Mitigating Bias in Artificial Intelligence

Hannah Hayes

Florida State University

ENC 2135

Professor Bridgette Sanders

1 June 2023

Introduction

As the use of artificial intelligence grows, so do concerns about how ethical these systems truly are. Areas such as the medical field, legal system, career recruitment, and many more are implementing AI systems into their everyday processes. While it may be expected of these systems to evaluate each person equally, they do not. Questions about ethics have recently begun to surface in regards to the usage of AI technology, but it has been known for many years now that biases are embedded in the systems. Artificial intelligence is growing in popularity each day and systematic biases within these systems have grown more prevalent over the last few decades. Now that artificial intelligence is a part of everyday life for many people across the world, it is concerning how little action has been taken towards eliminating system biases in systems. It is evident that artificial intelligence algorithms will continue to be a prominent part of modern society, but the risks of biases should be addressed before the issue gets out of hand. Research indicates that systemic biases have been unconsciously implemented in artificial intelligence in the past and present systems and they need to be mitigated.

The Beginning of Biases in Artificial Intelligence

Systematic biases have been influencing the results of artificial intelligence systems for decades. Even though technology does not have the ability to think or act as a human does, it is still able to analyze certain occurrences to create shortcuts in its decision making process when given the opportunity. Developers choose the data that is used and determine how the algorithm will use that data to get results, which can allow bias to enter the model if it is not properly tested (Marr). If developers do not carefully review the data that is implemented in systems, they are opening a window for human biases to enter the systems unintentionally. In the past, the concept of artificial intelligence was relatively new and technological resources were much more limited than they are now. The earlier algorithms were more susceptible to the influence of systemic biases as a result of the lack of knowledge and resources. In 1988, a British medical school stirred up controversy when, “the computer program it was using to determine which applicants would be invited for interviews was determined to be biased against women and those with non-European names” (Manyika). While the school had implemented the system to eliminate the biases that come along with having humans review applications, they failed to recognize the potential of unconscious biases entering their algorithms. The data set that was being used to evaluate students and determine if they were eligible for an interview was misrepresenting woman and non-European names, resulting in bias. This was one of the first indicators that biases surrounding gender, race, and individual backgrounds were being identified in algorithms. Almost thirty years ago this problem was brought to light and it has yet to be solved. In fact, biases have only grown more prominent in systems as the popularity of artificial intelligence has grown.

Modern Day Bias in Artificial Intelligence

Bias in artificial intelligence has become a greater issue in the modern world as these systems have become implemented in everyday tasks. Currently, there has been controversy within the legal system and medical field in regards to systemic biases found in algorithms. The utilization of artificial intelligence in criminal court has become the poster child for racial discrimination in algorithms because of its continual misrepresentation of black defendants. The court system in Broward County, Florida, found itself in troubled waters when journalists found that, “...after controlling for defendants’ criminal history, prior recidivism, age, and gender, Black defendants were 77% more likely than White defendants to be scored as high risk of committing a future violent crime” (Gudis). This is a major problem considering the Black community is being unfairly treated when it comes to serious matters, such as criminal offenses and jail sentencing. There is no logical reason why race should constitute a higher risk of commiting a future crime. This statistic is an obvious result of the negative impacts that human bias in artficial intelligence systems has on society. The unfortunate truth is this is not the only area that racial discrimination is being found in artificial intelligence algorithms. In the medical field, it has been found that algorithms seem to issue more resources to White patients than Black patients who have the same degrees of illness (Gudis). It can be difficult to understand how a system that cannot see or understand systemic biases on its own freewill, can begin to discriminate solely based on race. Not only is the black community being falsely labeled as criminals, but they are also receiving less medical treatment in comparison to white people. The, "problems can arise when data sets used to train the algorithms lack demographic diversity or when the data sets themselves reflect disparities in outcomes or bias in human behavior" (Noseworthy). It seems a common denominator in system biases found in algorithms are a result of a poor data set that the system is given to make important decisions with. When White people are overexposed in the data sets that the algorithm is using to make important decisions, such as medical resources and jail sentencing, it has a negative impact on other racial groups who are lacking in representation. These are just two examples of biases in systems, but they make a clear statement that bias is an issue and should be addressed.

