United Nations University

TECHNOLOGYBRIEF
No. 4, FEBRUARY 2024

Avoidable and Unavoidable Algorithmic Bias

Tshilidzi Marwala, United Nations University, Tokyo, Japan

Policy recommendations

Avoidable algorithmic bias

1. Maximize the diversity of AI systems' training data to accurately represent diverse populations.
2. Maximize algorithmic transparency to allow inspection of decision-making processes.
3. Perform routine audits of AI systems to detect and address biases, with the outcomes publicly disclosed to ensure accountability.
4. Ensure that inclusive AI development teams incorporate diverse viewpoints when designing AI systems.

Unavoidable algorithmic bias

5. Ensure stakeholders are fully informed about AI systems' inherent limitations, including any existing unavoidable bias and the steps to mitigate it.
6. Establish policy frameworks that explain acceptable trade-offs and decision-making criteria when bias is unavoidable.
7. Establish independent ethical oversight committees to evaluate and authorize AI systems in critical situations, ensuring that unavoidable bias is addressed ethically.
8. Continuously monitor and improve AI algorithms to align with evolving cultural norms.

Introduction

Algorithmic bias has become a significant issue in the fast-growing field of artificial intelligence (AI) and machine learning.[1,2] Even when it is unintentional, algorithmic bias can appear in several ways, resulting in discriminatory outcomes that unjustly disadvantage particular individuals and groups.

Understanding the distinction between avoidable and unavoidable algorithmic bias will become increasingly important for policymakers, developers and consumers as they navigate the evolving relationship between technology and ethical norms. Interestingly, studies suggest that people are currently less troubled by algorithmic bias than by human bias, but this attitude may change over time.[3]

Avoidable algorithmic biases are tendencies that can be reduced through meticulous system design, data analysis and inclusive development.[4] Avoidable biases can be caused by skewed data sets, faulty algorithmic design and negligence in the development process. Addressing these biases requires a proactive strategy that emphasizes diversity, openness and continual oversight in deploying AI systems.

Unavoidable algorithmic bias, in contrast, refers to bias that is difficult to eliminate due to conflicting fairness principles, the complexity of the data or the limitations of current AI technology. These biases result in complex ethical and practical difficulties, requiring a careful equilibrium between conflicting interests and acknowledging that a certain degree of bias may endure despite earnest attempts.

Addressing algorithmic bias, whether avoidable or unavoidable, is both a technical problem and a societal obligation. Policymakers, engineers, ethicists and society must work together to ensure that AI systems possess intelligence, efficiency and fairness. At this pivotal moment in technological advancement, our policy decisions can shape the long-term impact of AI on our societies.

Causes of avoidable algorithmic bias

Algorithmic biases will become increasingly significant as decision-making processes become more automated.[5,6] These biases arise from limited data collection, assumptions during algorithm design and insufficient consideration of how diverse groups may be treated by an algorithm.[7]

For example, if an AI system is trained using historical hiring data that mirror previous discriminatory practices, the system may perpetuate similar prejudices by favouring specific demographic groups and discriminating against others. As AI becomes more prevalent in different aspects of society, it is crucial to identify and remove algorithmic biases that can be avoided to promote trust, justice and equality in automated systems.

Strategies for avoiding bias

A proactive strategy that focuses on transparency, diversity and ongoing monitoring is necessary to address avoidable algorithmic bias.[8] First and foremost, it is essential to ensure that the data utilized for training AI systems are comprehensive, diverse and include all relevant demographic segments. Design teams should be multidisciplinary — composed of ethicists, sociologists, and domain specialists that can offer a range of perspectives and aid in the early detection of potential biases.

Ensuring transparency of AI algorithms is also crucial. By making the decision-making processes of AI systems accessible and explainable, biases may be identified and resolved. However, a technologic breakthrough may be required to solve the accuracy versus transparency dilemma, wherein — under current AI systems — the more accurate an algorithm is, the less transparent it is (and vice-versa).[9]

Periodic audits and updates of AI systems are also vital to consistently monitor for biases and prejudice. It is essential to have a robust legal and ethical framework that establishes explicit norms and principles for the development and implementation of AI. By cultivating a culture that prioritizes accountability and inclusiveness, we may effectively leverage the advantages of AI technologies while substantially mitigating the potential risks of preventable algorithmic bias.

