Zawayatul Fikr
• Journal of Islamic Education •
Zawayatul Fikr: Journal of Islamic Education
Vol 1, No 2, December 2025
E-ISSN: XXXX-XXXX
Page: 181-190

Assessing Algorithmic Fairness: Bias Analysis and Equity Principles in AI-Based Assessment and Personalized Learning Systems

Faricha Nidaul Hanifa

Universitas Darussalam Gontor, Indonesia

Corresponding Author:
Faricha Nidaul Hanifa, Universitas Darussalam Gontor, Indonesia; farichahanifa@gmail.com
DOI: -

Received:	Revised:	Accepted:	Published:
26-12-2025	19-12-2025	30-12-2025	31-12-2025

Abstract

The application of artificial intelligence (AI) in education can improve efficiency and consistency through automated assessment and personalized learning, yet it can also introduce algorithmic bias that amplifies educational inequities. This study examines how algorithmic bias manifests in AI-based assessment systems and personalized learning, and it proposes an algorithmic fairness framework to support equitable implementation of educational technology. This study applies a library-based desk study using a qualitative descriptive-analytical approach. It systematically collects and documents evidence from indexed journal articles (Scopus, Web of Science, IEEE Xplore, ACM Digital Library), textbooks on AI ethics and educational technology, and policy reports from UNESCO and OECD using targeted keywords, then it selects sources based on relevance and credibility. The analysis uses content analysis that (1) identifies core themes on bias sources, injustice manifestations, and fairness frameworks, (2) categorizes findings across technical, social, and ethical dimensions, (3) critically interprets patterns and gaps, and (4) synthesizes a conceptual framework with practical recommendations. The findings show that bias arises in automated essay scoring (AES) and adaptive learning through unrepresentative training data, proxy variables, and feedback loops that reinforce prior disparities. The proposed framework integrates individual, group, and counterfactual fairness, and it clarifies the ethical trade-offs that emerge when these criteria conflict. Bias mitigation strategies include stronger transparency and documentation, participatory design with affected stakeholders, routine algorithmic audits, human-in-the-loop review for high-stakes decisions, and AI literacy training for educators. This study offers practical guidance for developers, educators, and policymakers to build fairer and more transparent educational AI systems.

Keywords: algorithmic bias, algorithmic fairness, AI in education, automated assessment, personalized learning, educational equity, algorithmic fairness

INTRODUCTION

The artificial intelligence (AI) revolution has transformed the global education landscape by delivering automated assessment systems and personalized learning that promise high efficiency, objectivity, and adaptability to individual student needs. AI implementations in education include intelligent tutoring systems, automated essay scoring, adaptive learning platforms, and learning analytics, which have been adopted by thousands

of educational institutions worldwide to improve the quality of learning and reduce the administrative burden on educators (Alsagri & Sohail, 2024). Research shows that the global AI-based EdTech market is projected to reach $20 billion by 2027, with exponential growth particularly in developing countries that see technology as a solution to improve access to and quality of education. However, behind the promise of this positive transformation lies serious concerns about algorithmic bias that could reproduce and even reinforce systemic inequities in education, threatening the equity principles that underpin democratic education systems (Bird et al., 2024). Research reveals that machine learning algorithms trained on historical data tend to inherit and amplify racial, gender, and socioeconomic biases embedded in that data, resulting in discriminatory predictions and recommendations against minority and marginalized groups. This phenomenon demands a critical analysis of algorithm fairness in the educational context to ensure that AI technology does not become an instrument that exacerbates educational disparities but instead promotes inclusivity and equal learning opportunities for all students without discrimination.

