DEALING WITH BIAS IN ARTIFICIAL INTELLIGENCE

Hari Priya Kandasamy

School of Information, The University of Texas at Austin

Dealing with Bias in Artificial Intelligence

Initially intrigued by how the movie 'Interstellar' portrays a robot that performs tasks similar to a crew member and the growing Artificial Intelligence (AI) advances such as surgical robots that maximize hospital efficiency and smart assistants which understands our voices and assist in booking cabs on Uber, I understood the significance of AI in improving the way we live and work. Moreover, there are also other AI and Machine Learning systems that recommend us books on Amazon, similar to the ones we've liked in the past. AI in dating apps has become more progressive which allows us to track behaviors outside of the dating apps to find which websites we visit, the television shows we watch, and uses it to enhance our matches. All of these are great advances but potentially harmless applications of AI. If our voice assistant doesn't understand us, we could just open the Uber app and book a cab ourselves. If Amazon recommends a book that we might not like, there are numerous resources available on google to research about the book and discard it if we don't like it. If we end up on a blind date based on suggestions from the app, we might even end up having a good time meeting them. Things get rough, however, when we use AI for more serious tasks like giving out loans, filtering job candidates for interviews, or even for medical diagnosis. All of the previous decisions, partially assisted or completely taken care of by AI systems could have an impact on people's lives.

Algorithms are being used all the time to make decisions about who we are and what we want. We are reinforcing our bias in how we interact with AI. Voice assistants like Alexa, Siri are all females that are designed to be obedient, helping out people with ordering food and products on social media while male AI Assistants like IBM Watson are used to make high-level decisions. Sharma, K(2019) talks about the decisions being made about us by AI based on our gender, race, or background He gives examples of bias induced in society. A black/Latino person is less likely

than a white person to pay off the loan on time. A person called John makes a better programmer than a person called Mary. A black man is more likely to be an offender than a white man. This actually sounds like a racist/sexist but these are some real decisions made by AI based on the biases it has learned from the humans. Let's imagine an AI is helping a hiring manager to find the next tech leader and most of the former leaders have been men, so it learns that a woman is less likely to succeed as a tech leader. This leads to screening out of female candidates. AI is being used to help decide whether or not we get the job interview, how much we pay for car insurance, how good the credit score is, and even the rating we get in our annual performance review but these decisions are all being filtered through its assumptions about our race, identity, gender, and age.When kids growing in today's world around AI google "CEO" would see most images of men and on googling "personal assistant" will find females. The brightest minds in the world are creating this technology with all the biases. Even a company as powerful and technologically advanced as Google can encounter these kinds of problems, imagine the hundreds of thousands of other businesses that develop AI-based software and applications without such expertise. This is a good example of how challenging it can be to train AI software to be consistent and robust.

The data needs to be fed into Machine learning (ML) systems to perform ML tasks which needs to be trained in representative samples of a population. Models are sometimes trained on data containing human judgments or on data that reflect societal or historical inequity. Training on data that has been labeled with some sort of bias, mimics the same bias in the resulting AI system.To improve the effectiveness of these systems, we need to identify and eliminate biases. Lewis (2020) discusses the measures that business leaders and policymakers can take to mitigate bias. Talking about the possibility of humans and AI collaborating together, and focusing on diversity are some steps he mentions to eliminate exclusivity and unfairness. People's experiences of dealing with and

interacting with the technology could serve as a starting point for building platforms that use inclusive coding by factoring in fairness by collecting their experiences. It is crucial that algorithm developers are mindful of the social impact of the technology they are developing.

Burrell, J. (2016) tells about the development of predictive technologies to predict future events, such as the likelihood that an individual will default on a loan or the likelihood that a consumer will buy a specific product online. Recent decades have seen a proliferation of algorithmic technologies within the criminal justice system in the United States. Police departments now rely on predictive software programs to target potential victims and offenders and predict when and where future crimes will occur. (Brayne 2017). The use of predictive technologies raises a number of questions about the justice system's fairness and equity. Critics argue that algorithms tend to embed bias and reinforce social and racial inequalities, rather than reducing them (Benjamin 2019).