Solutions to Mitigate Bias in Artificial Intelligence

There are preventative measures that can be taken during the development process to avoid the infiltration of biases in artificial intelligence, as well as solutions to mitigate biases in pre-existing systems. The original goal of these algorithms was to eliminate the effects that human decision-making has on processes that require the evaluation of individuals to determine a certain outcome. Whether it is choosing the best candidate for a job, or determining someone's likelihood of being a continuous offender, AI systems have fallen short of its original goal. The good news is, it is never too late to implement new processes to eliminate biases. Experts say that, “part of the problem is that companies haven’t built controls for AI bias into their software-development life cycles, the same way they have started to do with cybersecurity” (Bousquetter). Since there is a standard protocol when implementing a cybersecurity system, there should be a similar standard for creating new artificial intelligence algorithms. It seems that since a smaller scale of individuals are being impacted by this problem, it is often brushed under the rug since the majority of users are not being impacted. Perhaps creating a legal obligation to mitigate bias in artificial intelligence may push developers to take this problem more seriously. The three different, “approaches for bias mitigation can be categorized into: (a) preprocessing methods focusing on the data, (b) in-processing methods focusing on the ML algorithm, and (c) post-processing methods focusing on the ML model” (Ntoutsi). Essentially, these processes address how to mitigate biases in each process of the machine learning models generation. From the beginning, middle, and end, it is never too late to implement a solution. Using the preprocessing approach, a developer would need to thoroughly analyze the data being used in the model to ensure that no biases are present and none can unconsciously enter the system. During the in-processing method, a developer would need to redefine how the data is interpreted in the system. Finally, if a developer is using the post-processing approach, they will need to alter the system's entire decision-making process and reverse its previous process to eliminate the potential for bias (Ntoutsi). While all of this is easier said than done, it provides a way for developers to reverse or avoid biases in artificial intelligence algorithms. Although it can be costly and time consuming, it is important for every individual to be accounted for appropriately.

Discussion

System biases found in artificial intelligence systems is not something that can be denied or taken lightly. After thirty years of evidence that individuals are being discriminated against based on race, gender, and other sensitive information, it should be a standard now that anti-bias processes be implemented in the system development process. Although there is not evidence of intentional human manipulation in systems to create bias, it still falls on engineers to reverse the negative effects that have been created by unintentional bias. There are examples of students being denied opportunities based on bias, individuals struggling with health care due to bias, and even people being labeled as criminals as a direct result of biased AI systems. All of the research indicates that a solution should be made, and a set of strict standards should be created to prevent people from experiencing discrimination. Seeing how much artificial intelligence has grown since it was created, it is reasonable to assume that it will only continue to grow in popularity. With all of the biases that are already present, the issue will become widespread as more algorithms are created and implemented into industries across the world. Perhaps it may be easier to mitigate biases if a law or regulation was placed on these systems to prohibit developers from ignoring the potential for biases in their systems. A simple regulation stating that new systems should be tested to a certain extent, or a law making it a punishable offense to allow bias to dictate system's decisions, may lesson the effects AI bias has on society. There are plenty of solutions and methodologies available that can aid in the creation of non-biased systems, so there is no true explanation for why unconcious biases are continuously being found in algorithms.

Conclusion

As the use of artificial intelligence continues to grow in businesses across the world, it is important to address the systemic biases that may be present. Studies and surveys by professionals in the field have shown that this bias has been present in systems for decades. Despite the evidence that systems unfairly evaluate some individuals based on biases, there has not been much advancement towards getting rid of this issue. There are solutions and preventative measures that can be taken to mitigate the negative impact bias has on society. Considering the historical and present issues with systemic bias in artificial intelligence, it may be time to consider implementing solutions and preventative measures to stop them from spreading even more.

Works Cited

Bousquette, Isabella. “Rise of AI Puts Spotlight on Bias in Algorithms.” *The Wall Street Journal*, 9 Mar. 2023, www.wsj.com/articles/rise-of-ai-puts-spotlight-on-bias-in-algorithms-26ee6cc9.

Gudis, David A., et al. “Avoiding Bias in Artificial Intelligence.” International Forum of Allergy & Rhinology, vol. 13, no. 3, 2023, pp. 193–95, https://doi.org/10.1002/alr.23129.

Manyika, James, et al. “What Do We Do about the Biases in Ai?” *Harvard Business Review*, 17 Nov. 2022, hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.

Marr, Bernard. “The Problem with Biased AIS (and How to Make Ai Better).” *Forbes*, 8 Nov. 2022, www.forbes.com/sites/bernardmarr/2022/09/30/the-problem-with-biased-ais-and-how-to-make-ai-better/?sh=516c22b34770.

Noseworthy, Peter A., et al. “Assessing and Mitigating Bias in Medical Artificial Intelligence: The Effects of Race and Ethnicity on a Deep Learning Model for ECG Analysis.” Circulation. Arrhythmia and Electrophysiology, vol. 13, no. 3, 2020, pp. e007988–e007988, https://doi.org/10.1161/CIRCEP.119.007988.

Ntoutsi, Eirini, et al. “Bias in Data-driven Artificial Intelligence Systems—An Introductory Survey.” *WIREs: Data Mining & Knowledge Discovery*, vol. 10, no. 3, May 2020, pp. 1–14. *EBSCOhost*, https://doi.org/10.1002/widm.1356.