BIASES IN THE DEVELOPMENT OF ARTIFICIAL INTELLIGENCE ALGORITHMS

Rudolf Petrušić[1]
[1]Internacionalni univerzitet Travnik u Travniku, Aleja Konzula Meljanca bb Bosna i Hercegovina
e-mail: rudolf.petrusic@iu-travnik.com

Review article
https://doi.org/10.58952/nit20241202043
UDK/UDC 004.4:004.6:006.3

Abstract

One of the major issues that can have far-reaching consequences on society is bias in the development of AI algorithms. Algorithms that cannot guarantee fairness result in discrimination and inequality, and this is very evident in sectors such as healthcare, the legal system, and finance. This paper discusses the different biases arising in AI systems, their causes, impacts, and possible mitigation strategies. In the interest of fair application of AI technology, we focus on ethical dilemmas and strategies that include more transparency and diversity in data and algorithms.

Keywords: *biases, development, artificial intelligence, algorithms*

1 INTRODUCTION

Artificial Intelligence, or AI for short, is the field of computing that focuses on creating computers that can perform tasks that typically would require human intelligence, such as recognizing speech, making decisions, and recognizing pictures and languages.

It was in the 1950s, by pioneers like Alan Turing, that thinking machines entered the hypothesis stage, paving the way for the commencement of artificial intelligence development to date. Throughout the years, artificial intelligence has developed through several stages: expert systems, machine learning, up to the dominant deep learning. (Hinton et al., 2012).

Learning about the issue of bias: Even though artificial intelligence has the potential to improve many aspects of life, the existence of algorithms becomes a big issue. Algorithms can reflect and even amplify societal inclinations because of a lack of variety in the data or due to design flaws. (O'Neil, 2016).

2 TYPES OF BIAS IN AI

The data that serves as the basis for training artificial intelligence models often contains inherent biases that can arise from historical inequalities, stereotypes, and discrimination. Namely, the data used to train recruitment algorithms can favor certain demographic groups, while marginalized groups are ignored. (Barocas & Hardt, 2019).

Algorithmic bias

Algorithmic bias occurs when the model itself, even if the data is neutral, does not make fair decisions. This can happen because of the way the algorithms are designed or implemented. For example, algorithms that use inappropriate metrics or are optimized for efficiency may ignore important factors such as fairness or balance (Angwin et al., 2016).

Model training bias

One of the key sources of availability in AI systems is the training model phase, which depends on the quality and representativeness of the data. If a model is not trained on sufficiently diverse data, its ability to make accurate and fair decisions is seriously compromised (Sweeney, 2013).

Effect of bias on results

Ethnic, gender and other social prejudices. One of the most prominent problems associated with the presence of AI is discrimination based on ethnicity, gender, age and other social factors. For example, in facial recognition systems, algorithms have shown a tendency to be more accurate in recognizing white faces compared to black faces. (Buolamwini & Gebru, 2018)**.**

Effects in real AI applications

Bias in AI can have serious consequences in the real world. In the justice system, algorithms that assess the likelihood of recidivism can be ethnically biased, leading to unfair convictions. Similarly, in financial applications, AI can unfairly favor certain users in accessing loans, based on biased historical data. (Angwin et al., 2016).

Bias across industries

Health care

In healthcare, accessibility can have serious consequences for diagnoses and treatments. For example, algorithms that analyze medical images may be less accurate in identifying diseases in minority patients due to a lack of representation in the training data. (Obermeyer et al., 2019).

Judiciary

In justice systems, algorithms for predicting the risk of reoffending, such as COMPAS, face criticism for ethnic bias that can affect the sentencing and rehabilitation of suspects (Angwin et al., 2016).

Fintech

In the financial industry, AI is used to make lending decisions and assess risk. Bias in these systems can lead to discrimination against certain social groups, especially those with lower incomes or weaker access to capital markets (Huang et al., 2020).

3 METHODS FOR REDUCING BIAS IN AI

Transparency

One of the key ways to reduce availability in AI is to increase transparency in the development process. This includes a clear understanding of the data used, the training model, and the metrics applied to evaluate performance (Raji & Buolamwini, 2019)**.**

Diversity of data

One of the most powerful tools for combating presence is the use of diverse representative datasets. This ensures that the AI system can handle a wide range of users and situations (Mehrabi et al., 2019).

