Mitigation of Algorithmic Bias to Improve AI Fairness

Kathy Wang
William & Mary

Mitigation of Algorithmic Bias to Improve AI Fairness

Kathy Wang

William & Mary

Research Mentor:

Dr. Kazi Islam, Research Assistant Professor and Research Scientist, ODU School of Cybersecurity

Table of Contents

Abstract

As artificial intelligence continues to evolve rapidly with emerging innovations, mass-scale digitization could be disrupted due to unfair algorithms with historically biased data. With the rising concerns of algorithmic bias, detecting biases is essential in mitigating and implementing an algorithm that promotes inclusive representation. The spread of ubiquitous artificial intelligence means that improving modeling robustness is at its most crucial point.

This paper examines the omnipotence of artificial intelligence and its resulting bias, examples of AI bias in different groups, and a potential framework and mitigation strategies to improve AI fairness and remove AI bias from modeling techniques.

Keywords: algorithmic bias, artificial intelligence fairness, machine learning, robustness

I. Introduction

The increasing use of artificial intelligence (AI) and algorithmic-based models in both public and private sectors increases the risk of prejudiced decisions based on demographic factors. Beginning as a neutral and objective alternative in decision-making, algorithmic-based models have been found to be increasingly biased through their collection and processing of data, creating a feedback loop of systematic inequalities in society. As one of the most transformative technologies, artificial intelligence and machine learning models have reached a critical inflection point where negative impacts can no longer be ignored.

In June 2019, a study of 3.2 million mortgage applications and 10 million refinance applications from two federally backed institutions, Freddie Mac and Fannie Mae, found evidence of racial discrimination in face-to-face lending and algorithmic lending. Of these 13.2 million applications, the overall likelihood of rejection is 49.6%. Further deconstruction of these numbers found that minority groups face a rejection rate of greater than 60.6% compared with the 47.6% from everyone else. In this one instance alone, AI bias and algorithmic consumer-lending discrimination from Fannie Mae and Freddie Mac cost African American/ Latinx borrowers 765 million dollars per year in extra interest in the United States.[1]

As AI systems and machine learning algorithms increasingly evolve to automate decisions dealing with sensitive information, existing flawed algorithms can amplify unconscious socioeconomic, ethno-racial, and gender bias. Only diligently identifying and

[1] Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019, June 17). *Consumer-lending discrimination in the Fintech Era*. NBER. Retrieved November 25, 2022, from https://www.nber.org/papers/w25943

acknowledging current instances of AI bias and examining flawed algorithmic systems can improve AI modeling fairness/robustness and ultimately alleviate AI bias from modeling techniques.

II. What is AI Bias, and Why is AI Fairness Essential?

Artificial intelligence bias, also known as machine learning bias, refers to the tendency of algorithms to reflect human bias or the phenomenon that arises when algorithms deliver systematically biased results due to erroneous assumptions.[2] Unfortunately, sources of AI bias can be unconsciously influenced by underlying data rather than the algorithm itself. This may be through data collection techniques or through a biased feedback loop. For example, in criminal justice, searches for African-American identifying names tended to result in more online ads featuring the word "arrest" than searches for white-identifying names. This is the direct result of algorithmic models being trained on data containing human-centered choices or data from social disparities. [3] This is incredibly substandard as algorithms like the one mentioned in Sweeney's research on racial differences in online ad targeting is reinforcing biases in algorithmic- based models that were previously thought to be neutral and fair.

In order to mitigate AI bias, it is essential to begin with the concept of fairness in the context of artificial intelligence. Fairness is broadly defined as the absence of prejudice or

---

[2] Levity AI GmbH. (2012, November 16). *Ai bias - what is it and how to avoid it?* Levity. Retrieved December 3, 2022, from https://levity.ai/blog/ai-bias-how-to-avoid

[3] Sweeney, L. (2013, March 1). Discrimination in online ad delivery: Google ads, black names and white names, racial discrimination, and Click Advertising: Queue: Vol 11, no 3. Queue. Retrieved December 3, 2022, from https://dl.acm.org/doi/10.1145/2460276.2460278

preference for an individual or group based on its characteristics. [4] Being polysemous, it is also important to understand that most algorithms cannot be completely bias-free because of various perspectives of what "fairness" entails.[5] Nonetheless, there are still several approaches to enforcing fairness on AI models.