Not all bias can be eliminated

Some level of algorithmic bias is inherent in the interplay of data, AI technology and human behaviour, so it cannot be eradicated entirely; at best, it can be minimized.[10] The enduring nature of some biases can be attributed to the manifestation of societal prejudices in historical data, from which AI systems acquire knowledge, unintentionally reinforcing preexisting biases. As noted above, significant progress can be made through data curation, transparent algorithmic design and ongoing monitoring, but the only realistic objective is to reduce the extent of bias, rather than eliminate it.

Acknowledging this constraint is essential for promoting a more responsible and conscientious approach to AI systems. The emphasis should be on continuous improvement and the diligent reduction of bias, while recognizing that it is usually not possible to achieve fully unbiased algorithms.

Dealing with unavoidable algorithmic bias

While some algorithmic bias can be avoided, a level of residual bias will be unavoidable due to competing moral frameworks and the deeply rooted structures of our cultural and technical systems. These inherent biases arise from the complexities of social phenomena, the diversified concept of justice, and always-evolving societal norms. The idea of fairness, which has existed since the dawn of humanity, is inherently subjective. Different groups and individuals have diverse perspectives on what constitutes fairness. Algorithms often encounter conflicting concepts as they strive for justice. The pursuit of perfect equity can seem impossible, as optimizing for one type of fairness may inadvertently result in biases towards another.

Moreover, societal norms undergo perpetual evolution.[11, 12] An algorithm that currently embodies fairness may eventually become biased because of changes in societal norms. The ever-changing landscape of this terrain turns the pursuit of equity into an ongoing expedition rather than a final goal, a constant progression rather than a solitary accomplishment.

To tackle inherent biases, we need to fundamentally change our approach to ensuring algorithmic fairness. Rather than a singular resolution, we must think in terms of a flexible, ongoing procedure. We must continuously monitor and improve algorithms to align with evolving cultural norms and understandings of fairness.

The pharmaceutical side effect approach to algorithmic bias

Examining the pharmaceutical industry's strategy for handling side effects can provide a helpful perspective on tackling algorithmic bias.[13] Within the pharmaceutical field, it is widely recognized that no medication is entirely devoid of adverse effects. This recognition is analogous to the notion that no algorithm can be utterly devoid of bias. Like drugs, AI systems should undergo meticulous testing phases to detect, measure and mitigate potential adverse effects before their release. AI systems require comprehensive examination and ongoing supervision to uncover and reduce biases.

The principle of informed consent in medicine, which involves informing patients about possible adverse consequences, has a parallel in the context of AI, known as algorithmic transparency.[14] This requires that users are provided with information regarding the decision-making process of AI systems and the potential biases that these systems may have.

Furthermore, akin to pharmacovigilance, which entails continuous monitoring of pharmaceuticals' real-world performance after they are on the market, AI systems also require comparable processes to acquire knowledge, and adjust and enhance their capabilities consistently.[15] This is crucial to detect and rectify any biases that may arise. This approach emphasizes the significance of a proactive, transparent and ethically based framework in addressing algorithmic bias.

Procedure for handling algorithmic bias

A proposed procedure for addressing algorithmic bias is outlined in **Figure 1**. The first question is whether a trained AI model is biased. If the answer is negative, then the AI model is deployed. If the answer is positive, the next question is whether the algorithmic bias is avoidable or not.

If the answer is positive, then mechanisms to minimize algorithmic bias are implemented, and the residual bias is quantified to be reported to the user during the deployment. If the answer is negative, the extent of bias is quantified and the AI model is deployed with a report of the results to the user.

