do things such as identify criminals from photos, train self-driving cars to recognize pedestrians, and analyze potential of vision loss due to diabetes (Abrams, 2019, n.p.). Machine learning is built upon the existence of data. In order to train computer systems, engineers first need to analyze a data set and extrapolate meaning from the data. Collecting data, however, is typically done with some human involvement, which introduces bias. For example, facial recognition programs are trained mainly using white people as data points (Hardesty, 2018, Introduction). Despite limiting the training data to white people, the solutions these algorithms create are being applied to various populations around the world. This has led to a staggering decrease in the accuracy of facial recognition algorithms when applied to people of color. These algorithms can have heavy consequences such as identifying criminals or predicting diseases, which makes the accuracy of their predictions much more important.

II. Supporting Argument No. 1: Problem Definition: *Inconsistencies in Recognition Programs*

Machine learning algorithms are failing to uphold fairness in the results of their algorithms. For the purposes of this paper, unfairness will be referred to as bias when discussing machine learning algorithms. Unfairness is defined as a difference in treatment or equality between people. Within machine learning, this extends to how the results of an algorithm impact different people and cultures unequally, with results potentially favoring one culture over another (Mehrabi et al., 2021, p. 10).

Bias is a big issue in recognition programs, where the applications of these algorithms can have serious consequences if the results are inaccurate. Three main areas of application that have issues with bias are algorithms being used to identify criminals, loan application and determination algorithms, and medical imaging algorithms. In the United States, there are laws against which characteristics can and cannot be used for decision making. This is most commonly seen in anti-discrimination laws where employers cannot discriminate against employees or potential hires based on attributes such as race, gender, religion, disability status, etc. (Hentze & Tyus, 2021, n.p.). However, these laws often do not impact how algorithms are trained. This means that algorithms can potentially discriminate in their decision making without suffering legal consequences.

Facial recognition and prediction methods for things such as crime can inherently introduce racial profiling if programs are not trained on a widely representative data set that includes minorities. In addition to crime prediction, algorithms that analyze the viability of loan

applications lack ethical judgment and knowledge of systematic oppression that may have prevented certain populations from receiving loans in the past (Budds, 2017, Hard Questions and Hard Math Problems). Medical imaging algorithms suffer from similar consequences as facial recognition algorithms, where a lack of diverse ethnic representation causes inaccuracies in results. However, these algorithms can suffer when the algorithm is trained on data whose quality is not applicable to less technologically developed countries (Abrams, 2019, n.p.).

It is essential to find a balance between necessary bias in the data and the introduction of blatant racial bias. Bias can never be eliminated in machine learning; it is essential for the algorithms to learn trends and patterns. Removing all bias would prohibit the algorithm from knowing what it should be learning, and what data is irrelevant (Jaton, 2021, p. 3). This tasks engineers with finding a balance in their data that eliminates racial bias without compromising baseline results. The research seeks to understand the different types of bias present in various image recognition algorithms and the steps that can be taken to prevent bias.

III. Supporting Argument No. 2: Methods: *Marginalized Communities Impacted by Machine Learning Algorithms*

Mehrabi et al. (2021) write about several different types of bias that machine learning algorithms can exhibit. Algorithms that identify criminals are most susceptible to representation bias and measurement bias. Representation bias results from data that does not mirror the entire population being used to train the algorithm. Measurement bias stems from an inaccurate weighting of features that the data set portrays (p. 5). In regards to facial recognition algorithms, representation bias poses an issue when algorithms are trained on data sets consisting mainly of

white people. This allows for high accuracy of prediction on white faces, but low accuracy on other ethnicities. Hardesty (2018) tested the accuracy of prediction of different ethnicities using publicly available commercial facial recognition algorithms. They reported that the typical error of detecting the gender of an individual based on a provided image was only 0.8% for light-skinned males, but rose above 34% for darker-skinned females. Bias in this case arose from the fact that the data set was 83% white (Introduction). Inaccuracies in facial recognition can lead to wrongful arrests of innocent citizens when the suspect is not white. Measurement bias is found in many risk assessment algorithms, which predict the chances that a given suspect will go on to commit further crimes. The risk assessment system Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) suffers from many inaccuracies in predictions despite being used widely across the United States. With COMPAS, African Americans are twice as likely to be labeled as high risk for becoming repeat offenders despite never actually committing another crime. On the other hand, white people are twice as likely as African Americans to be labeled as low risk for becoming repeat offenders but continue on to commit more crimes (Angwin et al., 2016, n.p.). The unreliable predictions from COMPAS are a by-product of the algorithm using the criminal history of family members to help predict an individual's risk level. This feature was weighed heavily in the decision-making process, despite the fact that minority communities suffer from more arrests on average due to them being policed more than other communities (Mehrabi et al., 2021, p. 5).

Loan application and prediction algorithms typically suffer from aggregation bias and historical bias. Mehrabi et al. (2021) defines aggregation bias as inaccuracies resulting from generalizations about a population being applied to an individual. Historical bias is when social systems and pre-existing prejudices influence the data, either intentionally or unintentionally (p.

5-8). Aggregation bias is prominent in loan determinations and is frequently associated with the practice of redlining, where financial institutions withhold loans and other financial benefits to people based on where they live. When deciding on loan amounts, algorithms often consider the zip code that an applicant resides within. Certain zip codes may be marked as undesirable due to the general population that lives there (Rice, 1996, p. 613). Generalizations about the overall population of an area can impact the ability of an individual application to receive a loan. Another common, but troubling, issue with machine learning algorithms used for loan predictions is their ability to prolong systemic racism and racist practices. Prior to the Federal Fair Housing Act of 1968, money lenders in the United States were not prevented from discriminating against people based on race, religion, or sex when giving loans. This results in many years of data where people in these categories were given less compared to white men (Quillian & Honoré, 2020, p. 1-2). Despite the act, discrimination of all kinds has persisted for many years after, and still exists on some level today. Algorithms do not have the same awareness as people do that historically certain populations were systematically disadvantaged. The algorithm purely sees a pattern of certain people receiving less money and replicates it, directly creating historical bias.

