Stakeholder Trust and AI in Education: A Policy Perspective on Data Privacy, Bias, and Decision-Making

Chukwudum Collins Umoke
Department of Science Education,
Alex Ekwueme Federal University Ndufu-Alike, Ebonyi State.
umoke.chukwudum@funai.edu.ng

Sunday Odo Nwangbo
Department of Political Science,
Alex Ekwueme Federal University Ndufu-Alike, Ebonyi State.
snwangbo@gmail.com

Oroke Abel Onwe
Department of Computer Science Education,
Ebonyi State College of Education, Ikwo orokeabel@gmail.com
DOI: 10.56201/rjpst.vol.8.no3.2025.pg54.68

Abstract
Artificial Intelligence (AI) is increasingly integrated into educational systems, shaping student assessments, admissions, and personalized learning. However, ethical concerns related to trust, fairness, transparency, accountability, and data privacy remain significant barriers to its responsible deployment. This study explores these issues through the SAFE-T Framework (Stakeholder-Aligned Fairness, Ethics, Transparency in AI-Education), a model designed to enhance ethical AI governance in education. Employing a qualitative research design based on secondary data analysis, this study examines AI policies, governmental regulations, and scholarly literature to assess AI governance effectiveness. Findings highlight persistent challenges in transparency, particularly in AI-driven decision-making processes, as well as algorithmic biases that reinforce educational inequities. The study underscores the need for fairness-aware AI models, participatory AI policy frameworks, and accountability mechanisms such as fairness audits and regulatory oversight. The implications of this research emphasize the necessity for educational institutions to integrate explainable AI, ethical oversight, and AI literacy programs to build trust among stakeholders. Proposing structured governance mechanisms, this study contributes to the discourse on responsible AI adoption in education and offers recommendations for ensuring equitable and transparent AI-driven learning environments.

Keywords: *AI governance, transparency, fairness, algorithmic bias, ethical AI*

Introduction

Artificial Intelligence (AI) has become an integral component of modern education, influencing student assessments, admissions, and personalized learning experiences (Knowles et al., 2022). However, while AI-driven technologies promise efficiency and innovation,

concerns regarding trust, fairness, transparency, accountability, and data privacy continue to challenge their ethical implementation in education (Patidar et al., 2024; Westover, 2024). The SAFE-T Framework (Stakeholder-Aligned Fairness, Ethics, Transparency in AI-Education) provides a structured approach to addressing these challenges by ensuring that AI governance in education prioritizes fairness-aware algorithms, regulatory compliance, and stakeholder engagement (Chaudhry et al., 2022; Peney et al., 2024). The purpose of this study is to analyze the role of transparency, fairness, and accountability in AI-driven education, emphasizing the need for ethical AI governance frameworks to ensure equitable and responsible AI adoption in educational settings (Akinrinola et al., 2024; Zhang, 2024).

AI decision-making processes in education remain largely opaque, with students and educators struggling to understand how AI-generated recommendations and assessments are derived (Chaudhry et al., 2022). The lack of transparency contributes to distrust, particularly when AI-based grading, admissions, and scholarship allocations exhibit biases that disproportionately disadvantage marginalized groups (Mangal & Pardos, 2024; Bogina et al., 2021). Scholars argue that explainable AI (XAI) plays a crucial role in mitigating these concerns by making AI outputs interpretable and justifiable to stakeholders (Patidar et al., 2024; Peney et al., 2024). The SAFE-T Framework integrates transparency as a foundational principle, advocating for AI systems that are both interpretable and aligned with ethical standards to foster trust among users (Westover, 2024; Manias et al., 2023).

Fairness in AI-powered education is another pressing issue, as biased algorithms can reinforce social and economic disparities, further entrenching existing educational inequalities (Chinta et al., 2024; Nazeer, 2024). Studies highlight that algorithmic bias often stems from training datasets that fail to represent diverse student populations, leading to discriminatory outcomes in student assessments and admissions (Baker & Hawn, 2021; Udoh et al., 2024). Implementing fairness-aware AI models, diverse training datasets, and bias detection frameworks can help mitigate these disparities (Fu et al., 2020; Hacker et al., 2020). The SAFE-T Framework underscores the importance of fairness-aware AI policies that incorporate continuous monitoring and human oversight to prevent bias-related injustices in education (Angerschmid et al., 2022; Wong, 2019).

