Prospectus

Analysis of Tools for the Detection and Mitigation of Algorithmic Bias

(Technical Topic)

Mitigating Algorithmic Bias in United States

(STS Topic)

By

Caroline Hickey

November 5, 2020

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.

Signed: Caroline Hickey

Technical Advisor: Aaron Bloomfield

STS Advisor: Sean Ferguson

Introduction

It has been 70 years since the term artificial intelligence was first introduced, and in those 70 years, AI has become an integral part of society. It's uses are wide spread such as medical devices, chat bots, predictive policing, and facial recognition to name a few. As products were developed using machine learning algorithms, a new issue soon came to light: the bias that appeared in these algorithms. These biases have appeared everywhere from Facebook advertisements to the court sentencing. Some instances of algorithmic bias, such as advertisements for higher paying jobs appearing on men's Facebook advertisements, are incredibly problematic, others, such as the algorithmic bias that has been used in US courts for guiding sentencing to determine the likelihood that defendants would commit another crime, could ruin and potentially may have ruined lives. These algorithms, especially those used in situations where bias could be life altering, need to be regulated, reevaluated, and taken with a grain of salt. For my technical project, I will be researching the development of a software that could determine if an algorithm has bias, if a dataset that an algorithm is being trained on is diverse enough, and any software that can mitigate algorithmic bias. As of now, we are aware that algorithmic bias exists, but we need to find out which algorithms are, so that we can fix them. Once we can figure out which algorithms need to be fixed, we then need to figure out how exactly we should approach the solution to the bias problem. For my STS portion, I will be researching what is currently being done to mitigate algorithmic bias on a societal level, specifically within the government and within various community programs. I will be looking into legislation and organizations that are doing just that. "There is rarely a technical solution to social and political problems. Technology might be helpful in many ways, but only if its interrelations with social and material circumstances are part of the picture" (Sætnan, 2018).

Continuing from my technical portion which focuses on the technical solutions to algorithmic bias, my STS portion will look into what society can do to mitigate algorithmic bias.

Technical Topic

For my technical project, I plan on analyzing existing research on how to mitigate algorithmic bias through technical means. I have seen a lot of research highlighting the problems artificial intelligence has created, mostly the bias it has caused. In these papers calling attention to algorithmic bias, I have not seen much about trying to create or find tools to fix these biases. Some of the research that I will present attempts to fix algorithmic bias by training the algorithms on unbiased datasets, running algorithms to detect bias, and creating open toolkits for easy detection and mitigation. Many organizations have determined algorithms as biased after the deployment and wide use of the product using the biased algorithms, but I plan to research tools that should be used before the deployment of a product. With my research, I will consolidate the findings of other researchers who have found ways to fix algorithmic biases through various software and algorithms.

While bias is present in many algorithms, it is not impossible to remove such bias. In fact, "such algorithmic biases are new kinds of bugs, specific to modern, data-driven applications, that programmers should proactively check for, debug, and fix with the same rigor as they apply to other security and privacy bugs" (Tramer et al., 2017). Regulation of these kinds of algorithms should be an integral part of the development lifecycle, and tools such as the ones that I am researching should help.

Many researchers and software developers have created algorithms and products to combat algorithmic bias. A certain approach proposed by many is simply just to remove sensitive attributes, such as race, from the training dataset in order to avoid racial bias. This

approach does not work, however, because many other attributes, such as income or education level, can be highly correlated with race(Calders et al., 2010). One team developed a new approach for discrimination free classification of datasets using naive bayes approaches for discrimination-free classification by adding probablity to discriminated sentistuve atributes and decreasing probability of favored sensitive attributes(Calders et al., 2010). Another approach has been from Aequitas, a company that is tackling the problem of biased datasets by providing an auditing tool for developers to run their datasets to determine if the datasets will cause bias if used as training sets for their algorithms (Stevens et al., 2018). One of the most comprehensive approaches to mitigating algorithmic biases through the development of software has been IBM's AI Fairness 360 toolkit. This open source toolkit provides tools to detect bias in algorithms as well as a wide range of algorithms to mitigate the biases found (Bellamy et al., 2019). These tools provide easy access to lessen the bias in applications and create fairer algorithms.

These new tools may be able to solve algorithmic bias, but they pose their own set of problems. Ideally, these tools should be able to catch these biases with no issues. A problem could be if the tool itself becomes biased. It is nearly impossible to determine true neutrality, so any determination of bias versus unbiased will be very difficult. Another potential problem is privacy. In order to produce these tools, large amounts of personal data are needed. This has caused privacy issues that are difficult to solve due to the lack of legislation and regulations on artificial intelligence and data privacy, especially within the United States.

Over the course of this technical project, I will attempt to find the best approach to mitigating algorithmic bias. I will look at what is already in place, what other research is in progress, and what still needs to be done in the future.