Adjustment of algorithms

Another strategy is to adjust the algorithms themselves so that they are less sensitive to the social and demographic factors that can drive attraction. This is achieved by developing new techniques for proper learning and data balancing (Zemel et al., 2013). they understand how algorithms work and on the basis of which data they make decisions. Without adequate transparency, it is difficult to assess whether algorithms take into account relevant factors and whether there is a possibility of discrimination based on race, gender, ethnicity or other demographic characteristics (Pasquale, 2015).

AI ethics also requires accountability in system architecture. Any biases, either in the algorithms or in the data used, should be considered by researchers and developers. For example, an algorithm could be innocently designed but still support and extend inequalities in society if it has been trained on historical data reflecting those very inequalities. The concept of design ethics in the development of systems in AI is important to be brought forth to ensure technologies are not only serving the interests of the most powerful, but also contributing towards building a more equitable society.

Because of this complexity, ethical approaches to artificial intelligence involve a wide range of issues, including fairness, privacy, accountability, and security. Each of these areas requires detailed consideration and, most importantly, the active cooperation of technology experts, legislators, regulators, and technology users themselves...

4 THE NEED FOR REGULATIONS AND STANDARDS

It is high time that suitable laws and uniform standards be established, considering the increased impact of artificial intelligence in daily life, so as to ensure its equitable, responsible, and secure application. The legislative framework regarding AI should include rights such as privacy, data protection, justice, nondiscrimination, and culpability in case of harm arising out of the use of AI systems. However, the regulation of AI is complicated, as it has to be balanced between the advancements that make the technology possible and the user protection that's required.

Establishing suitable laws and uniform standards is imperative in light of artificial intelligence's increasing impact on daily life in order to ensure its equitable, responsible, and secure application. The rights to privacy, data protection, justice, nondiscrimination, and culpability for potential harm resulting from the usage of AI systems should all be covered by the legislative framework for AI. Regulating AI is a complicated topic, though, because it requires striking a balance between the advancements that make the technology possible and the user protection that is required. This regulation also sets strict guidelines for auditing, testing and mandatory certifications, as well as a clear division of responsibilities between different actors - from designers to end users.

In a similar effort, the OECD (Organization for Economic Co-operation and Development) has developed guidelines for the responsible use of artificial intelligence, which include recommendations regarding transparency, training people, avoiding discrimination and preserving human rights. The OECD called for international cooperation in the development of ethical standards, with the aim of minimizing the negative effects of artificial intelligence, while at the same time encouraging innovation and economic growth (OECD, 2019).

Globally, in 2021, UNESCO adopted the Recommendations on Ethics in Artificial Intelligence, which provide a framework to guide countries in developing responsible and fair AI policies. These recommendations call for global cooperation and exchange of best practices in managing the ethical challenges of artificial intelligence, as well as the inclusion of different social groups in the policy development process, in order to ensure equitable access to technologies.

Challenges in the implementation of regulations

Although there is significant progress in the development of regulatory frameworks, their implementation and alignment with global standards is a major challenge. Different states have different approaches to regulating AI. While the EU favors strict regulation, there is skepticism in the United States about excessive regulation, which they believe could slow down innovation. There are also challenges related to global standards, as AI technology expands rapidly and does not respect borders. This means that every regulation should be harmonized with international laws and practices in order to avoid contradictions and disagreements between countries.

Because of these challenges, many experts argue that regulations should be flexible and able to adapt quickly to rapid technological advances. Also, there is a need for international cooperation to develop global standards, as suggested by many initiatives such as the Global Partnership for Artificial Intelligence (GPAI), which brings together countries from around the world in developing the responsible use of artificial intelligence.

User rights and protection against damage

In addition to the regulations related to the technical aspects of AI, it is important to pay attention to the rights of users. Users who rely on AI systems, whether for financial institution services, healthcare or educational services, have the right to be protected from potential damages that may arise as a result of unfair or discriminatory algorithmic decisions. Also, it is important to ensure that users have access to information about how their data is used, how decisions are made and what are the criteria for making those decisions. This information should be easily accessible, understandable and transparent**.**

Education and community engagement

One of the key factors in creating responsible artificial intelligence is education and community awareness. The development of educational programs that also cover the ethical aspects of artificial intelligence is crucial to ensure that future experts and decision makers are AWARE of the impact of their technologies on society. It is also important to involve the wider community in the decision-making process, allowing citizens to have a voice in the regulation of artificial intelligence, which can ensure a wider representation of interests and values.

