Bangladesh Journal of Bioethics

Published by: Bangladesh Bioethics Society
https://bjbio.bioethics.org.bd/index.php/BJBio
ISSN: p2226-9231 e 2078-1458
BJBio 2025; 16 (1):8-14
Submitted:26.12.2024
Accepted: 08.02.2025
Published:01.03.2025

Review Article

Tackling Racial Bias in AI Systems: Applying the Bioethical Principle of Justice and Insights from Joy Buolamwini's "Coded Bias" and the "Algorithmic Justice League"

Etaoghene Paul Polo[1] and Donatus Osatofoh Ailodion[2]

https://doi.org/10.62865/bjbio.v16i1.129

Abstract: This paper examines the issue of racial bias in artificial intelligence (AI) through the lens of the bioethical principle of justice, with a focus on Joy Buolamwini's *Coded Bias* and the work of the *Algorithmic Justice League*. AI technologies, particularly facial recognition systems, have been shown to disproportionately misidentify individuals from marginalised racial groups, raising profound ethical concerns about fairness and equity. The bioethical principle of justice stresses the importance of equal treatment and protecting vulnerable populations. Through qualitative research, including content analysis of Buolamwini's works and case studies of AI bias, this paper assesses the efforts of the *Algorithmic Justice League* to combat racial bias in AI. It emphasises their advocacy for developing fair, equitable algorithms and calls for systemic reform in AI development to ensure justice for marginalised communities.

Keywords: AI bias, racial bias, bioethics, justice, Algorithmic Justice League, Coded Bias, Joy Buolamwini, facial recognition, equitable algorithms

Introduction: Artificial intelligence refers to the ability of machines, devices, or apparatuses to mimic natural human intelligence by performing tasks or exhibiting traits that are typically characteristic of human behaviour[1]. Artificial intelligence (AI) technologies have rapidly proliferated across various sectors, significantly impacting decision-making processes in a multitude of domains, including healthcare, finance, law enforcement, and employment practices. These technologies possess the potential to enhance efficiencies, improve accuracy, and streamline operations, promising a future that is both innovative and transformative. However, alongside their numerous benefits,

1. BA., MA., Lecturer, Department of Philosophy, University of Delta, Agbor, Delta State, Nigeria
Email: etaoghenepaulpolo@gmail.com ORCID ID: https://orcid.org/0009-0001-1606-0273
2. BA., Postgraduate Student, Austin Peay State University, U.S.A. Email: ailodiondon@gmail.com

Corresponding Author: Etaoghene Paul Polo, Email: etaoghenepaulpolo@gmail.com

Content of this journal is licensed under a Creative Commons Attribution- CC BY-NC 4.0 International License.

AI technologies have raised profound ethical concerns, particularly regarding racial bias[2]. This bias manifests particularly prominently in facial recognition systems, which have been documented to misidentify individuals from marginalised racial groups at alarming rates. For instance, a landmark study conducted by Joy Buolamwini and Timnit Gebru in 2018 revealed that commercial facial recognition systems misidentified Black women with a failure rate of 34.7%, compared to just 0.8% for white men[3]. Such discrepancies not only reflect technological shortcomings but also exacerbate systemic inequalities, reinforcing existing societal biases and discrimination.

The implications of these findings are both alarming and significant. When AI systems misidentify individuals based on their race, they can lead to dire consequences, including wrongful arrests, disproportionate surveillance, and exclusion from essential services[4]. This creates a cycle of distrust between affected communities and institutions that deploy these technologies. Moreover, the reliance on AI systems that perpetuate bias contributes to a broader societal discourse on fairness, justice, and human rights. The stakes are particularly high in areas such as law enforcement, where biased AI technologies can result in racially targeted policing practices, further entrenching social injustices[5]. As such, the ethical considerations surrounding AI technologies, particularly in relation to racial bias, are increasingly coming under scrutiny from scholars, policymakers, and civil society alike.

This paper aims to explore the issue of racial bias in AI through the lens of the bioethical principle of justice, which underscores the importance of fairness and equitable treatment for all individuals, especially those from vulnerable populations. Central to this exploration is the work of Joy Buolamwini, particularly her seminal project, *Coded Bias*, which exposes the limitations and biases inherent in AI systems. Furthermore, the paper examines the efforts of the *Algorithmic Justice League* (AJL), an advocacy group founded by Buolamwini, dedicated to promoting fairness and accountability in AI. This paper contributes to the ongoing discourse by examining the intersection of racial bias, AI technologies, and bioethics, with a specific focus on the insights gleaned from Joy Buolamwini's *Coded Bias* and the efforts of the *Algorithmic Justice League*.

Methodology: This paper, which was drafted from September 2024 to December 2024, adopts a qualitative research approach. With particular focus on content analysis of Joy Buolamwini's works and case studies/real-world examples of AI bias, the paper explores the ethical implications of racial bias in AI systems. This approach provides a deep understanding of the phenomenon and its impact on marginalised communities through the lens of the bioethical principle of justice. Google Scholar, ResearchGate, Academia, Semantic Scholar, PhilPapers, Web of Science, and other institutional repositories were the search engines used for this research. Articles were searched by following keywords such as AI bias, racial bias, *Algorithmic Justice League*, *Coded Bias*, Joy Buolamwini, facial recognition, equitable algorithms, *et cetera*.

Discussions:

Racial Bias in AI: Racial bias refers to the unfair treatment or prejudiced attitudes towards individuals based on their race or ethnicity[6]. It can manifest in various forms, from social interactions to institutional practices, often resulting in disadvantageous outcomes for marginalised racial groups. This bias is typically rooted in historical and systemic inequalities, reinforcing stereotypes and perpetuating discrimination[7]. In other words, racial bias occurs when people are treated unfairly or judged based on race or ethnicity[8]. Racial bias often leads to negative outcomes for marginalised groups. Rooted in historical inequalities, it perpetuates stereotypes and systemic discrimination[9].

In artificial intelligence (AI), racial bias can manifest in various forms, primarily through data bias and algorithmic bias. Data bias arises when the datasets used to train AI systems are not representative of the diverse population they aim to serve. This lack of representation often leads to inaccurate predictions and decisions, particularly for marginalised racial groups[10]. For example,

facial recognition systems trained predominantly on images of white individuals tend to misidentify people of colour, resulting in disproportionately high error rates for these groups. Buolamwini and Gebru highlight this issue in their research, which shows that commercial facial recognition systems had a misidentification rate of 34.7% for dark-skinned women compared to an error rate of just 0.8% for light-skinned men[11]. Such significant discrepancies underscore the critical need for diverse and representative datasets in AI training.

Algorithmic bias, on the other hand, occurs when algorithms are designed in ways that unintentionally favour certain groups over others[12]. This bias can arise from flawed assumptions made during the development process or from biased historical data on which the algorithms are trained[13]. For example, an algorithm designed to predict recidivism rates (how likely individuals are to reoffend after release) may unintentionally reinforce existing biases within the justice system if trained on historical crime data. As such, data often reflects systemic prejudices, the algorithm could lead to unfair treatment of certain racial groups, perpetuating cycles of discrimination and further marginalising already disadvantaged communities[14].

Algorithmic bias can be found in the hiring algorithms used by companies to screen job applicants[15]. These algorithms are often trained on historical hiring data that may reflect past hiring decisions favouring certain demographic groups over others. For instance, if a company's historical hiring data shows a preference for male candidates in technical roles, an AI hiring system trained on this data might perpetuate that bias by systematically favouring male applicants, while disadvantaging equally or more qualified female candidates[16]. This issue was highlighted in a case where an AI recruitment tool developed by Amazon was found to downgrade resumes that included the word "women" or references to female-related activities, resulting in gender discrimination during the hiring process[17]. The tool's bias stemmed from the historical data used to train the algorithm, which primarily consisted of resumes from male candidates, thus reinforcing gender imbalances in hiring decisions. This example illustrates how AI systems can perpetuate existing biases, even when there is no explicit intention to discriminate.

Impact on Marginalised Groups: The consequences of biased AI systems extend beyond technological failures; they can lead to real-world discrimination, wrongful identification, and violations of civil rights. For instance, facial recognition technology has been used in law enforcement, where misidentifications can result in wrongful arrests or excessive surveillance of specific communities[18]. A particularly alarming case involved the wrongful arrest of Robert Williams, a Black man who was misidentified by facial recognition technology and detained for over 30 hours[19]. Such instances not only undermine trust between marginalised communities and law enforcement agencies but also perpetuate a cycle of systemic injustice.

Furthermore, the ethical implications of AI bias can contribute to a broader societal discourse on race, equity, and justice, prompting calls for urgent reforms in AI development and deployment. The negative effects of racial bias in AI are not limited to law enforcement; they can infiltrate various sectors, including employment, healthcare, and education. In hiring practices, AI systems might favour candidates from historically privileged backgrounds, thereby exacerbating inequalities in the job market. Similarly, biased algorithms in healthcare can lead to disparities in treatment recommendations, affecting the quality of care received by individuals from marginalised groups.

Addressing racial bias in AI is not merely a technical challenge; it is an ethical imperative that demands a commitment to fairness and accountability in all aspects of AI system design and implementation. To mitigate these biases, it is crucial for AI developers and stakeholders to actively engage with diverse communities, ensuring that the perspectives and experiences of marginalised groups inform the development of AI technologies[20]. Additionally, implementing rigorous testing and validation processes can help identify and

correct biases in AI systems before they are deployed[21]. Essentially, these processes should involve: comprehensive audits that assess how algorithms perform across different demographic groups, publishing the data sources used for training AI, and providing clear explanations of how algorithms make decisions. These will help ensure that all individuals receive equitable treatment. As AI continues to play an increasingly significant role in shaping our society, it is crucial that AI systems are developed with a commitment to justice, ensuring that they uplift and empower all individuals, particularly those from historically marginalised communities[22].

The Bioethical Principle of Justice: The bioethical principle of justice focuses on the fair distribution of benefits and burdens, advocating for equal treatment and protection for all individuals, particularly those from vulnerable or marginalised populations[23]. This principle holds that equity is not merely an ideal but a necessary ethical obligation. In healthcare, for instance, justice calls for equitable access to medical resources and interventions, ensuring that no group is disproportionately disadvantaged[24].

In the context of AI, the principle of justice becomes increasingly relevant, as AI significantly impacts numerous facets of everyday life, including access to resources, safety, and civil rights, determining the outcomes of critical decisions, from loan approvals to job hiring and criminal justice sentencing. If the algorithms that power AI systems are biased, they can perpetuate systemic inequalities, with marginalised groups frequently bearing the brunt of negative outcomes. For example, biased AI technologies in facial recognition have been shown to misidentify individuals from specific racial and ethnic backgrounds, leading to harmful consequences such as wrongful arrests or denied services[25].

Ultimately, the principle of justice calls for a critical examination of how AI systems are designed, deployed, and governed, highlighting the need for ethical oversight throughout the entire lifecycle of AI technologies.

Joy Buolamwini's *Coded Bias:* Joy Buolamwini's project, *Coded Bias,* serves as a critical examination of the inherent biases present in AI systems, particularly in facial recognition technologies. Through her research, Buolamwini revealed that many commercial AI systems exhibited significant accuracy disparities based on race and gender, disproportionately misidentifying individuals from marginalised groups[26]. Her groundbreaking work began when she noticed that facial recognition systems struggled to correctly identify her own face, a phenomenon she later realised was indicative of a much larger problem[27].

In her research, Buolamwini conducted a series of experiments using various facial recognition technologies from major tech companies. She discovered that while these technologies are often touted for their objectivity and precision, they are, in reality, reflective of the biases embedded in their training data and algorithms. For instance, she found that facial recognition software demonstrated an error rate of 34.7% for dark-skinned women, compared to just 0.8% for light-skinned men[28]. Such significant discrepancies not only reveal the technological limitations of these systems but also expose the ethical dilemmas involved in their deployment, particularly when these technologies are used in sensitive areas like law enforcement, hiring practices, and public surveillance.

In addition to her research on facial recognition, Buolamwini has explored other areas of AI bias. In one study, she examined the disparities in gender classification algorithms, revealing that AI systems consistently misclassified individuals of darker skin tones and women at higher rates than lighter-skinned men[29]. Another research by Buolamwini focused on the ethical implications of AI bias in healthcare systems, where biased algorithms can lead to unequal treatment, particularly for marginalised communities[30].

Buolamwini's works underscore her broader focus on the socio-technical challenges of AI fairness across various domains. They also

highlight the ethical implications of relying on flawed AI systems, drawing attention to the potential harms inflicted on already vulnerable populations. Her findings indicate that when AI technologies are employed without a critical understanding of their limitations, they can perpetuate existing inequalities and exacerbate societal injustices. The project, *Coded Bias,* not only raises awareness about the limitations of AI technologies but also serves as a call to action for greater scrutiny and accountability in AI development. In essence, *Coded Bias* challenges the prevailing narrative that AI technologies are inherently objective and infallible, urging stakeholders to confront the reality that biases can be encoded in the very fabric of these systems.

Moreover, Buolamwini's works align with the broader discourse on digital ethics, prompting a re-evaluation of how technology interacts with societal norms and values. As AI systems increasingly influence critical areas of life, the implications of bias in these technologies become even more pronounced. Buolamwini emphasises that technology does not exist in a vacuum; rather, it is shaped by the cultural and social contexts in which it is developed and deployed.

