UCLA

Recent Work

Title

Bias in Artificial Intelligence Models and its Effect On Health Outcomes of Vulnerable Groups

Permalink

https://escholarship.org/uc/item/2f59r9zb

Author

Blanco, Andres

Publication Date

2025-03-17

eScholarship.org Powered by the California Digital Library University of California

Bias in Artificial Intelligence Models and its Effect On Health Outcomes of Vulnerable Groups

Andres Blanco[1]

[1] *UCLA, Computer Science Department and Electrical and Computer Engineering Department, 410 Westwood Plaza, Los Angeles, CA, 90095, USA*
**blancoandres45@g.ucla..edu*

Abstract: AI models progress rapidly and are widely used for knowledge, but they inherit biases from their data and creators. Though outputs appear objective, these inherent biases can cause harm if treated as unbiased truth.

INTRODUCTION

The world of artificial intelligence began with the advent of computing and the Von Neumann model of computer architecture in the 1950s. Alan Turing created the "Turing Test" in 1949, a method of experimentation used to glean if computers were as intelligent as humans. While for much of the history of computers, a computer passing the Turing Test seemed part of a distant sci-fi future, various versions of the Turing Test have been passed by artificial intelligence models today. The progress of these models has occurred exponentially quickly. Just 10 years ago chatbots could only output rudimentary, predetermined responses or compiled information from a search engine. Today, chatbots can output stories that seem human-made at a glance and emotionally connect to its users in a way seen like science fiction not too long ago.

The exponential progress of AI models stems from breakthroughs in machine learning algorithms and through the exploitation of the enormous amounts of data generated from the internet and its users. The progress of these artificial intelligence models has been too rapid for people to keep up with. Many people take the output of these models as fact, not realizing that these models were made by imperfect humans and internet data scraped from web pages through machine learning algorithms. This data often contains the cultural, historical, and social biases of modern society and may perpetuate discriminatory beliefs and stereotypes. While these biases might seem relatively harmless when AI models are asked simple questions, as AI becomes increasingly integrated with modern society, these biases can quickly become harmful. In a hospital setting, biases in AI models may lead to inadequate treatment of racial or sexual minorities. In a policing setting, biases may lead to discriminatory targeting of racial minorities and physical harm if given the tools. This paper will focus on the dangers of the biases of AI models, specifically in the healthcare industry, showing how AI can perpetuate the inequalities of modern society. We will tackle how AI can be changed to better reflect realities in a way that is inclusive, culturally aware, and less prone to perpetuating structural inequalities. "This work is in partial fulfillment of the ENGR184 course using the blueprint curriculum in Ref.[1,2] and captured in a collection [3]"

METHODS

We will be investigating the health outcomes of historically disadvantaged minority groups and relating these disparities to how AI models can further exacerbate these disparities and produce unequal outcomes.

Today, the health outcomes of minority groups like black and hispanic people are often worse than those of white people. The healthcare industry has been optimized to serve the privileged who often have easy access to its services and pay lots of money through insurance and individual health care costs.

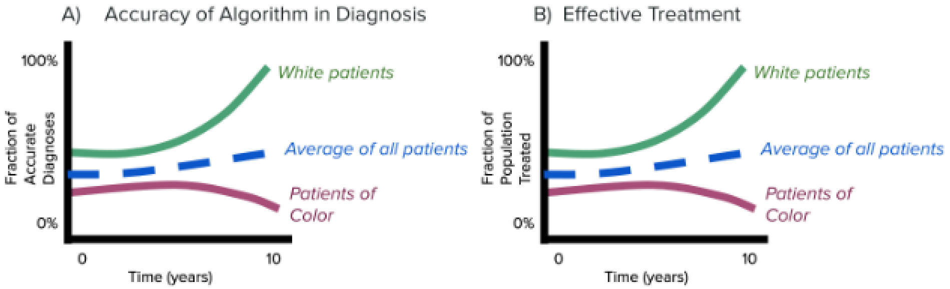

Fig. 1. Algorithmic Bias in Healthcare leads to unequal outcomes (Ref. [4], Fig. 1)

This figure provided by the organization Data for Black Lives shows the long term effects of misdiagnoses due to algorithmic biases. The negative effect of misdiagnoses compounds over time leading to distrust of the healthcare system by minority groups, and the damage it can cause to the health outcomes of disadvantaged minorities.

