Algorithmic Bias: Invasion of Human Bias into Algorithm

Ankita Gaonkar[1], Soham Akhave[2], Mona Deshmukh[3]

[1]PG Student, Dept. of MCA, Vivekanand Education Society's Institute of Technology, Mumbai, India
[2]PG Student, Dept. of MCA, Vivekanand Education Society's Institute of Technology, Mumbai, India
[3]Asst. Professor, Dept. of MCA, Vivekanand Education Society's Institute of Technology, Mumbai, India

Abstract - *With the changing times, AI and Machine Learning have become a big part of our daily lives. Decision making aided by big data has become essential in many sectors like healthcare, finance, hiring and many more. This data-driven decision-making system is expected to be a neutral player who will give us the results which are not corrupted by human cultural and societal discrimination. But we have to ask ourselves if the systems are truly unbiased. Though expected to be fair in their workings, algorithms have also proven to be prone to have a bias. Amazon's Hiring Algorithms, for example, have shown to have a preference for men than women[1].*

Algorithmic bias exists and sometimes it is necessary. The algorithms are designed to filter out the data so that humans can work more efficiently. But the filtering in an algorithm can sometimes be discriminatory from an ethical standpoint. This kind of bias is mainly driven by human and societal bias that had created the need for neutral systems initially. Bias can be avoided through careful and conscious planning with the help of various tools which are created to check for discrimination in AI-based decision making. In this paper, we discuss the definition of algorithmic bias, necessary filtering, unfair bias and what could be its causes.

Key Words: **bias, algorithm, discrimination, AI, algorithmic bias**

1. INTRODUCTION

The term bias can be used to refer to favouring one over the other. But not all bias is discriminatory. Example, if a bank were to give a customer a lower credit rating because they have a history of unpaid credit loans then the bank is within its rights to do so. The bank would not want to lose money when they can prevent the loss. But in a similar scenario, if the cause for lower ratings is the customer's ethnicity or other unrelated factors then the bank is said to be unfairly biased. Another example would be the use of alphabetical ordering for a search result on a booking website, where the price and the facilities remain the same but the topmost search result would unfairly benefit from this[2].

The term algorithmic bias to refer to computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favour of others[2]. The bias may come from the already corrupt training data which was created from decisions made by humans. The data is then biased to start with and the algorithm will follow the suit by making decisions with the same metrics that humans used.

In some situations, the algorithmic bias is written in the algorithm, like filtering algorithms in social media platforms. The feed that a user gets on social media platforms is made of the posts from the people that they are connected with or have a shared interest with. Such filtering may also lead to interactions among people who are like-minded. These kinds of interactions lead to uncritical conformity, where another opinion is considered invalid. The tailor-made news feed for users may create "filter bubbles"[3] or "echo chambers"[4]. While filtering makes the user's experience on the app pleasant it festers some ideals that may go unopposed by other users. For example, user A who is an anti-immigration supporter is using the platform to publicize their views. Another user B with the same views interacts with the user A. Both the users now have someone who conforms their views on the matter. The original post will attract many other like-minded users. The users who do not agree with the ideals usually avoid interacting with the post and if they do they might not be considered as an intellectual who is stating their opinion, which is opposite of what the user A believes in, but rather are treated as trolls. This kind of filtering further magnifies the problem of an echo chamber when used with identifiers like hashtags.

With the widespread use of the internet and the services provided on it, some organisations are in possession of enormous data. Organisations use this data to determine which ad-placement would be most profitable. These placements are determined by the data generated by the user interacting with the services. Example if a user has liked numerous pages related to travel on Facebook then they are more likely to get an advertisement for travel-related services like booking sites. This type of filtering is widely accepted despite the concerns of data privacy.

2. CATEGORIES OF BIAS IN COMPUTER SYSTEMS

Friedman and Nissenbaum categorized bias, in computer systems, in the following types: Pre-existing Bias, Technical Bias and Emergent Bias[2]. These categories were determined by examining the typology of existing systems.

Pre-existing is a biased way of thinking that has been ingrained in us as individuals and as a member of society. This kind of bias generally stems from society, its culture and subculture. For example, associating a certain profession with a particular gender. Google search results for images for the keyword "teacher", the majority of the images are of females. Same can be observed with the search for the keyword "nurse". We know that these professions are not limited to women. But Google image search results are ranked by an algorithm which ranks them most by most site visits.

Technical bias arises from technical constraints or technical considerations like a transport booking system which displays the available transport by alphabetical listing rather than listing them by the order of their departure time.

Emergent bias is the prevalent bias in social circumstances that could not have been relevant at the time of the system's conception. For example, a system which was tested on a particular group of people but is aimed towards another group. Example the biometric recognition systems at airports which were tested on a training data consisting of a majority of white people and hence the systems have difficulty in recognising people of colour. This kind of bias would also be created due to a change in social view by the general society. Facebook had to inject a hate-speech monitoring algorithm after being criticized for allowing content which targeted minorities. The 2017 policy was made to be colour-blind, where all the users, disregarding their age, race or religion would be equally protected. But the company again was criticized by experts when reports indicated that it protects the already advantageous group i.e. white men[5]. The company has since modified the way its policies on hate speech and improved its hate speech detection system[6] but still has a long way to go to a near-perfect system.

3. CAUSES

3.1 Data reflects existing bias

When it comes to Machine Learning and AI, the training data-set plays an important part in 'teaching' the system on how to behave. The training data is a set of parameters, relevant and non-relevant, and the decisions that were made based on the parameter. The system scans the training data for patterns. Once it detects the pattern, it follows the same pattern to make decisions. But the problem with the system is that the data that was fed to it was created by humans and thus was muddled with human conscious and unconscious bias. Amazon's hiring algorithm had this particular problem. The training data fed to it consisted of the resumes submitted for a ten year period. The corporate industry at the time, from which the data was used, was known to be biased against hiring female candidates and also the number of female candidates was less. The algorithm was blind to the data of the applicant's gender but the algorithm found a way to derive it. Candidates who went to two or more women's colleges were filtered out and the technology was reported to prefer masculine language[7].

3.2 Unbalanced Classes

Unbalanced data is where the system was not tested vigorously through all possible demographics. The most common example of such systems is biometric systems like face recognition software which are biased against people of colour or speech-controlled systems which have had difficulty understanding accents[8]. In 2016, the UK employed a system that used facial recognition for passport renewal. The system was found to have difficulties in detecting faces which were very light or were very dark when it was tested. Despite the apparent flaws in the system, it was still implemented with the understanding that it was obviously biased[9]. In a similar scenario, the system used by the New Zealand government for passport renewal couldn't identify a person of Asian descent. The system here wrongfully kept registering that the person's eyes were closed[10].

3.3 How do we measure non-quantifiable qualities

A non-quantifiable quality includes human traits like competence or bravery. These characteristics that are present in the human world are hard to measure with machines. If a job applicant were to say they are a diligent worker in their application, the hiring manager has no way to know how diligent they are. But some employers may take another route to measure the applicant's perceived productivity, like asking the applicant if they smoke, with a deep-rooted assumption that smokers are less productive than non-smokers. This puts smokers at a disadvantage over non-smokers even if they are the best fit for the job description. Another example would be an essay grading system for schools and entrance exams. These grading systems are trained to check for sentence length and vocabulary with no way to check if the sentence makes sense in the context of the essay. These AI can hence be manipulated to give higher grades by using complex vocabulary and constructing long sentences.

3.4 Existing bias in data is reinforced due to the positive feedback loop

The already biased training data produced biased results. These biased results are fed into the system thus amplifying the bias. To explain this with an example we can look at a crime prediction system. Crime prediction systems, like PredPol, are used by law enforcement to auto-deploy police to the location where they can prevent the occurrence of crime, efficiently using the limited police officials available. The theory being the system would be unbiased to race, ethnicity and economic background of the neighbourhoods. But the systems like PredPol work with feedback, meaning the action taken after its prediction and the subsequent outcome are fed into the system. The data on which the prediction was based on would definitely be biased to a certain degree. Hence, so is the prediction. Say the police go to the neighbourhood that the system predicted and successfully make an arrest, then the prediction was right and this is let known to the system. But the issue with this is just because the prediction was correctly made pertaining to certain neighborhoods doesn't necessarily mean that the other neighborhoods are crime-free. It just means that the system has no data of any known crime committed in the other neighbourhood. In a 2017 paper, the authors argued that "the problem of feedback makes traditional batch learning frameworks both inappropriate and (as we shall see) incorrect."[11]

3.5 Malicious attack on the training data

The AI chatbots are susceptible to a malicious attack by the users. Some AI chatbots are initially trained for basic conversation and are designed to supposedly get better at more meaningful and intelligent conversations as they are used more. Microsoft released a chatbot for Twitter in 2016, named Tay. Tay was preceded by XiaoIce, a chatbot released in 2014 that was successful in China. But Tay had issues - it was released on Twitter, which is arguably a cesspool of internet trolls. The chatbot had to be shut within 24 as it went from tweeting out "humans are super cool" to racist, anti-semitic and sexist tweets. This was because the bot had a feature of 'repeat after me' where the user would tweet something at the bot and the bot would return the sentiment. The users exploited the feature and the bot reflected those views without realizing that some of those tweets were inappropriate[12].

4. CONCLUSIONS

Inclusion of AI lead technology in our daily lives is a foregone conclusion. But we need to know that the systems are treating everyone fairly. The humans are biased and so is the data created by them and biased data results in biased decisions. The bias in AI-assisted systems has long been a talking point in the AI, Machine Learning and Data Mining community. Algorithmic bias can lead to systemic discrimination of the groups who have been long been the victim of human bias. And while AI systems can not be designed to be completely unbiased as of yet, the least that can be done is that being aware that the bias exists.

REFERENCES

[1] "Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, Oct. 10, 2018.
[2] B. Friedman and H. Nissenbaum, "Discerning bias in computer systems," in INTERACT '93 and CHI '93 conference companion on Human factors in computing systems - CHI '93, Amsterdam, The Netherlands, 1993, pp. 141–142, doi: 10.1145/259964.260152.
[3] E. Pariser, The Filter Bubble: What The Internet Is Hiding From You. London: Penguin Press, 2011.
[4] C. R. Sunstein, Republic.com 2.0. Princeton University Press, 2007.
[5] J. Angwin, ProPublica, H. Grassegger, and special to ProPublica, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children," ProPublica. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms (accessed Jun. 07, 2020).
[6] "Facebook's AI for Hate Speech Improves. How Much Is Unclear," Wired, May 12, 2020.
[7] "Insight - Amazon scraps secret AI recruiting tool that showed bias against women," Reuters, Oct. 11, 2018.
[8] G. McMillan, "It's Not You, It's It: Voice Recognition Doesn't Recognize Women," Time, Jun. 01, 2011.
[9] Oct 10 and 2019 | Luana Pascu, "UK knowingly deploys biased facial recognition passport checking system," Biometric Update, Oct. 10, 2019. https://www.biometricupdate.com/201910/uk-knowingly-deploys-biased-facial-recognition-passport-checking-system (accessed Jun. 07, 2020).
[10] "New Zealand passport robot tells applicant of Asian descent to open eyes," Reuters, Dec. 07, 2016.
[11] D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian, "Runaway Feedback Loops in Predictive Policing," ArXiv170609847 Cs Stat, Dec. 2017, Accessed: Jun. 09, 2020. [Online]. Available: http://arxiv.org/abs/1706.09847.
[12] "Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day - The Verge." https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist (accessed Jun. 09, 2020).