The artificial intelligence (AI) revolution has begun to reshape education through automated assessment and personalized learning systems that aim to improve efficiency and adapt instruction to student needs, and market analyses have even projected AI-in-education revenues to surpass USD 20 billion by 2027. Yet the most visible gains and risks now surface in developing and low-connectivity contexts, where infrastructure constraints shape what "AI in education" actually looks like. In 2024, only 27% of people in low-income countries used the internet, and Africa averaged 38%, while globally just 40% of primary schools had internet connectivity, which pushes many programs toward low-bandwidth or offline delivery. Alongside these mixed impacts, AI systems can reproduce inequities through algorithmic bias in automated essay scoring, adaptive pathways, unrepresentative training data, proxy variables, and feedback loops (Taşkın, 2025) which makes fairness governance urgent in contexts where models often import assumptions from high-income settings. This study therefore conducts a library-based desk study using a qualitative descriptive-analytical approach, drawing on indexed literature and policy reports and applying structured content analysis and source triangulation to synthesize how bias manifests in AI-based assessment and personalization and to develop a practical algorithmic fairness framework for educators, developers, and policymakers.

Algorithmic bias in educational AI systems stems from various stages of the algorithm development cycle, from data collection and feature selection to model design and output interpretation, each of which is susceptible to both human and structural biases. The training data used to train machine learning algorithms is often unrepresentative of the diversity of the student population, with overrepresentation of majority groups and underrepresentation of minority groups resulting in models that underperform students from underrepresented backgrounds in the data. Research (Matthews et al., 2022) showed that word embeddings used in natural language processing inherit gender and racial stereotypes from the text corpus used to train them, which then influence automated essay scoring systems in evaluating student writing based on names or styles associated with

certain demographic groups. Bias can also arise from proxy variables that indirectly encode sensitive demographic information; for example, zip code can be a proxy for race and socioeconomic status, allowing algorithms using these variables to make discriminatory decisions even without explicitly using racial or social class categories. Furthermore, feedback loops in adaptive systems can reinforce initial biases, where students initially rated low by the algorithm receive less challenging content, further hindering their progress and confirming the algorithm's initial predictions in a self-fulfilling prophecy. The study confirmed that adaptive learning systems tend to provide different recommendations to students based on the algorithm's perception of their abilities, which is often influenced by demographic bias, resulting in disparities in learning opportunities and educational outcomes.

The principle of algorithmic fairness in the educational context encompasses multiple complexes and often conflicting dimensions, requiring a comprehensive ethical framework to evaluate and mitigate bias in AI systems. Fairness can be defined through various lenses: individual fairness, which requires that similar individuals be treated similarly; group fairness, which requires that different demographic groups have equivalent outcomes; and counterfactual fairness, which requires that algorithmic decisions remain unchanged if sensitive individual attributes are changed. Research (Bogina et al., 2021) studies have shown that these various definitions of fairness often cannot be met simultaneously, creating trade-offs that require consideration of values and ethical priorities in the specific context of educational applications. In AI-based assessment systems, procedural fairness requires transparency in how decisions are made, accountability when errors occur, and the opportunity for students to understand and challenge algorithmic assessments. Distributive justice in personalized learning requires that all students, regardless of their demographic background, have equal access to high-quality learning resources and the opportunity to reach their full potential.

The impact of algorithmic bias on students from marginalized and minority groups can be multidimensional and long-term, affecting not only academic achievement but also their motivation to learn, self-efficacy, and educational aspirations. Biased automated assessment systems can systematically underestimate the abilities of students from certain groups, generating negative feedback that undermines self-confidence and creating stereotype threat, where students feel threatened by negative stereotypes about their group. Research (Cooke et al., 2022) reveals that intelligent tutoring systems that use algorithmic predictions to adapt content difficulty can create "algorithmic tracking" where students from disadvantaged backgrounds are systematically directed down less challenging learning paths, limiting their access to advanced content and reducing their academic mobility. Biased personalized learning can also narrow the curriculum students' access, limiting their exposure to diverse topics and perspectives based on algorithmic assumptions about "relevance" that are often influenced by demographic stereotypes. Furthermore, high-stakes decisions such as dropout predictions, program placement recommendations, or the identification of gifted students made by biased algorithms can have long-term consequences

for students' educational and career trajectories, perpetuating structural inequalities across generations.