There's a problem for systems of classification and ranking that have social consequences, such as spam filters, credit card fraud detection, search engines, news trends, market segmentation, and advertising, as well as insurance and loan qualifications. In many cases, computational algorithms, and in other cases machine learning algorithms, are used to perform these classifications. Algorithms for machine learning are highly effective predictors and generalizers. Due to the fact that the accuracy of these algorithms improves when more data is available, the increasing availability of such data in recent years has renewed interest in these algorithms. Kosinki (2017) examined how inconsistent data affected the machine's predictions. Its findings, however, were not conclusive. Generally speaking, women have a bigger forehead, a thinner jaw, and a longer nose than men. The computer then gave him a list of 100 faces it thought were gay or straight, and he averaged the proportions of each. Gay men's faces exhibited more feminine proportions than

women’s. It could support the theory that testosterone levels, which are already known to affect facial features, also have an effect on sexuality. Kosinski's findings suggest something even stranger: that artificial intelligence can often excel by developing new ways of seeing, or even thinking, that are inexplicable to us. We are dealing with a more profound version of what is sometimes called the "black box" problem, which refers to our inability to discern exactly what machines do when they are learning new skills. Silberg (2019) argues that a more diverse AI community will be better able to anticipate, spot, and review unfair bias issues and engage communities likely to be affected by bias. ML developers should be mindful while preprocessing data, tailoring available models accordingly. It is crucial to teach them how to deal with the biases in the AI pipeline. This could take the form of running algorithms alongside human decision-makers, comparing results, and examining possible explanations for differences, requiring investments on multiple fronts, but especially in AI education and access to tools and opportunities.

Increasing numbers of practitioners are focusing their work and careers on translating these calls to action into their areas of practice due to an increased awareness of algorithmic bias and the need to responsibly build and deploy Artificial Intelligence (AI).In order for the mathematical models to be regulated, the first step must be taken by the modelers themselves. Model designers should avoid using overly complex mathematical tools that obscure their simplicity and explainability. These models should be built only based on carefully chosen data, and the use of dangerous proxies should be avoided.In addition, they should always keep in mind the final goal of the models, i.e. making people's lives easier, providing value to the community, and improving our overall quality of life, rather than focusing on Machine Learning metrics like accuracy or mean squared error. If the model is designed for a particular business, another usual success metric will probably have to

be put on a second plane: economic profit. Aside from profit, it is important to examine the results of the model in terms of the decision it is making: given our insurance example, the creators of the model should find who is getting rejected, and try to understand why. We have a responsibility as AI practitioners to reevaluate the ways in which we collect and use data with the goal to develop non-black-box, explainable models, audit these models, and track their results carefully, taking the time to manually analyze some of the results. It is crucial that technology is used in a responsible manner in such a way that it has a positive impact on humans and represents the values we believe in. Our AI systems of the future will have to be fair, robust, illustrative, transparent, and aligned with the values of the society for which they are designed.

Bibliography

1. Lewis,M(2020).Tackling Bias Issues in Artificial Intelligence

2. Silberg,M(2019). Tackling bias in artificial intelligence (and in humans) https://www.mckinsey.com/featured-insights/artificial-intelligence/tackling-bias-in-artificial-intelligence-and-in-humans

3.Schwartz, D., Jonas (2021). Identifying and Managing Bias in Artificial Intelligence

4. Stevens,N & Keyes,O (2021). Seeing infrastructure: race, facial recognition and the politics of data, *Cultural Studies*, DOI: 10.1080/09502386.2021.1895252

5. Rao (2020) Artificial Intelligence risks in criminal justice system

6. Sharma ,K (2019) How to keep Huamn Bias out of AI https://www.youtube.com/watch?v=BRRNeBKwvNM

7. Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms

8.Kosinki,M (2017). The history of Artificial Intelligence

8. Brayne, S., & Christin, A. (2017). Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts