UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

CHALMERS

Introduction to Data Science and AI

AI Ethics Essay - It is impossible to create AI that is free of bias

Neha Devi Shakya
University of Gothenburg
CY Tech (Exchange student)
gusshane@student.gu.se

Sarvesh Meenowa
University of Gothenburg
CY Tech (Exchange student)
gusmeesa@student.gu.se

Date Last Edited: May 13, 2022

Human biases are well-documented, ranging from implicit tests revealing predispositions we may not know to research showing how these biases can affect outcomes. Although humans are prone to error and bias, this does not imply that algorithms are inherently superior. Over the last few years, society has grappled with how these human prejudices might have disastrous consequences infiltrating artificial intelligence systems. Being aware of the risks and striving to mitigate them is a top priority as many companies consider deploying AI systems across their operations. AI bias is a variation in the output of machine learning algorithms caused by biased assumptions made during the algorithm development process or biases in the training data.

Nonetheless, these systems can be biased by how they are designed, built, and ultimately used, a phenomenon known as algorithmic bias. It is difficult to discern how systems might be prone to algorithmic bias because this technology frequently functions in a corporate black box. When considering machine learning tools and AI systems, it is better to consider training, which entails exposing a computer to a large amount of data and then teaching that computer to make judgments, or predictions, about the information it processes based on its observation patterns.

For example, suppose a model is trained to determine whether an object is a car based on a few parameters such as shape, weight, and size. A human might be able to do it, but a computer might do it faster. The model is trained with the metrics for each parameter and whether the object is a car or not. After continued testing and refinement, the model is expected to learn what identifies a car and, presumably, forecast whether an object is a car in the future based on those metrics without the need for human intervention.

This example is rather basic and straightforward if the first batch of data was appropriately categorised and comprised various metrics covering eclectic car models. However, similar algorithms are frequently used in situations with far more severe repercussions than this work and where there is not always an "objective" result. Many of these classification models are trained or tested using data that is not complete, balanced, or adequately selected, which can be a significant cause of algorithmic bias.

We can think about algorithmic bias in two primary ways: accuracy and impact, i.e. an AI can have different accuracy rates for different demographic groups or can make vastly different decisions when applied to different populations. There are mainly two causes of biases in AI systems: cognitive biases and the lack of complete data.

Cognitive biases are unintentional thinking errors that affect people's judgments and decisions and result from the brain's attempt to simplify processing information about the world. Cognitive biases could enter machine learning algorithms accidentally introduced by the designers or through a training data set including those biases. On the other hand, incomplete data may not be representative and contain bias. Most psychological research papers, for instance, incorporate results from undergraduate students, which is a specific group that does not reflect the entire community.

When we think of data, we might think of formal studies in which demographics and representation are carefully considered, limitations are weighed, and peer-reviewed results. That is not necessarily the case with the AI systems that might be used to decide for us. Let us take one data source that everyone has access to, i.e. the internet. Teaching artificial intelligence to crawl through the internet and just reading what humans have already written would result in prejudices such as racism and sexism.

An example of how training data might cause racism in an algorithm occurred a few years ago in 2016, when Microsoft was ready to launch its new Twitter chatbot, Tay [4]. Tay, described as a "conversational understanding experiment," was created to engage people in conversations via tweets or direct messages while imitating the style and lingo of a teenage girl. Tay was a machine learning, natural language processing, and social network experiment. Unlike other chatbots, such as Joseph Weizenbaum's Eliza, Tay was designed to learn more about English and allow her to engage in conversations about virtually any topic over time.

When Tay was initially released on Twitter, she engaged in light banter and terrible jokes with her rising number of followers. However, after only a few hours, Tay began tweeting incredibly offensive things, with a worrisome percentage of her tweets being abusive and offensive. What the company intended to be a fun experiment in "conversational understanding" became their golem that spiralled out of control through the force of language. Many reports also detailed precisely how a bot that was supposed to mimic the language of a teenage girl became so vile. Another thing to keep in mind is: Just because a tool is evaluated for bias against one group does not mean it is tested for bias against another, which is particularly true when an algorithm simultaneously considers numerous identification criteria. In some circumstances, finding bias-free training data may be impossible.

Amid discussions about algorithmic biases, companies utilising AI may claim to be taking safeguards, such as using more representative training data and routinely reviewing their systems for unintended bias and uneven impact against specific groups. We could aim to create more representative datasets that could be part of the solution, though it is important to note that not all efforts to improve data sets are ethical. It is also not only about the data as an AI might be programmed to define a problem in a fundamentally flawed manner.

It is clear that AI systems are not objective and are subject to bias, reflecting their developed context. There are countless ways in which potentially significant biases might skew these systems' decisions and predictions. Nonetheless, some people dispute if this is any worse than regular human decision making. Indeed, others have claimed that, while significant biases can influence the working of AI systems, we at least have the ability to analyse, understand, and correct these biases [3]. In contrast, biased people can do little to nothing.

Humans are incomprehensible in ways that algorithms are not. Our reasons for our actions alter and are developed after the fact. To assess human racial prejudice, we must build controlled environments where only race differentiates. We can develop an equally controlled algorithm by feeding it the correct inputs and observing its behaviour. While there is some truth in the assumption that we can more readily audit and thereby eliminate the biases of AI systems when compared to individual human decision-makers, there are several issues that this reasoning overlooks. In theory, biased algorithms are easier to repair, but in practice, nearly no researchers, much alone anyone harmed by these systems, have access to them [5]. Indeed, we rely on the goodness of the organisations implementing these systems to ensure that they do not produce biased results. Furthermore, even when researchers or auditors have access, enhancing this system is a complex statistical and sociological problem.