This research aims to conduct a comprehensive analysis of bias and equity principles in AI-based assessment and learning personalization systems to develop a framework and practical recommendations that can ensure algorithmic fairness in educational technology implementation. The urgency of this research is based on the increasingly widespread adoption of AI systems in education without an adequate understanding of the risks of bias and mechanisms to mitigate them, which can result in detrimental consequences for millions of students, especially from groups already marginalized in conventional education systems. By integrating perspectives from computer science, education, ethics, and social justice, this research seeks to bridge the gap between technological innovation and inclusive and equitable pedagogical values. Theoretically, this research contributes to the development of an algorithmic justice framework that is contextually relevant for the educational domain, which differs from AI applications in other fields such as criminal justice or hiring due to the developmental and formative nature of the educational process. Practically, the findings of this research are expected to provide guidance for EdTech developers, educators, school administrators, and policymakers in designing, evaluating, and implementing fair, transparent, and accountable educational AI systems. Thus, this research is expected to contribute to the realization of an educational vision that harnesses the power of AI technology to promote equal learning opportunities and social justice, rather than to reinforce existing inequities.

METHODS

This study applies library research with a qualitative descriptive-analytical approach to examine algorithmic bias and fairness principles in AI-based learning assessment and personalization systems. The study draws data from indexed journal articles (Scopus, Web of Science, IEEE Xplore, ACM Digital Library), textbooks on AI ethics and educational technology, and policy reports from UNESCO and OECD. The literature selection follows explicit criteria. First, inclusion criteria cover publications that discuss AI systems used for educational assessment or personalization and provide empirical findings, technical explanations, or normative frameworks related to algorithmic bias, fairness, equity, transparency, or accountability. Second, the study prioritizes peer reviewed articles and official institutional reports, and it excludes non-reviewed opinion pieces, promotional industry content, and sources without a clear methodology. Third, the review focuses on recent literature to capture current model designs and governance debates, while it retains seminal works that define core fairness concepts when later studies repeatedly cite them. Fourth, the study includes cross-disciplinary sources from computer science, education, and ethics to ensure conceptual completeness, and it excludes studies that address fairness in unrelated domains without clear relevance to educational contexts. Data collection uses systematic documentation by searching titles, abstracts, and keywords with the terms "algorithmic bias", "fairness in AI education", "automated assessment", and "personalized

learning equity", followed by screening and full-text reading to confirm eligibility. Data analysis uses content analysis through four steps: identifying core themes on bias sources, injustice mechanisms, and justice frameworks; categorizing findings into technical, social, and ethical dimensions; critically interpreting patterns and gaps; and synthesizing results into a conceptual framework and practical recommendations. The study strengthens validity through source triangulation across disciplines and cross-checking claims across multiple publications to reduce single-source bias and improve analytical comprehensiveness.

FINDINGS AND DISCUSSION

Manifestations of Algorithmic Bias in AI-Based Assessment and Learning Personalization Systems

Algorithmic bias in AI-based automated assessment systems, such as automated essay scoring (AES), is evident in the unfair assessment of certain groups of students. AES rely on features that often reflect non-universal academic norms, such as syntactic complexity and vocabulary, which tend to marginalize students with strong ideas but limited skills in formal academic writing. This exacerbates inequities in access to quality education, as these assessment results influence important decisions like college admissions (Yang et al., 2024). If assessment systems do not reflect diverse ways of thinking and writing, they will exacerbate inequalities in the education system and hinder the academic growth of students from marginalized groups. This unfairness can reinforce existing social inequalities, hindering the academic mobility of students who have potential but are limited by non-inclusive standards. Therefore, it is crucial to continuously evaluate and design assessment systems that are fairer and more representative of the diversity of student thinking.

AI-based personalized learning creates a more subtle but dangerous bias, as it impacts students' learning opportunities through opaque systems. Adaptive learning systems utilize data to predict student abilities, but their results are often biased by assumptions built into historical data. This creates "filter bubbles" that limit topic exploration, and algorithms often assign students deemed "low-performing" less challenging content, ultimately reinforcing the algorithm's initial predictions and creating a self-fulfilling prophecy (Munappy et al., 2022). As a result, personalized learning no longer provides equal opportunities for all students, but instead reinforces their limitations, ultimately creating a wider gap in educational opportunities. This system, which fails to accommodate individual differences, limits students' potential for growth and diminishes the quality of an education that should be inclusive. More flexible and adaptive learning is needed so that technology can support, rather than hinder, student development.