Medical imaging suffers extensively from population bias. Population bias occurs when the data used to train an algorithm represents an entirely different population from the one the algorithm will be applied towards (Mehrabi et al., 2021, p. 8). Google's Automated Retinal Disease Assessment tool is a recently developed machine learning-based and data driven program that uses images to identify early signs of vision loss in diabetic patients. In the United States it produced promising results, but it fails to work properly in less technologically advanced countries such as India. When testing their program in India, Google failed to

accommodate for lower resolution images, which are a common occurrence in countries whose technology is not as developed as the United States (Abrams, 2019, n.p.). Google failed to account for all of the populations that their solution would be given to, and instead focused on a population representative of themselves which impacted the accuracy of the algorithm.

IV. Supporting Argument No. 3: Results: *Changing Approaches to Combat Bias in Machine Learning*

There is a lot of talk about the existence of bias in machine learning, but few discussions on how to actually prevent bias. The issue is that there is not a fix-all solution to eliminate bias entirely from all machine learning algorithms. Machine learning algorithms operate using various weights and parameters, which help them determine what data is important and what is not. If the parameters are adjusted to improve the performance of the program by making it produce a more general approximation, it is possible that the complexity of the program may be compromised, which could remove important patterns found in a minority of the data (Veale & Binns, 2017, p. 2-3). This creates a complex balance between ensuring the algorithm has enough bias to create an accurate prediction without infringing on the rights and representation of minority groups.

In order to create more accurate and fair algorithms, a variety of solutions need to be applied depending on the scenario. With representation bias, as discussed in the first example of algorithms used to identify criminals, a relatively simple solution exists. Training these algorithms on a wider data set that equally represents all ethnicities would greatly increase the

accuracy of the algorithm. Eliminating measurement and aggregation bias is more difficult. A detailed analysis of the factors and variables being used to train a system is necessary. Human evaluation is the only way to determine which variables are ethical to include when training a system and which are not. Computers cannot do this on their own as they have no inherent sense of morals. The solution to fixing historical bias is relatively easy to implement, but difficult in practice to adopt. The purpose of machine learning is to find algorithms that can optimize the work of humans, but in order to combat historical bias, human review of the results an algorithm produces is required. Many engineers want to minimize human involvement once algorithms are trained, which thereby increases the chances of historical bias occurring. Much like with measurement and aggregation bias, humans can see patterns of systematic oppression and racial discrimination while computers can only see the data they are given and not the ethics or history behind it. Population bias can be resolved by ensuring that the data a system is trained on is representative of the entire population and not just a subgroup of it.

V. Conclusion: *The Future of Machine Learning*

Machine learning is a promising field as society looks towards automation as a solution for many jobs and practices beyond human capability. An often-overlooked issue with machine learning is the presence of racial bias in algorithms. Minority communities are regularly either forgotten or ignored in machine learning algorithms. This can further racial bias already present in a society when these programs are used to make generalizations about a population they do not represent. It is essential that when using these algorithms, we remember that their predictions

may not be representative of the entire population. Bias will always be present, but it is the duty of engineers to ensure that this bias creates fair predictions and does not infringe on the rights of any individual.

VI. References:

Abrams, C. (2019, January 26). Google's effort to prevent blindness shows AI challenges. *The Wall Street Journal.* https://on.wsj.com/39tnDLo

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016) Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica.* https://www.propublica.org/article/machine-biasrisk-assessments-in-criminal-sentencing

Budds, D. (2017, July 25). Biased AI is a threat to civil liberties. The ACLU has a plan to fix it. *Fast Company.* https://bit.ly/3zAAeae

Hardesty, L. (2018, February 11). Study finds gender and skin-type bias in commercial artificial-intelligence systems. *MIT News.* https://bit.ly/3lVW4R4

Heaven, W. D. (2021, October 22). How AI is reinventing what computers are. *MIT Technology Review.* https://bit.ly/3m92Uni

Hentze, I., & Tyus, R. (2021, August). *Discrimination and harassment in the workplace*. National Conference of State Legislatures. https://www.ncsl.org/research/labor-and-employment/employment-discrimination.aspx

Jaton, F. (2021). Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. *Big Data & Society, 8*(1), 1-15. https://doi.org/g3mn

McKendrick, J. (2021, September 27). AI adoption skyrocketed over the last 18 months. *Harvard Business Review.* https://bit.ly/30N8EuK

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), Article 115 1-35. https://doi.org/10.1145/3457607

Metz, C. (2021, March 15). Who is making sure the A.I. machines aren't racist?. *The New York Times.* https://nyti.ms/3jztsfP

Quillian, L., Lee, J. J., & Honoré, B. (2020). Racial Discrimination in the U.S. Housing and Mortgage Lending Markets: A Quantitative Review of Trends, 1976–2016. *Race and Social Problems*, *12*(1), 13-28. https://doi.org/10.1007/s12552-019-09276-x

Rice, W. E. (1996) Race, gender, redlining, and the discriminatory access to loans, credit, and insurance: An historical and empirical analysis of consumers who sued lenders and

insurers in federal and state courts, 1950-1995. *San Diego Law Review, 33*(2), 583-700. https://advance-lexis-com.proxy01.its.virginia.edu/api/document?collection=analytical-materials&id=urn:contentItem:3S3V-1770-00CW-F0SD-00000-01&context=1516831.

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2), 1-17. https://doi.org/gdcfnz