We must recognise this information gap and enact legislation to impose transparency measures to make AI systems auditable. There are numerous ideas for these transparency measures, ranging from public AI system registers to documentation approaches such as Datasheets for Datasets and Model Cards for Model Reporting. It becomes apparent that AI systems will be incapable of avoiding detrimental bias in whatever form such controls take. Furthermore, they will not help in reducing prejudice in human discrimination if they are kept concealed from examination and audit due to worries about trade secrets.

On the other hand, transparency will not be a cure-all: it is possible that systems are unfixable and should not be employed in certain situations. Using an AI system to address an issue is not a neutral option; it comes with several risks and externalities. Although several companies have created technical methods to detect and prevent prejudice, these models continue to be hampered by the challenge of selecting what to aim for instead of bias. Either aim to maximise fairness' or prioritise other metrics, such as justice.

Understanding and assessing "fairness" is one of the most challenging steps. Various technical definitions of fairness have been proposed by researchers, such as demanding that models have similar predictive values across groups or that models have identical false positive and false negative rates between groups. However, this poses a substantial issue because the models cannot achieve several fairness standards concurrently. Examining the politics behind these definitions reveals that no single term is unequivocally correct and that complex political issues influence our choice. Diverse formalisations of fairness requirements are incompatible, demonstrating the need for trade-offs between different criteria.

An interactive game built by MIT Technology Review [2] that replicates a judicial algorithm to assess whether a criminal should be given bail is an excellent tool for demonstrating this trade-off dilemma. It is a good illustration of the difficult trade-offs that we must make when attempting to attain fairness in these systems, and it begs whether fairness is a suitable paradigm. It seems impossible to please everyone and satisfy all requirements. However, we can try and go beyond the limited lens of algorithm tuning to determine if we can address the actual problems and inequities through other means.

Aside from disputes regarding bias mitigation, several people have pointed out the limitations of framing the harms caused by AI systems as bias and fairness issues. The terms bias and stereotypes are rooted in the theory that emphasises individual perception rather than structural oppression. As a result, when discussing algorithmic prejudice by using the terminology of bias, we may wind up overly focused on the individual intentions of the engineers engaged rather than the structural power of the institutions to which they belong.

There are limitations of fairness solutions, i.e. optimisation systems, systems that interact with the environment in which they are deployed and optimise over constantly changing variables. In addition to bias difficulties, such systems generate significant externalities: situations in which the activities of some groups have severe consequences for people outside of that group. Some systems optimise for majorities but negatively affect minority users. For example, navigation apps such as Waze and Google Maps that maximise travel time for app users guide heavy traffic onto generally quiet residential streets.

Facial recognition systems perform poorly on darker-skinned people, emphasising that an unbiased algorithm can yet have unjust outcomes or externalities. After all, even the most "absolutely impartial" facial recognition system can be exploited by a racist police force or a government that seeks to oppress minorities. Striving for fairness can, in some situations, aggravate the issue. For example, suppose we ensure "fairness " for a model developed to determine "creditworthiness" and trained to maximise profit. In that case, it may result in the model authorising predatory subprime loans, resulting in unjust outcomes.

To effectively remedy the problems created by these systems, we must approach them from a perspective that considers the whole spectrum of historical and sociological factors relevant to each situation. Without historical or sociological depth, computational depth is just shallow learning. Deep learning that is antisocial can capture and contain, but it can also damage individuals. A historically and sociologically informed can provide new perspectives and is capable of generating new settings. It can encode new values and build on crucial intellectual traditions that have consistently created justice-based ideas and solutions. While it is evident that we must do all possible to keep present AI systems from inflicting harm, we must also continue to question the need for AI systems in all sectors in order to combat AI inevitability [1].

Bias and discrimination are societal and human issues. Replacing poor human decision making with AI systems will not solve the problems' core causes and will likely shift these social and political issues into technological battles. When these challenges are framed as technical issues of bias mitigation, non-technical perspectives are excluded from the discussion, which often means omitting the individuals who are most affected by these systems.

In conclusion, AI systems are not objective, and biases can skew their predictions and calculations in various ways. While it is crucial to enhance the models we use and guarantee that they do not duplicate or create biased and discriminatory outcomes, it is also necessary to question their existence and necessity. Even achieving flawless algorithmic justice in AI systems will not solve the complicated social and political difficulties we confront.

References

[1] Ruha Benjamin. *Race after technology: abolitionist tools for the new Jim code*. Medford, MA: Polity, 2019. ISBN: 9781509526390 9781509526406.

[2] *Can you make AI fairer than a judge? Play our courtroom algorithm game*. en. URL: https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/.

[3] Jon Kleinberg et al. *Discrimination in the Age of Algorithms*. en. SSRN Scholarly Paper ID 3329669. Rochester, NY: Social Science Research Network, Feb. 2019. URL: https://papers.ssrn.com/abstract=3329669.

[4] James Vincent. *Twitter taught Microsoft's friendly AI chatbot to be a racist asshole in less than a day*. en. Mar. 2016. URL: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

[5] *Weapons of Math Destruction*. en. Page Version ID: 1081667593. Apr. 2022. URL: https://en.wikipedia.org/w/index.php?title=Weapons_of_Math_Destruction&oldid=1081667593.