Bias in training data is a significant source of algorithmic inequity. Algorithms trained with unrepresentative data tend to be poor at predicting the learning patterns of minority students. For example, data on student achievement often reflects existing structural inequities, such as school segregation or teacher bias, causing algorithms to reproduce these inequities in their predictions and recommendations (Kizilcec & Lee, 2020). Fairness in Education. *ArXiv*, abs/2007.05443 This suggests that while technology is designed

for efficiency and objectivity, it actually exacerbates existing inequalities, creating an education system that is more exclusive than inclusive. When historical data is used to train models, the biases embedded in this data are passed on, resulting in recurring inequities in the education system. Therefore, it is crucial to ensure that the data used to train algorithms reflects the diversity of society, so that AI systems can function more equitably for all students, especially those who have traditionally been underrepresented.

Feature selection in machine learning algorithms also has the potential to introduce demographic bias, even when sensitive variables like race or gender are not explicitly used. Variables like zip code or time spent on a platform can be proxies for socioeconomic status or race, allowing discriminatory decisions to be made without specifying demographic categories. Indiscriminate feature selection can also reinforce bias, such as using standardized test scores that are biased against certain groups (Nezami et al., 2024). This highlights that even if AI systems appear objective, the use of certain untested variables can magnify inequities, further worsening the opportunities for already marginalized students. The use of seemingly neutral variables that actually reflect social inequalities can exacerbate existing educational disparities. Therefore, a thorough evaluation of the tools used is necessary to ensure they do not become tools for perpetuating structural inequities.

Feedback loops in adaptive AI systems can exacerbate bias by reinforcing the algorithm's initial assessments. When low-scoring students receive unchallenging content, they become trapped in a cycle where their initial assessments are increasingly confirmed, reducing their opportunities to demonstrate higher abilities. This can lead to large differences in learning trajectories that begin with small differences in the algorithm's initial assessments (Bagunaid et al., 2022). These undetected feedback loops can lock students into limited academic pathways, denying them the opportunity to develop to their full potential, severely undermining education, which should be fair and equitable for all. These feedback loops not only limit student learning but also exacerbate stereotypes and stratification within the education system. Therefore, it is crucial to design systems that allow students to develop beyond initial assessments, giving them the opportunity to demonstrate their true potential.

Algorithmic bias in AI-based assessment and personalized learning systems can exacerbate educational inequities. Systems like automated essay scoring (AES) and adaptive learning systems tend to unfairly grade certain students, particularly those from minority or disadvantaged groups. This bias arises because algorithms often rely on unrepresentative data or normative features, such as language complexity or learning speed, that do not reflect the diversity of students' learning styles. Furthermore, bias in training data and feature selection can lead to deeper inequities, reinforcing existing educational disparities. To ensure a more equitable AI-based education system, it is crucial to design and implement technology that accommodates diverse student backgrounds and conducts ongoing evaluations to mitigate bias and ensure equal learning opportunities for all students.

Algorithmic Fairness Framework and Bias Mitigation Strategies in Educational AI Systems

The principle of algorithmic fairness in education reveals significant challenges in achieving equality of outcomes across groups. Different definitions of fairness, such as individual fairness and group fairness, often conflict, requiring trade-offs that must be navigated based on ethical values and pedagogical priorities in specific educational contexts. This means that in educational applications, decisions about which type of error is more acceptable-false positives or false negatives—must be made with careful ethical consideration (Idowu, 2024). Implementing algorithmic fairness in education means making balanced decisions, ensuring that every student is given an equal opportunity to succeed, without sacrificing the quality or integrity of the education. In education, every student should be treated fairly according to their abilities, not based on assumptions or potentially biased profiles. Therefore, a fair assessment system will ensure that educational decisions are more reflective of students' needs and not simply meeting predetermined metrics.

Transparency and explainability of algorithms are prerequisites for ensuring accountability in educational AI systems. Transparently explainable systems allow educators and students to understand why decisions are made, which is crucial for identifying and addressing bias. Simpler models, such as decision trees, should be prioritized for their ease of understanding over complex and difficult-to-explain "black box" models (Khosravi et al., 2022). By ensuring transparency, we can ensure that technology-based education remains accountable and responsible, giving educators and students greater control over the learning process. Transparency also increases trust between developers, educators, and students, and allows for more constructive interactions between technology and the parties involved in education. Therefore, every decision made by the system must be explainable and understandable to all parties involved to avoid distrust in technology.