CONCLUSION

Bias in the development of artificial intelligence (AI) algorithms is a serious challenge that goes beyond the technical aspects of technology development. AI is developing rapidly and is being integrated into almost every aspect of our lives, including health, education, employment, justice and finance. While these technologies have great benefits in terms of efficiency, productivity and innovation, they also carry the risk of profound social, economic and political inequalities if not properly designed and implemented.

Large data sets are employed by AI systems, which often inherit and further propagate societal biases. For example, facial recognition algorithms related to security have exhibited high error rates in recognizing faces with darker skin tones, thus often causing harm to the rights of minorities (Buolamwini & Gebru, 2018). Similarly, credit-scoring algorithms in the financial domain may favor specific demographic groups more than others, thereby marginalizing the already underprivileged existing populations even more than before, and this according to O'Neil 2016. These examples now show that in artificial intelligence, access is a multidimensional issue influenced by social norms and practices, which are, in turn, likely to be furthered unconsciously by algorithms.

Irrespective of these challenges, it is important to underline that the development of responsible and innovative AI technology is possible, desirable, and indeed necessary. Therein lies a great deal of responsibility with users, lawmakers, regulators, and inventors of artificial intelligence systems. Efforts such as those developed by the OECD and UNESCO, and regulations like the EU AI Act, which provides clear-cut rules on high-risk AI systems, therefore stand to provide a guideline for the ethical handling of AI in an effort to reduce the desirability of the technology and abuse. These regulations also promote transparency, accountability and fairness, ensuring that AI is used for the benefit of all social groups, regardless of their race, gender, ethnicity or socioeconomic status.

However, international cooperation is essential to the development of valid legislation. Because AI knows no borders, international standards and guidelines should be harmonized to avoid legal conflicts and ensure users' safety around the globe. Involvement of all relevant stakeholders in the processes of regulatory and technological developments is also important. That includes the users themselves, who are to be informed and allowed to be a part of the decision-making on AI use, and experts from other areas, like ethics, law, and sociology. This would also mean education and training on the responsible use and development of technology for the future engineer in AI, policymakers, and the general public, if such a society is ever to be created.

In addition to regulation, the focus should also be on preventive measures that will enable recognition and removal of accessibility already in the design and training phase of artificial intelligence systems. One way to achieve this goal is to develop and implement tools for auditing algorithms, such as systems for checking compliance with ethical principles and testing AI systems for potential discrimination before they are put into use. Such tools make it possible to identify weaknesses in the system and provide an opportunity for improvement before access becomes a serious problem

Bias in artificial intelligence is not a problem that can be solved by technical fixes or a uniform legislative framework alone. It is a long-term challenge that requires cooperation among all sectors of society – from academia and industry to political leaders and users themselves. Only by working together, directing technological development towards fairness and equality, can we ensure that AI technology becomes a tool that benefits all people, regardless of their personal characteristics.

In conclusion, bias in the development of artificial intelligence is a problem that must be solved through a multidisciplinary approach, with an emphasis on ethical guidelines, transparency and accountability. Considering the potential that artificial intelligence has in shaping the future, it is important that in this process all risks of discrimination are recognized and minimized, so that the technology serves the well-being of the whole society.

REFERENCE

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica.

Barocas, S., & Hardt, M. (2019). Fairness and Machine Learning. Cambridge University Press.

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency (FAT*).

Crawford, K. (2021). Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence. Yale University Press.

European Commission (2021). Proposal for a Regulation on Artificial Intelligence. https://ec.europa.eu

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A. R., & Jaitly, N. (2012). Deep neural networks for acoustic modeling in speech recognition. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 1-10.

Huang, K. W., & Li, Z. (2020). The Unintended Consequences of Machine Learning in Finance. Journal of Finance, 75(6), 2587-2626.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys, 52(1), 1-35.

O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases

Inequality and Threatens Democracy. Crown Publishing.

Obermeyer, Z., Powers, B. W., Vogeli, C., & Mullainathan, S. (2019). Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science, 366(6464), 447-453.

Russell, S. J. (2016). Artificial Intelligence: A Modern Approach. Pearson.

Raji, I. D., & Buolamwini, J. (2019). Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

Sweeney, L. (2013). Discrimination in Online Ad Delivery. Communications of the ACM, 56(5), 44-54.

Zemel, R. S., Wu, Y., Yu, T., & Weinberger, K. Q. (2013). Learning Fair Representations. Proceedings of the 30th International Conference on Machine Learning.