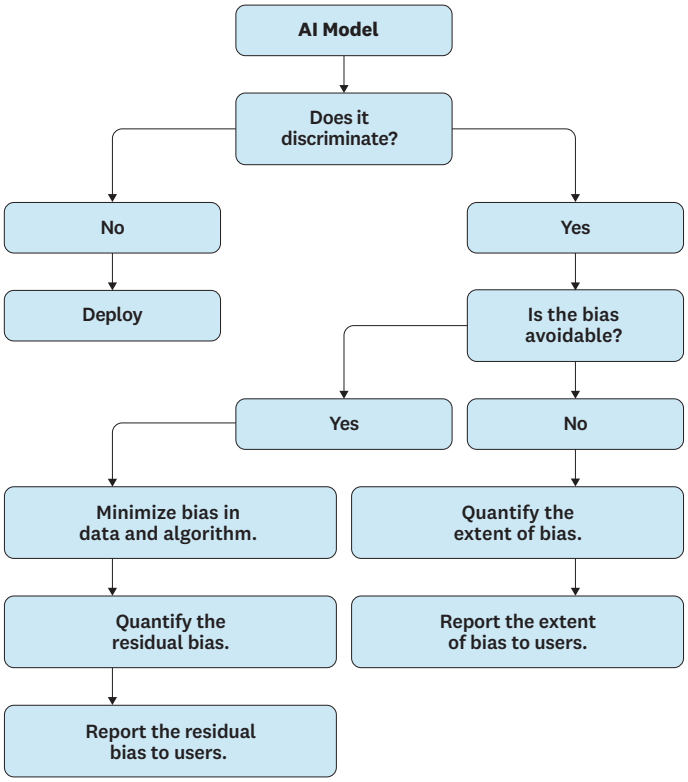

Figure 1 Procedure for handling avoidable and unavoidable algorithmic bias

Conclusion

As AI brings about significant changes in society, addressing both avoidable and unavoidable algorithmic biases will become increasingly important. By differentiating between these two types of bias — and applying specific strategies to address each type — policymakers can help create AI systems that are intelligent, efficient and just. A balanced approach is essential to upholding public confidence in AI and fully using its potential to advance society.

ENDNOTES

1 T. Marwala (2023) Algorithm Bias — Synthetic Data Should Be Option of Last Resort When Training AI Systems. Daily Maverick.
https://unu.edu/article/algorithm-bias-synthetic-data-should-be-option-last-resort-when-training-ai-systems

2 T. Marwala (2024) The Dual Faces of Algorithmic Bias — Avoidable and Unavoidable Discrimination. Daily Maverick.
https://unu.edu/article/dual-faces-algorithmic-bias-avoidable-and-unavoidable-discrimination

3 Bigman, Y.E., Wilson, D., Arnestad, M.N., Waytz, A. and Gray, K., 2023. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), p.4.

4 Kordzadeh, N. and Ghasemaghaei, M., 2022. Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems, 31*(3), pp.388-409.

5 Breidbach, C.F., 2024. Responsible algorithmic decision-making. *Organizational Dynamics*, p.101031.

6 Obermeyer, Z., Nissan, R., Stern, M., Eaneff, S., Bembeneck, E.J. and Mullainathan, S., 2021. Algorithmic bias playbook. *Center for Applied AI at Chicago Booth*.

7 Marwala, T., 2014. Missing Data Approaches for Rational Decision Making: Application to Antenatal Data. Artificial Intelligence Techniques for Rational Decision Making, pp.55-71.

8 Hastings, J., 2024. Preventing harm from non-conscious bias in medical generative AI. *The Lancet Digital Health, 6*(1), pp.e2-e3.

9 von Eschenbach, W.J., 2021. Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*(4), pp.1607-1622.

10 Aysolmaz, B., Iren, D. and Dau, N., 2020. Preventing algorithmic Bias in the development of algorithmic decision-making systems: A Delphi study.

11 Acemoglu, D. and Jackson, M.O., 2015. History, expectations, and leadership in the evolution of social norms. *The Review of Economic Studies, 82*(2), pp.423-456.

12 Young, H.P., 2015. The evolution of social norms. *economics, 7*(1), pp.359-387.

13 Belenguer, L., 2022. AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics, 2*(4), pp.771-787.

14 Nijhawan, L.P., Janodia, M.D., Muddukrishna, B.S., Bhat, K.M., Bairy, K.L., Udupa, N. and Musmade, P.B., 2013. Informed consent: Issues and challenges. *Journal of advanced pharmaceutical technology & research, 4*(3), p.134.

15 Trenque, A., Rabiaza, A., Fedrizzi, S., Chretien, B., Sassier, M., Morello, R., Alexandre, J. and Humbert, X., 2024. Evaluation of a simplified pharmacovigilance tool for general practitioners: 5 years of insight. *Scientific Reports, 14*(1), p.1766.

EDITORIAL INFORMATION

About the research
This technology brief is part of a UNU series highlighting specific areas of global technology governance related to the Global South and sustainable development.

Author biographies
Professor Tshilidzi Marwala is the Rector of United Nations University, headquartered in Tokyo, and Under-Secretary-General of the United Nations. He was previously the Vice-Chancellor and Principal of the University of Johannesburg. Prof. Marwala has published over 300 research papers and articles, over 250 articles in newspapers and magazines, 27 books on AI and related topics, and holds five patents. He is a member of the American Academy of Arts and Sciences, the Chinese Academy of Sciences, the World Academy of Sciences (TWAS) and the African Academy of Science.

Citation
Tshilidzi Marwala, "*Avoidable and Unavoidable Algorithmic Bias*", UNU Technology Brief 4 (Tokyo: United Nations University, 2024).

ISBN 978-92-808-9152-2