STS Topic

As a society, we have progressed to the point that artificial intelligence has become an integral part of everyday life. It has become such an important aspect of society that there is very little possibility of ever going back to a life without intelligent algorithms. While these algorithms have done a lot to progress society and make life better and easier for humans, it has also brought to light issues such as the bias that these algorithms produce. Most notably, a risk-assessment algorithm that predicts the likelihood that defendants will commit another crime, COMPAS, has been used by United States courts as a factor to decide sentencing (Desmarais et. al, 2016). Despite its wide use in courts for many years, COMPAS was found to be biased against Black people, giving higher risk scores to Black individuals with the same profile as White individuals. This is just one example of the bias produced by unfair algorithms. While it might not be possible to completely eradicate the bias caused by algorithms, through legislation and community program efforts, we may be able to reduce bias and the implications that come with them.

Artificial Intelligence is widely used in the United States; in the US court systems, in banking systems, and in nearly every facet of life. Despite its use and promotion, there is little to no policy on artificial intelligence and how it should or shouldn't be used. In 2019, the AI in Government Act was put into place in order to promote the use of artificial intelligence within the government (H.R. 2575, 2019). In terms of regulation, the Algorithmic Accountability Act was introduced in 2019. It states that large companies that control and collect large amounts of personal data must conduct impact assessments in order to determine if any of their systems have a high risk of showing bias (S. 1108, 2019). The issue with this act is that it does not require companies to make these impact assessments available to the public, defeating the purpose of

making companies accountable for their algorithmic biases. In recent years, government funding of the development and research of artificial intelligence has been proposed, potentially providing $6.5 billion. Part of this funding is supposed to go towards regulating and creating standards for artificial intelligence (S. 1558, 2019) . Acts like this will help decrease algorithmic bias, but it is necessary to first create these standards before the government can continue in the AI race. “Governments across the globe should agree on global principles and regulations for AI systems that also incorporate the prevailing standards of democracy and human rights”(Wirtz et al., 2019). Even though in recent years more policy regarding artificial intelligence has been introduced, there are very few that have been actually passed and enacted.

Since there is little policy or action within the US government to mitigate bias in artificial intelligence, the US government should look to community projects that have been successful at mitigating algorithmic bias. The Algorithmic Justice League is an organization that has focused on decreasing bias in facial recognition. This organization provides research surrounding algorithmic bias, an auditing service for companies to submit their algorithms to ensure there is no bias, and is actively trying to increase policy on artificial intelligence (Artificial Justice League, n.d.). These organizations not only help mitigate bias through their services, they have increased rhetoric surrounding bias in AI that has created the awareness of the need for regulation and change. Organizations such as AJL play an important role in changing the way we approach artificial intelligence.

Bias exists everywhere, and while there is algorithmic bias, we must remember that this bias is a direct result of human bias. “Machines are constructed by humans, with a certain social setting in mind and manifesting an objectivity and significance of numbers that often is not warranted” (Sætnan, 2018). Even though bias may be near impossible to completely erase

because true neutrality is near impossible to achieve, there are steps that must be taken to mitigate this bias. Society has progressed to a point where artificial intelligence is indispensable, so ensuring that algorithms that dictate parts of our lives are fair is an integral part of the growth of AI.

Next Steps

Next spring, I will begin the technical portion of my thesis as well as the STS portion of my thesis. For my technical project, I will research various bias detection and mitigation algorithms to be used on datasets and software that uses machine learning. I will explore the strengths and weaknesses of various software and try to determine what else still needs to be done. For my STS portion, I will continue my in-depth research on ways to mitigate algorithmic bias through government intervention and organization programs, focusing more specifically within the United States.

References

Artificial Intelligence Initiative Act, S. 1558, 116th Cong. (2019).

AI in Government Act, H.R. 2575, 116th Cong. (2019).

Algorithmic Accountability Act, S. 1108, 116th Cong (2019).

Algorithmic Justice League - Unmasking AI harms and biases. (n.d.). Retrieved November 05, 2020, from https://www.ajl.org/

Bellamy, R. K., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., . . . Mehta, S. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development, 63(4/5). doi:10.1147/jrd.2019.2942287

Calders, T., Verwer, S. Three naive Bayes approaches for discrimination-free classification. Data Min Knowl Disc 21, 277–292 (2010). https://doi.org/10.1007/s10618-010-0190-x

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in US correctional settings. Psychological Services, 13(3), 206–222. https://doi.org/10.1037/ser0000075.supp (Supplemental)

Sætnan, A. (Ed.), Schneider, I. (Ed.), Green, N. (Ed.). (2018). The Politics and Policies of Big Data. London: Routledge, https://doi.org/10.4324/9781315231938

Stevens, A., Anisfeld, A., Kuester, B., London, J., Saleiro, P., and Ghani, R. Aequitas: Bias and fairness audit, 2018. URL https://github.com/dssg/ aequitas. Center for Data Science and Public Policy, The University of Chicago.

Tramer, F., Atlidakis, V., Geambasu, R., Hsu, D., Hubaux, ` J.-P., Humbert, M., Juels, A., and Lin, H. FairTest: Discovering unwarranted associations in data-driven applications. In IEEE European Symposium on Security and Privacy, pp. 401–416, 2017. doi: https://doi.org/10.1109/EuroSP.2017.29

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2018). Artificial Intelligence and the Public Sector—Applications and Challenges. International Journal of Public Administration, 42(7), 596-615. doi:10.1080/01900692.2018.1498103