RESULTS AND INTERPRETATION

The algorithms used to power AI Models often fail to take into account the diverse backgrounds of the people they serve, resulting in a discriminatory outcome. The data of medical patients and medical studies have used computers since the 1970s and algorithms have played an important role in assessing medical diagnoses and providing treatment. Vulnerable groups like racial or sexual minorities have often been missing, misrepresented or underrepresented in the datasets of healthcare providers (Natalia). Since current AI models train off of these data sets, the insufficient data for vulnerable groups leads to worse outcomes for them and distrust of the healthcare system. It has been well documented that the worse health outcomes for minorities has led to them visiting the doctor less often as they don't trust them to provide the right treatment or diagnoses (Kuhlberg, Jill A).

It is important to note that AI models are not omnipotent. They do not output the truth 100% of the time and fall into the same traps and fallacies as normal humans. As a doctor can misdiagnose a black patient who shows similar symptoms to a well documented case more common in white patients, an AI model who uses the same dataset as the doctor, can misdiagnose as well. AI is not separated from our society and our human shortcomings, it is a product of our very creation parroting the same discriminatory institutions and systems in our society.

During the covid pandemic AI models were used by Pfizer and Moderna in the creation of the covid vaccine (Sharma). These models helped these pharmaceutical companies develop a vaccine at a much faster rate than would have otherwise been possible. However, the mortality rates for black and hispanic patients were consistently higher than those of white patients (Macias-Konstantopoulos, Wendy L). While the higher mortality rates could be partially explained through the lower vaccination rates of black and hispanic people, the data set of minority groups is often insufficient and could have led to insufficiencies in the vaccine that did not account for the immune systems of vulnerable groups. The lower vaccination rates of minority groups also shows the distrust of the healthcare system by minority groups which leads to worse health outcomes.

All of these effects compound over time. Accurate diagnoses of medical conditions leads to better outcomes of those with the condition. The treatment is tailored toward the right ailment and the patient can begin their road to recovery. Any subsequent diagnoses of the same condition can be found through previous diagnoses and the treatments can adapt as more patients are diagnosed and their outcomes recorded. A misdiagnoses on the other hand can lead to inadequate treatment that can harm the patient. This can lead to the patient having distrust of the specific hospital or medical industry and choosing not to seek further treatment on their medical condition. Their medical condition in turn is less well known and future diagnoses may lead to the same inadequate treatment. Many minority groups are stuck in this cycle of receiving inadequate care as they are more frequently misdiagnosed and given a treatment plan that does not work. This cycle is hard to break as the trust between minorities and the healthcare industry has long been broken, so efforts to restore this trust and obtain more accurate data on the medical problems of vulnerable groups must be done to ensure better health outcomes.

CONCLUSIONS

It is clear that today's AI models are ill equipped to navigate through the inequalities of today's society. These models were trained on biased training data that misrepresents or underrepresents vulnerable groups. A solution I propose to improve AI models in helping to serve a diverse group of people is to have a grassroots data collective. Working with current day advocacy groups like the ACLU or NAACP to create alternative training data sets can help train algorithms to be more aware of the issues facing vulnerable peoples. This could have algorithms propose multiple solutions based on the intersecting identities and experiences of the individuals they serve and more accurately assess their needs. Anybody can contribute data into this grassroots collective that can be filtered and packaged by a committee like the NAACP, that can be used so AI systems can serve everybody equally, rather than relying on biased data sets that harm minority groups. It may be hard to push this initiative to the mainstream however, especially in the U.S government where this may be seen as D.E.I or "woke". A grassroots movement helps mitigate this as it is not reliant on government or corporate entities, however involving data collection as a public good can help further the cause of removing bias in AI models

Doing medical experiments involving disadvantaged groups and compiling all that data into training sets that AI can use will help models more accurately diagnose and treat any ailments they may face. A group that is currently tackling this issue is the Data for Black Lives who use data science to improve the lives of black people. Algorithms play an ever increasing role in today's society, and recognizing that minority groups are discriminated against in algorithms and AI models is important to address the issue. Focusing on obtaining data that can help disadvantaged groups can help provide equity in AI.