Participatory and inclusive design strategies can help reduce bias in AI systems by ensuring that perspectives from marginalized groups are integrated early in the development process. This includes involving students, parents, and minority communities in defining solutions and evaluating technology applications. By involving them in the design process, we can identify potential biases that might be invisible to developers from the majority group. It is crucial to create a truly inclusive education system, where all voices, especially those of marginalized groups, are recognized and heard at every stage of technology development. This ensures that every group impacted by technology has the right to contribute to decisions that will affect their education. This way, we can ensure that the resulting solutions are relevant and beneficial to all segments of society.

Algorithmic audits and ongoing monitoring are crucial for detecting and mitigating bias in deployed AI systems. Through ongoing testing and monitoring, we can identify disparities between demographic groups and ensure that the system remains fair over time. This process should be accompanied by a feedback mechanism that allows users to report any bias they encounter and introduce improvements based on those findings (Murikah et al., 2024).Ongoing monitoring is key to ensuring the system continues to evolve toward a

more equitable and accountable system, ultimately supporting diversity and inclusivity in education. This process also ensures that technology is not implemented once, but rather through continuous evaluation and improvement, leading to a better and more inclusive education system. Thus, monitoring is a crucial instrument in maintaining the quality and equity of education in the digital age.

A human-in-the-loop approach and educator empowerment are crucial steps to ensure that AI in education does not replace human decisions, but supports them. These systems enable educators to review and adjust algorithm recommendations, reducing the risk of automation bias and ensuring decisions remain reflective of students' specific contexts. AI education is also crucial to ensure that educators understand the limitations of the technology and can use AI systems critically and informedly (López-Meneses et al., 2025). By empowering educators to play an active role in every algorithmic decision, we can ensure that technology in education will always serve as a supporting tool, not a substitute, in educating future generations. Human-in-the-loop ensures that while technology plays a significant role in increasing efficiency, the final decision remains with educators who better understand the context and needs of their students. Therefore, collaboration between technology and educators must be fundamental to creating equitable and effective educational experiences.

Implementing algorithmic fairness in education is a significant challenge that requires a balanced approach between various fairness principles and ethical priorities in the educational context. Decisions made by AI systems must consider acceptable types of errors and ensure that every student has an equal opportunity to succeed without compromising the quality or integrity of education. By increasing algorithm transparency and developing more explainable systems, we can ensure that AI decisions are understandable not only to developers but also to educators and students. This transparency process will increase trust in the technology and enable educators and students to interact with the system more openly and responsibly.

Furthermore, it is crucial to integrate perspectives from marginalized groups into the educational technology development process through inclusive, participatory design. By involving students, parents, and minority communities in the design process, we can identify and mitigate potential biases in AI systems. Continuous algorithmic audits are also necessary to detect inequities that may emerge over time, ensuring that systems remain fair and functioning for all students. A human-in-the-loop approach that empowers educators to play an active role in every algorithmic decision will ensure that technology does not replace, but supports, human decisions, resulting in more inclusive AI-based education and a more equitable learning experience for all students.

CONCLUSION

The application of artificial intelligence (AI) in education holds substantial potential to improve learning quality and operational efficiency, yet it also introduces serious challenges related to algorithmic bias. Bias can arise across the full system lifecycle, from

data collection and feature design to model training and interpretation, and it can translate into inequitable outcomes, especially for students from marginalized or minority groups. Educational AI therefore requires inclusive and equitable design that accommodates diversity in learners' backgrounds, thinking patterns, and learning styles. Transparent and explainable systems also matter because students, educators, and institutions need to understand how AI reaches conclusions, and they need a clear path to question and correct decisions. Bias mitigation should combine participatory design that involves affected groups, algorithmic audits, and human-in-the-loop practices that keep educators empowered to review, override, and contextualize AI outputs so the technology supports decisions rather than replaces them. In addition, sustained fairness cannot rely on one-time validation because student cohorts, curricula, language use, and platform behaviors change over time. Continuous evaluation and monitoring are essential to detect performance drift, emerging proxy biases, and feedback loops that can gradually reintroduce inequities, and they should operate through scheduled audits, ongoing outcome tracking across demographic groups, and iterative model updates aligned with ethical governance and educational goals. With sustained evaluation, transparency, and educator oversight, AI-based education systems can move closer to providing equal opportunities for all learners to progress and succeed.