Artificial intelligence must be designed to minimize bias and promote inclusivity as it becomes increasingly pervasive. It is highly connected to AI models' interpretability and transparency of the model creation process.[6] Furthermore, AI fairness is crucial as its outcome could be detrimental to communities when left unchecked.

[4] Reagan, M. (2021, April 2). *Understanding bias and fairness in AI Systems*. Medium. Retrieved December 3, 2022, from https://towardsdatascience.com/understanding-bias-and-fairness-in-ai-systems-6f7fbfe267f3

[5] Fu, R., Huang, Y., & Singh, P. V. (2020, October 16). *Ai and algorithmic bias: Source, detection,mitigation and implications*. SSRN. Retrieved December 3, 2022, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3681517

[6] Lapach, Y. (2022, November 28). *Fairness in ai*. 2021.AI. Retrieved December 3, 2022, from https://2021.ai/fairness-in-ai/

III. Instances of AI Bias

AI Bias Towards Race:

In 2016, research done on over 10,000 criminal defendants found that the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), an algorithm used in United States court systems to assess a criminal defendant's likelihood of becoming a recidivist, would find black defendants far more likely to a be higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk. In fact, the COMPAS tool would misclassify black defendants as higher risk at 45 percent compared to white counterparts at 23 percent.[7]

AI Bias Towards Gender:

In 2014, artificial intelligence specialists began building a résumé filtering computer program to automate the search process. However, by 2015, Amazon realized its system was not inclusively rating candidates. Due to poor observations and a male dominance across the tech industry, computer models were trained to vet applicants by detecting patterns in résumés submitted to the company over ten years, most résumés from men. As a result, Amazon's algorithm taught itself that male candidates were preferable and penalized résumés that included the word "women's", as in "women's chess club captain". Amazon attempted to edit the

[7] Larson, J., Angwin, J., Kirchner, L., & Mattu, S. (2016, May 23). *How we analyzed the compass recidivism algorithm*. ProPublica. Retrieved December 3, 2022, from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

programs to make them as neutral as possible, but there was no guarantee that the algorithm would not devise other ways of sorting candidates that could prove discriminatory.[8]

AI Bias Towards Disabilities:

If data used to train a pedestrian recognition system does not include representations of people using bicycles or wheelchairs, it is extremely likely that these individuals will not be recognized as pedestrians. For example, in 2018, an autonomous Uber in Arizona killed Elaine Herzberg, a pedestrian who was pushing a bicycle when she was killed. A recent National Transportation Safety Board investigation found significant problems with Uber's autonomous system. The investigation found that Uber's system had a hard time classifying Elaine Herzberg: " When the car first detected her presence, 5.6 seconds before impact, it classified her as a vehicle. Then it changed its mind to 'other,' then to vehicle again, back to 'other,' then to bicycle, then to 'other' again, and finally back to bicycle." [9] Would it similarly misclassify people on wheelchairs?[10]

---

[8] Reuters. (2018, October 10). *Amazon ditched AI recruiting tool that favored men for technical jobs*. The Guardian. Retrieved December 3, 2022, from https://www.theguardian.com/technology/2018/oct/10/amazon-hiring-ai-gender-bias-recruiting-engine

[9] Marshall, A. (2019, November 6). *Uber's self-driving car didn't know pedestrians could jaywalk*. Wired. Retrieved December 4, 2022, from https://www.wired.com/story/ubers-self-driving-car-didnt-know-pedestrians-could-jaywalk/

[10] Whittaker, M. (2019, November). *AI Now Institute*. Retrieved December 4, 2022, from https://ainowinstitute.org/disabilitybiasai-2019.pdf

