E-ISSN: 1115 – 8492 P-ISSN: 3027 – 0693
MPRI-JNSAR
INTERNATIONAL JOURNAL OF NATURE AND SCIENCE ADVANCE RES.
VOL. 10 NO. 9 OCTOBER-2025

PROMOTING ETHICAL AI, STRATEGIES FOR MITIGATING BIAS IN MACHINE LEARNING MODELS

***MAYOWA AROWOSAFE; **MUHAMMAD BELLO; ***OLATUNDE AWOLOLA; & ****MICHAEL BINUYO**

*Department of Electrical and Computer Engineering, University of Windsor, Ontario, Canada.
**Department of Computer Science, Ulster University, United Kingdom. ***School of Health Information Management, University College Hospital, Ibadan, Nigeria. ****Department of Computer Science, University of Hertfordshire, United Kingdom

Corresponding Author: arowosam@uwindsor.ca

DOI: https://doi.org/10.70382/mejnsar.v10i9.073

Abstract

As artificial intelligence (AI) becomes an increasingly integral part of our daily lives, the need for ethical AI practices, particularly in addressing bias, has never been more crucial. This article provides a closer examination of the complex issue of bias in AI, examining how it arises from data, algorithms, and human decisions. It underscores why fairness, transparency, and accountability are essential to ensure that AI systems deliver fair outcomes for everyone. The article also discusses practical strategies for detecting and reducing bias, such as sourcing diverse data, applying fairness constraints during algorithm design, and using fairness metrics to evaluate models. Beyond technical solutions, it highlights the importance of involving stakeholders and complying with regulations to guide the ethical development of AI. By tackling these challenges, the article aims to support the creation of AI technologies that are not only innovative but also fair and responsible benefiting society as a whole.

Keywords: Ethical AI, Bias, Machine Learning, Artificial Intelligence (AI), and Data Preprocessing

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Introduction

Artificial intelligence (AI) has been advancing at an incredible pace in recent years, reshaping everything from healthcare to finance and pretty much everything in between. But with all these breakthroughs comes a big responsibility, that is, making sure AI is not just powerful but also ethical and fair.

Ethical AI is all about developing and using AI in a way that aligns with our values, making sure it's transparent, accountable, and free from harmful biases. It's a big deal because, without it, people won't trust AI systems, and that trust is key to making AI work for everyone (Floridi et al., 2018).

Bias in AI and machine learning is a serious issue. It can lead to unfair treatment and even discrimination against certain groups. This bias can creep in from multiple sources, like the data used to train models, the algorithms themselves, or even human choices made during AI development. If we don't tackle it, AI could end up reinforcing existing inequalities and making social injustices worse (Barocas et al., 2019). Take hiring systems, for example. If an AI is biased, it might unfairly reject candidates from underrepresented backgrounds. Or think about facial recognition. Studies have shown that these systems often make more mistakes when identifying people with darker skin tones (Buolamwini & Gebru, 2018). Fixing bias in AI isn't just a technical challenge, it's a moral responsibility.

Ethical AI means ensuring these technologies promote fairness and inclusivity rather than deepening divides. That's why we need strong strategies to detect, reduce, and monitor bias at every stage of AI development, from data collection to deployment. As AI plays a bigger role in critical decisions, prioritizing ethics isn't optional, but it is essential to creating fair outcomes for everyone.

Purpose of Research

This research aims to tackle the critical issue of bias in machine learning models, with the bigger goal of promoting ethical AI practices. As AI becomes more embedded in everyday life, it's crucial to ensure these systems are fair, transparent, and aligned with societal values. This study explores where bias in AI comes from, how it affects marginalized and vulnerable communities, and how well existing mitigation techniques actually work. Beyond that, it contributes to the ongoing conversation on ethical AI by offering recommendations to improve transparency, fairness, and regulatory compliance in machine learning. Finally, this research advocates for the responsible development and deployment of AI systems, the ones that promote social equity and minimize harm.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Types of Bias

Data Bias: Data bias occurs when the datasets used to train machine learning models are not representative of the real-world scenarios they are intended to reflect. This can happen due to historical inequalities, sampling errors, or lack of diversity in the data collection process. For instance, if a facial recognition system is trained predominantly on images of lighter-skinned individuals, it may perform poorly on individuals with darker skin tones, leading to higher error rates and potential discrimination (Buolamwini & Gebru, 2018).

Algorithmic Bias: Algorithmic bias arises from the design and implementation of the algorithms themselves. This type of bias can occur when algorithms inadvertently favor certain outcomes over others due to their inherent structure or the way they process data. For example, a hiring algorithm that prioritizes candidates based on historical hiring data may perpetuate existing gender or racial biases if those biases were present in the original data (Raghavan et al., 2020).

Human Bias: Human bias is introduced by the decisions and judgments made by individuals involved in the AI development process. This can include biases in selecting features, labeling data, or setting model parameters. Human biases can be subtle and often reflect societal prejudices, which can be inadvertently encoded into AI systems (Holstein et al., 2019).

Principles of Ethical AI

Ethical AI principles act as essential guidelines to ensure that AI technologies are built and used in ways that reflect societal values and moral standards. These principles focus on promoting fairness, transparency, accountability, and inclusivity. It helps to build trust and make sure AI benefits everyone. Let us consider a few of them.

- **Fairness**: AI systems should operate without favoritism or prejudice, ensuring equitable treatment for all individuals (Hort et al., 2022). This involves actively identifying and mitigating biases that may arise from skewed data or algorithmic processes. For instance, in hiring practices, AI should evaluate candidates based on merit, free from biases related to gender, race, or background.
- **Transparency:** Transparency is crucial for building trust in AI systems. It involves making AI processes and decision-making understandable and accessible to users and stakeholders. This includes providing clear explanations of how AI models work, the data they use, and the factors influencing their decisions (Lumenalta, n.d.). Transparency helps users make informed decisions and holds AI developers accountable for their systems.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

- **Accountability**: Organizations deploying AI must take responsibility for their systems' actions. This means establishing mechanisms to monitor AI performance, address errors, and rectify any harm caused. Accountability ensures that AI remains a tool for societal benefit, with clear recourse in cases of misuse.
- **Privacy and Security:** Protecting user privacy and ensuring data security are fundamental ethical considerations in AI. This principle involves implementing robust measures to safeguard personal data and prevent unauthorized access or misuse. AI systems should be designed to respect user privacy and comply with relevant data protection regulations.
- **Inclusivity:** Inclusivity in AI emphasizes the importance of involving diverse perspectives in the development and deployment of AI systems. This principle advocates for engaging stakeholders from various backgrounds, including underrepresented and marginalized communities, to ensure that AI technologies address the needs and concerns of all users.
- **Human-Centric Design:** AI systems should be designed with a focus on enhancing human well-being and autonomy. This principle involves creating AI technologies that complement human capabilities, support human decision-making, and prioritize user experience and satisfaction.
- **Sustainability:** Ethical AI also considers the environmental impact of AI technologies. This principle encourages the development of energy-efficient AI systems and the responsible use of resources to minimize the ecological footprint of AI deployments.

The Impact of AI Bias on Society

Bias in AI systems can have serious and far-reaching consequences. One major concern is social justice-biased AI can reinforce existing inequalities in society. For example, if a loan approval system is trained on historical data that reflects past lending discrimination, it may continue denying loans to marginalized groups, deepening economic disparities. Another critical issue is the legal and regulatory challenges that come with biased AI. Even if unintentional, discriminatory practices can violate anti-discrimination laws, leading to lawsuits and regulatory penalties. As governments become more aware of these risks, there is a growing push for comprehensive AI regulations to ensure fairness and protect citizens' rights (The Australian, 2025).

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Identifying Bias in Machine Learning Models

Identifying bias in machine learning models is a critical step in ensuring that AI systems operate fairly and equitably. By using various techniques and tools, we can detect and address biases that may emerge during both development and deployment.

Statistical Analysis

One common approach is statistical analysis, which helps uncover bias in datasets by examining data distribution and identifying imbalances or disparities. Techniques like disparity impact analysis and demographic parity can highlight biases that may influence model outcomes (Friedler et al., 2019).

Model Evaluation

Model evaluation is a key way of identifying bias in machine learning models, where fairness is assessed by analyzing performance across different demographic groups. Techniques like cross-validation and subgroup analysis can reveal inconsistencies in how a model treats diverse populations (Binns, 2018).

Tools and Frameworks

To support bias detection and mitigation, several tools and frameworks have been developed. Platforms like Fairness Indicators and AI Fairness 360 provide developers with metrics, visualizations, and algorithms to evaluate and improve fairness in machine-learning models (Bellamy et al., 2019). These tools make it easier to identify and correct biases, helping to build AI systems that are more transparent and inclusive.

Strategies for Mitigating Bias

Mitigating bias in machine learning models requires a multifaceted approach that encompasses data collection, algorithm design, and model evaluation.

- **Data Collection and Preprocessing:** In data collection and preprocessing, ensuring that datasets are diverse and representative of the populations they aim to serve is essential for reducing bias. This means gathering data from a wide range of sources and demographics to accurately capture real-world scenarios (Gebru et al., 2018). Also, various data-cleaning techniques can help minimize bias within datasets. Methods such as data augmentation, re-sampling, and de-biasing are effective in creating balanced datasets that do not disproportionately favor any specific group (Zhang et al., 2018). These approaches play a crucial role in promoting fairness and improving the overall reliability of AI models.

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

- **Algorithm Design:** When designing machine learning algorithms, it is crucial to incorporate fairness constraints during training to prevent biased outcomes. Techniques such as adversarial debiasing and fairness-aware learning help enforce these constraints effectively, ensuring that models treat different groups fairly (Zafar et al., 2017). Additionally, several algorithms have been specifically developed to reduce bias in machine learning models. Methods like re-weighting, re-sampling, and adversarial training work to minimize disparities in predictions and promote fairness across different demographic groups (Kamiran & Calders, 2012). By integrating these techniques, AI systems can become more equitable and inclusive.
- **Post-Processing Techniques:** Post-processing techniques play a crucial role in reducing bias in machine learning models after they have been trained. One effective method is using fairness-aware decision thresholds. By adjusting decision thresholds for different demographic groups, we can help achieve more equitable outcomes. For example, in medical diagnostics, setting different thresholds for various groups can ensure that all patients receive appropriate care recommendations (Buolamwini & Gebru, 2018). Another approach is implementing bias correction models, which adjust primary AI systems' outputs to correct biased predictions. These additional models analyze initial outputs and make necessary modifications to align with fairness criteria, ensuring more balanced and just outcomes (Curvature AI, n.d.).
- **Explainability:** In AI, explainability is essential for building trust and preventing harmful biases. Regulatory frameworks, like the EU's General Data Protection Regulation (GDPR), stress the 'right to explanation,' ensuring that individuals can understand how AI-driven decisions affect them (EU AI Act, 2023). To achieve this, developers can use tools like SHAP, LIME, and interpretable models to make AI systems more accountable and fairer. These techniques help reduce the risks of discrimination and bias, especially in high-stakes fields like healthcare, finance, and law enforcement.

Stakeholder Involvement

Engaging diverse stakeholders in the AI development process is crucial for creating AI systems that are inclusive and aligned with societal values. Stakeholders can include technologists, ethicists, policymakers, industry experts, and representatives from underrepresented and marginalized communities. By involving a wide range of perspectives, AI developers can better understand the potential impacts of their

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

technologies and address the needs and concerns of all users. Stakeholder involvement fosters collaboration and dialogue, ensuring that AI systems are designed with fairness, equity, and inclusivity in mind. This collaborative approach also helps identify potential ethical issues early in the development process and promotes the creation of AI technologies that benefit society as a whole.

Regulatory and Compliance Aspects

The regulatory and compliance landscape for AI ethics is evolving rapidly as governments and organizations recognize the need to establish guidelines and standards for responsible AI development and deployment. Regulations and standards related to AI ethics aim to ensure that AI systems are developed and used in a manner that respects human rights, privacy, and societal values. Key regulatory frameworks, such as the European Union's General Data Protection Regulation (GDPR) and the proposed EU AI Act, set out requirements for transparency, accountability, and fairness in AI systems. Compliance with these regulations is essential for organizations to maintain trust with users and avoid legal and reputational risks. As the field of AI ethics continues to grow, ongoing collaboration between regulators, industry leaders, and stakeholders will be vital in shaping effective and comprehensive regulatory frameworks that promote ethical AI practices globally.

Transparency and Explainability

Transparency and explainability are fundamental components of ethical AI, as they ensure that AI models are not only effective but also understandable and trustworthy. Transparency involves making the processes and decision-making mechanisms of AI systems accessible and clear to users and stakeholders. This includes providing insights into how AI models function, the data they utilize, and the criteria influencing their decisions. Explainability goes a step further by offering detailed explanations of AI outputs, allowing users to comprehend the rationale behind specific predictions or actions. By enhancing transparency and explainability, AI systems can build trust with users, facilitate informed decision-making, and enable accountability for AI developers and organizations.

Conclusion

In the push for ethical AI, addressing bias in machine learning is not just a technical challenge it’s a responsibility. A well-rounded approach is needed, starting with diverse and representative data, effective data-cleaning techniques, and fairness constraints built

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

into algorithm design. Regularly evaluating models with fairness metrics and continuously monitoring AI systems is equally important to ensure they remain fair and unbiased over time. These strategies lay the foundation for AI technologies that are equitable, transparent, and accountable helping to build trust and inclusivity. Looking ahead, the field of ethical AI is set for exciting advancements. Emerging techniques, such as explainable AI and fairness-aware algorithms, offer new ways to enhance transparency and fairness. But technology alone isn't enough collaboration between technologists, ethicists, policymakers, and other stakeholders will be key in shaping the ethical frameworks that guide AI development. As AI becomes more deeply embedded in society, prioritizing ethics is no longer optional it's essential. By embracing these principles and continuing to refine our approaches, we can create AI systems that are not just powerful, but also responsible and fair, ultimately benefiting society as a whole.

References

The Australian. (2025). *A question of ethics: Artificial intelligence faces its most important crossroads.*

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning.* fairmlbook.org.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM Journal of Research and Development, 63*(4/5), 4:1-4:15. https://doi.org/10.1147/JRD.2019.2942287

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, 149-159. https://doi.org/10.1145/3287560.3287582

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1-15.

Curvature AI. (n.d.). *From racist chatbots to wrongful arrests: Understanding AI bias consequences.*

European Commission. (2023). *The artificial intelligence act: Regulatory framework for AI in the EU.*

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689-707. https://doi.org/10.1007/s11023-018-9482-5

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for datasets. *arXiv preprint arXiv:1803.09010.*

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1-16. https://doi.org/10.1145/3290605.3300830

Hort, M., Chen, Z., Zhang, J. M., et al. (2022). Bias mitigation for machine learning classifiers: A comprehensive survey. *arXiv preprint arXiv:2207.07068.*

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]

Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems, 33*(1), 1-33. https://doi.org/10.1007/s10115-011-0463-8

Lumenalta. (n.d.). *Ethical considerations of AI: Fairness, transparency, and frameworks.*

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469-481. https://doi.org/10.1145/3351095.3372828

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 962-970.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335-340. https://doi.org/10.1145/3278721.3278779

EDITOR-IN-CHIEF
Prof. E. Mclenna [Australia]