The Algorithmic Justice League (AJL):

The *Algorithmic Justice League* (AJL), founded by Joy Buolamwini, is a prominent organisation dedicated to promoting fairness and accountability in AI. It is aimed at raising awareness about the ethical implications of AI technologies, particularly concerning their impact on marginalised communities[31]. The mission of the AJL is multifaceted, encompassing advocacy for equitable algorithmic practices, the development of inclusive datasets, and the implementation of policies that mitigate racial and gender bias in AI systems[32]. It is noteworthy that the AJL's efforts are not only focused on raising awareness but also on influencing the regulatory landscape surrounding AI technologies. Through educational initiatives, research dissemination, and public engagement, the AJL endeavours to hold tech companies accountable for their AI products and to push for systemic reforms that ensure justice in AI development. AJL engages a diverse array of stakeholders, including technologists, policymakers, and civil society, fostering a collaborative approach to addressing the challenges posed by biased algorithms. This collaborative framework is crucial in creating a collective understanding of the implications of AI technologies and in driving meaningful change.

The *Algorithmic Justice League* (AJL) has made significant strides in influencing AI policy and reform since its inception. One of its key achievements is the successful advocacy for greater transparency in AI systems, urging companies and developers to disclose the data sources and methodologies used in their algorithms. This transparency is crucial for identifying and addressing biases within AI systems, enabling independent scrutiny and evaluation[33]. The AJL also played a pivotal role in various forums, including congressional hearings and international conferences, where it influenced public perception of AI technologies by leveraging research findings such as those from Buolamwini's *Coded Bias*, and establishing partnerships with key organisations[34].

The AJL's efforts have not only exposed the risks associated with biased algorithms but have also educated the public on the broader implications of these technologies in relation to civil rights and social justice. This educational approach empowers underrepresented communities to voice their concerns and hold tech companies accountable.

Despite these successes, the challenge of addressing racial bias in AI remains ongoing. As technology rapidly evolves, the AJL's proactive work will be essential in sustaining momentum for reform and advocating for the rights of marginalised communities, ensuring AI development adheres to principles of equity and justice.

Conclusion: This paper has provided an in-depth examination of the pervasive issue of racial bias within artificial intelligence (AI), particularly through the lens of the bioethical principle of justice. A significant focus has been placed on the critical insights derived from Joy Buolamwini's influential work, *Coded Bias*, alongside the advocacy initiatives spearheaded by the *Algorithmic Justice*

League (AJL). Our findings reveal that AI systems, especially those employed for facial recognition, often exacerbate existing societal inequalities by disproportionately misidentifying individuals from marginalised racial groups. For example, Buolamwini's research indicated that commercial facial recognition technologies tend to misidentify darker-skinned individuals at rates far exceeding those of their lighter-skinned counterparts. This alarming trend raises profound ethical concerns surrounding fairness, equity, and justice in the development and deployment of AI technologies.

Moreover, the impact of racial bias in AI extends beyond mere technological inaccuracies; it fundamentally undermines the rights and dignity of affected individuals, thereby entrenching systemic injustice. This paper has highlighted that the ramifications of biased AI are not confined to misidentifications but also encompass wrongful arrests, excessive surveillance, and a general erosion of trust between law enforcement agencies and the communities they serve. The societal implications of these technologies necessitate an urgent and multifaceted response, one that advocates for transparency, accountability, and ethical considerations in AI development.

The AJL's efforts to raise awareness, advocate for accountability, and promote transparency in AI practices represent a vital response to the pressing challenges posed by racial bias in AI. By engaging diverse stakeholders, including technologists, policymakers, and community representatives, the AJL has emphasised the need for inclusive dialogue and systemic reform. The work of the AJL underscores the importance of developing equitable algorithms that consider the unique needs and experiences of all demographic groups, particularly those historically subjected to marginalisation and discrimination.