AI Bias Towards Religion:

GPT-3, an artificial intelligence system that generates text hailed for its potential to enhance creativity, disproportionately associated Muslim with violence. When researchers took out “Muslims” and put in “Christians” instead, the AI went from providing violent associations 66 percent of the time to giving them 20 percent of the time. GPT-3, created by the research lab OpenAI, is aware of the anti-Muslim bias. In fact, the original paper published on GPT-3 noted that they found that words, such as “violent”, “terrorism”, and “terrorist” co-occurred at a greater rate with Islam than with any other religions and were in the top 40 most favored words for Islam in GPT-3.[11]

[11] Samuel, S. (2021, September 18). Ai's Islamophobia problem. Vox Media. Retrieved December 4, 2022, from https://www.vox.com/future-perfect/22672414/ai-artificial-intelligence-gpt-3-bias-muslim

IV. Mitigation of AI Bias and Improving Robustness

With rising concerns in AI bias, a systematic evaluation of existing approaches to minimize bias is the first step to improving AI modeling fairness and robustness. Three broad categories have been divided to better examine the process of mitigating AI bias.

Understanding bias. Approaches that help understand how bias is created in the society and enters our socio-technical systems, is manifested in the data used by AI algorithms, and can be modeled and formally defined.

Mitigating bias. Approaches that tackle bias in different stages of AI-decision making, namely, preprocessing, in-processing, and post-processing methods focusing on data inputs, learning algorithms, and model outputs, respectively.

Accounting for bias. Approaches that account for bias proactively, via bias-aware data collection, or retroactively, by explaining AI-decisions in human terms. [12]

A way to understand and mitigate AI bias is by strengthening AI literacy throughout society as it would arm more of society with the skills needed to adapt to the ever-changing field of AI. For example, increasing investment in education at all levels could foster a more inclusive and diverse AI ecosystem and support mitigating misconceptions around AI. Creators of algorithmic models should also prioritize opportunities that advance equity in AI. These collaborations should include representatives from communities most impacted by inequities.

[12] Ntoutsi, E. (2020, February 3). *Bias in data driven artificial intelligence - An introductory survey*. Wiley Interdisciplinary Reviews. Retrieved December 4, 2022, from https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1356

Accounting for bias proactively includes establishing universal limitations of AI and adopting responsible licensing practices, which may help prevent high-risk AI systems from being overleveraged, irresponsible, or illegal in some cases. This may include establishing limitations to prohibit AI from racial profiling and violating basic human rights and freedoms. However, accountability also includes implementing mechanisms for consumer insights and feedback to help capture issues, concerns, or complaints from consumers related to automation and algorithms, as well as assessments and testing for high-risk AI systems to focus on protecting consumers from harm while enabling innovation. [13]

[13] Hobson, S., & Dortch, A. (2022, April 14). IBM policy lab: Mitigating bias in Artificial Intelligence. IBM Policy. Retrieved December 4, 2022, from https://www.ibm.com/policy/mitigating-ai-bias/

V. Concluding Remarks

Many of the underlying problems in algorithmic bias is encoded in human bias. As artificial intelligence continues to evolve, these machines and algorithmic models embedded in biased historical data must be approached in a manner that seeks to understand current flawed algorithms, tackle bias pre and post processing, and remain accountable proactively and retroactively.

The human tapestry is not as black-and-white as black and white, it is rich in complex subgroups and combinations.[14] It is true that due to the multifaceted nature of fairness there is no one ideal solution that is mutually agreed upon; however, biased data sets produce biased decisions which amplifies and threatens to perpetuate biased algorithms and machines. It is crucial that fairness in artificial intelligence and algorithms is maximized, as all sectors of the economy and society become increasingly algorithmic in decision-making and its resulting negative impacts must be addressed.

[14] Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2019, June 17). *Consumer-lending discrimination in the Fintech Era*. NBER. Retrieved November 25, 2022, from https://www.nber.org/papers/w25943