Another potential solution is to make the training sets used by AI algorithms more transparent so people using AI systems can see the biases of the training data. The train of logic of the output of an AI model should also be transparent so any biases are more apparent. This can help people realize the inadequacies of our current models and provide suggestions to them. This solution is incomplete as a lot of knowledge in technology is needed for a person to understand the logic of AI algorithms and to train their own algorithms with more unbiased data. However, suggestions to AI algorithms by experienced developers who come from a diverse set of backgrounds can help create better AI models that are better equipped to help a diverse set of people.

AI has been used in the medical industry in recent years, and its use for the most part has led to increased positive health outcomes for everybody. However, those positive outcomes are concentrated, and many vulnerable groups still struggle with trust in the medical industry and poor health outcomes from inadequate treatment. Having legislature that targets inadequate data for racial, sexual minorities, women and other disadvantaged groups can help lead to better health outcomes for everybody. The importance of AI models and algorithms and healthcare can not be understated and needs to be taken into account on a national and global level.

With AI models increasingly integrating with our society it is imperative that they serve all people and don't perpetuate existing systems of inequality. More radical means of change such as doing away with the capitalistic system that perpetuates these biases may ultimately be needed to root out these biases. AI models mimic the humans that created them and our society, and our fundamentally flawed society that has institutionalized discrimination and inequality leads to AI models perpetuating these same systems. Collectivizing our data and providing everybody with a means to sell their own data and obtain the useful data of others can lead to a system that has more accurate data provided by a diverse set of people. This data can then be used to train AI models that now have access to data that better mimics the realities of all its users and humanity as a whole. This would require a lot more than a simple grassroots movement or political legislation, but is useful to think about to tackle the inequities we face today. We need to think about how moving away from an unequal capitalist system can help reduce or do away with systemic inequities in our modern society. AI model bias is a symptom of wider inequities, but targeting it specifically can still help disadvantaged groups. AI tools are incredibly powerful and can be used by all people to gain knowledge and skills to help them in life.

REFERENCES

1) Lee, Ethan, et al. "Education for a Future in Crisis: Developing a Humanities-Informed STEM Curriculum." *arXiv preprint arXiv:2311.06674* (2023).
2) Y. Sergio Carbajo, Nurturing Deeper Ways of Knowing in Science, Issues in Science & Technology, 2025, v. 41, n. 2, p. 71, doi. 10.58875/jkrw4525
3) Z. Carbajo, Sergio. "Queered Science & Technology Center: Volume 3." (2025).
4) Kuhlberg, Jill A., et al. "Advancing community engaged approaches to identifying structural drivers of racial bias in health diagnostic algorithms." *arXiv preprint arXiv:2305.13485* (2023).
5) Carbajo, Sergio. "Queered Science & Technology Center: Volume 2." (2024).
6) Macias-Konstantopoulos, Wendy L., et al. "Race, healthcare, and health disparities: a critical review and recommendations for advancing health equity." *Western journal of emergency medicine* 24.5 (2023): 906.
7) Norori, Natalia, et al. "Addressing bias in big data and AI for health care: A call for open science." *Patterns* 2.10 (2021).
8) "Medical Devices and Technology Across the Years." *Yale Medicine Magazine*, Yale School of Medicine, 10 Nov. 2021, medicine.yale.edu/news/yale-medicine-magazine/article/medical-devices-and-technology-across-the-years/.
9) Sharma, Ashwani, et al. "Artificial intelligence‐based data‐driven strategy to accelerate research, development, and clinical trials of COVID vaccine." *BioMed research international* 2022.1 (2022): 7205241.