References

1. Olojede HT, Polo EP. In Praise of Normative Science: Arts and Humanities in the Age of Artificial Intelligence. International Journal of Social Sciences and Humanities. 2025 Jan;11(2):1-9. DOI: https://doi.org/27265774111101121-1
2. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018;81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed October 1, 2024).
3. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018;81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed October 1, 2024).
4. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018;81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed October 1, 2024).
5. Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. ProPublica 2016 May 23. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
6. Dovidio JF, Gaertner SL. Prejudice, Discrimination, and Racism. Orlando: Academic Press; 1986.
7. Bonilla-Silva E. Racism without Racists: Colour-blind Racism and the Persistence of Racial Inequality in America. 5th ed. Lanham: Rowman & Littlefield; 2014. ISBN: 9781442220553.
8. Pager D, Shepherd H. The sociology of discrimination: Racial discrimination in employment, housing, credit, and consumer markets. Annual Review of Sociology. 2008;34:181-209. DOI:https://doi.org/10.1146/annurev.soc.33.040406.131740
9. Bonilla-Silva E. Racism Without Racists: Color-Blind Racism and the Persistence of Racial Inequality in America. 5th ed. Lanham: Rowman & Littlefield; 2017.
10. O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing Group; 2016. ISBN: 9780553418811.
11. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018;81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed October 1, 2024).
12. Barocas S, Hardt M, Narayanan A. Fairness and Machine Learning: Limitations and Opportunities. Cambridge: MIT Press; 2021.
13. Mitchell M, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic fairness: Choices, assumptions, and definitions. Annual Review of Statistics and Its Application. 2021;7;8:141-63. DOI:http://dx.doi.org/10.1146/annurev-statistics-042720-125902
14. O'Neil C. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York: Crown Publishing Group; 2016. ISBN: 9780553418811.
15. Raghavan M, Barocas S, Kleinberg J, Levy K. Mitigating bias in algorithmic hiring: Evaluating claims and practices. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency pp 69-481.

DOI: https://doi.org/10.1145/3351095.3372828 .
16. Lambrecht A, Tucker C. Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. Management Science. 2019Jul;65(7):2966-81. DOI:http://dx.doi.org/10.2139/ssrn.2852260
17. Dastin J. Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women. Reuters [Internet]. 2018. https://www.reuters.com/article/us-amazon-com-jobs-automation-idUSKCN1MK08G(Accessed October 1, 2024).
18. Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. New York: St. Martin's Press; 2018. ISBN: 9781250074317.
19. Hill K. Wrongfully Accused by an Algorithm. The New York Times 2020. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html (Accessed October 1, 2024).
20. West SM, Whittaker M, Crawford K. Discriminating systems: Gender, race, and power in AI. AI Now Institute [Internet]. 2019 Apr [cited 2025 Mar 4]. https://ainowinstitute.org/discriminatingsystems.pdf (Accessed October 1, 2024).
21. Raji ID, Buolamwini J. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society 2019;429-435. DOI: https://doi.org/10.1145/3306618.3314244.
22. Noble SU. Algorithms of Oppression: How Search Engines Reinforce Racism. New York: NYU Press; 2018.
23. Beauchamp TL, Childress JF. Principles of Biomedical Ethics. 7th ed. Oxford: Oxford University Press; 2013. ISBN: 9780199924585.
24. Braveman P, Gruskin S. Defining equity in health. J Epidemiol Community Health. 2003;57(4):254-8.
25. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018;81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed October 1, 2024).
26. Buolamwini J. Coded Bias. [Video]. MIT Media Lab; 2018. https://www.media.mit.edu/videos/coded-bias-2018 (Accessed October 1, 2024).
27. Buolamwini J. Coded Bias. [Video]. MIT Media Lab; 2018. https://www.media.mit.edu/videos/coded-bias-2018 (Accessed October 1, 2024).
28. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018;81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed October 1, 2024).
29. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018;81:77-91. https://proceedings.mlr.press/v81/buolamwini18a.html (Accessed October 1, 2024).
30. Buolamwini J. AI for the People: Combatting Algorithmic Bias in Healthcare. Ethical AI Conference; 2020.https://www.ethicalaiconference2020.org (Accessed October 1, 2024).
31. Algorithmic Justice League. About AJL. https://www.ajl.org/about (Accessed October 1, 2024).
32. Buolamwini J. Algorithmic Justice League: Addressing Bias in AI Systems. Algorithmic Justice League. https://www.ajl.org (Accessed October 1, 2024).
33. Crawford K, Paglen T. Excavating AI: The Politics of Training Sets for Machine Learning. AI Now Institute [Internet]. 2019. https://www.ai-now.org/excavating-ai (Accessed October 1, 2024).
34. Benjamin R. Race After Technology: Abolitionist Tools for the New Jim Code. Cambridge: Polity; 2019. ISBN: 9781509526406.

Authors' Contribution: The first author Etaoghene Paul Polo developed the idea of the research project, did the literature review, and drafted the manuscript. The second author Donatus Osatofoh Ailodion read through and made useful contributions to the manuscript. Both authors meticulously checked the manuscript.

Conflict of Interests: The Authors disclose no conflict of interest.

Funding: None

Data Sharing: We hereby consent that our data be made available in this article for a broader audience by making them publicly accessible as 'open data'.