REFERENCES

Alsagri, H., & Sohail, S. (2024). Evaluating the role of Artificial Intelligence in sustainable development goals with an emphasis on "quality education." *Discover Sustainability*, *5*(1), 458. https://doi.org/10.1007/s43621-024-00682-9

Bagunaid, W., Chilamkurti, N., & Veeraraghavan, P. (2022). AISAR: Artificial Intelligence-Based Student Assessment and Recommendation System for E-Learning in Big Data. *Sustainability*, *14*(17), 10551. https://doi.org/10.3390/su141710551

Bird, K., Castleman, B., & Song, Y. (2024). Are algorithms biased in education? Exploring racial bias in predicting community college student success. *Journal of Policy Analysis and Management*, *43*(2), 456–485. https://doi.org/10.1002/pam.22569

Bogina, V., Hartman, A., Kuflik, T., & Shulner-Tal, A. (2021). Educating Software and AI Stakeholders About Algorithmic Fairness, Accountability, Transparency and Ethics. *International Journal of Artificial Intelligence in Education*, *32*(3), 808–833. https://doi.org/10.1007/s40593-021-00248-0

Cooke, F., Dickmann, M., & Parry, E. (2022). Building sustainable societies through human-centred human resource management: emerging issues and research opportunities. *The International Journal of Human Resource Management*, *33*(1), 1–15. https://doi.org/10.1080/09585192.2021.2021732

Idowu, J. (2024). Debiasing Education Algorithms. *International Journal of Artificial Intelligence in Education*, *34*(4), 1–31. https://doi.org/10.1007/s40593-023-00389-4

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S., Maldonado, R. M., Sadiq, S., & Gašević, D. (2022). Explainable Artificial Intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kizilcec, R. F., & Lee, H. (2020). Algorithmic Fairness in Education. In *arXiv preprint*. https://doi.org/10.48550/arXiv.2007.05443

López-Meneses, E., López-Catalán, L., Pelícano-Piris, N., & Mellado-Moreno, P. (2025). Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. *Applied Sciences*, *15*(2), 772. https://doi.org/10.3390/app15020772

Matthews, S., Hudzina, J., & Sepehr, D. (2022). Gender and Racial Stereotype Detection in Legal Opinion Word Embeddings. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 12026–12033. https://doi.org/10.48550/arXiv.2203.13369

Munappy, A., Bosch, J., Olsson, H. H., Arpteg, A., & Brinne, B. (2022). Data management for production quality deep learning models: Challenges and solutions. *Journal of Systems and Software*, *191*, 111359. https://doi.org/10.1016/j.jss.2022.111359

Murikah, W., Nthenge, J., & Musyoka, F. (2024). Bias and Ethics of AI Systems Applied in Auditing - A Systematic Review. *Scientific African*, *25*, e02281. https://doi.org/10.1016/j.sciaf.2024.e02281

Nezami, N., Haghighat, P., Gándara, D., & Anahideh, H. (2024). Assessing Disparities in Predictive Modeling Outcomes for College Student Success: The Impact of Imputation Techniques on Model Performance and Fairness. *Education Sciences*, *14*(2), 136. https://doi.org/10.3390/educsci14020136

Taşkın, M. (2025). Artificial Intelligence in Personalized Education: Enhancing Learning Outcomes Through Adaptive Technologies and Data-Driven Insights. *Human Computer Interaction*, *9*(1).

Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the Tapestry of Automated Essay Scoring: A Comprehensive Investigation of Accuracy, Fairness, and Generalizability. In *arXiv preprint*. https://doi.org/10.48550/arXiv.2401.05655