Accountability in AI-driven education is essential for safeguarding student rights and ensuring that AI decision-making processes remain ethical and legally compliant (Chaudhary, 2024). Many scholars argue that regulatory oversight, fairness audits, and participatory governance are necessary to hold AI developers and educational institutions accountable for the ethical deployment of AI technologies (Chakraborty & Gummadi, 2020; Akinrinola et al., 2024). The SAFE-T Framework advocates for a hybrid approach, combining strict regulatory mandates with stakeholder engagement to create AI governance policies that are both enforceable and adaptable to evolving educational contexts (Hong et al., 2022; Peney et al., 2024). Ensuring accountability also requires the implementation of transparency indices and AI literacy programs to educate students and educators about AI-driven decision-making processes (Bendechache et al., 2021; Mirishli, 2024).

The study is structured into several key sections. The literature review explores existing research on AI trust, fairness, transparency, and accountability, providing insights into the successes and challenges of AI governance in education. The conceptual framework presents the SAFE-T Framework as a model for ethical AI governance, emphasizing transparency, fairness, and stakeholder engagement. The methodology section details the research design, which employs secondary data analysis, document reviews, and thematic analysis to examine AI policy effectiveness and ethical considerations. The findings and discussion section analyzes the implementation of AI governance in education, drawing from case studies and scholarly insights to highlight best practices and challenges. Finally, the conclusion and recommendations provide policy and practice implications for AI adoption in education, advocating for ethical AI deployment strategies that prioritize student rights, equity, and accountability.

Literature Review

Algorithmic Bias in Education: Case Studies on AI-Induced Disparities in Learning Outcomes

AI-driven decision-making in education, while promising, has the potential to reinforce historical biases, leading to disparities in student outcomes. A case study on AI-powered grade prediction revealed that models systematically favored students from specific racial and economic backgrounds, raising concerns about fairness in algorithmic assessments (Mangal & Pardos, 2024). These biases often stem from historical data embedded in AI models, resulting in ethical dilemmas in automated decision-making processes (Nazeer, 2024). Addressing such biases requires targeted bias mitigation techniques, including data curation and algorithmic transparency, to ensure equitable AI-driven decisions (Nazeer, 2024).

Bias in AI-driven educational tools extends beyond race and economic background, affecting gender and socioeconomic status as well. Studies indicate that AI models used in student assessments and admissions may perpetuate existing disparities, necessitating fairness-aware algorithms and diverse training datasets to counteract such effects (Chinta et al., 2024). The disproportionate impact of algorithmic bias on marginalized student groups is particularly concerning, as predictive modeling in admissions and assessments often disfavors these populations. To address this, bias detection frameworks have been suggested as a means of evaluating fairness in AI-powered learning applications (Baker & Hawn, 2021).

Ensuring fairness in AI systems remains complex, requiring an intersectional approach that considers legal, social, and technical aspects. A proposed framework attempts to balance algorithmic fairness with predictive accuracy, aiming to reduce discriminatory outcomes in student evaluations while maintaining the reliability of AI-generated assessments (Udoh et al., 2024).

AI in Decision-Making: Policy Frameworks Ensuring Fairness and Equity in Education Policies

AI's increasing role in educational policy-making necessitates careful scrutiny, as it can inadvertently reinforce pre-existing inequities if fairness is not explicitly programmed into its decision processes. Policymakers must integrate fairness-aware AI models to ensure that all students, regardless of background, receive equitable opportunities in education (Wong, 2019). The role of AI-driven learning analytics further complicates fairness considerations, as biased evaluation metrics can lead to skewed interpretations of student progress. To mitigate these risks, fairness constraints must be embedded within AI-based educational analytics frameworks, ensuring that evaluations remain impartial and promote equity (Fu et al., 2020).

Developing flexible algorithmic fairness frameworks is essential to accommodate the varying legal, technological, and ethical dimensions of AI governance. A proposed adaptive algorithm suggests that fairness constraints should be tailored to different educational contexts, preventing the rigid application of one-size-fits-all policies that may not address specific institutional needs (Hacker et al., 2020). Additionally, regulatory policies should mandate fairness audits and transparency reports for AI-based learning systems, fostering an environment of accountability and trust (Chakraborty & Gummadi, 2020).

AI fairness policies must evolve to address the complexities of real-world educational settings. A participatory approach to AI governance, involving key stakeholders such as educators, students, and policymakers, is necessary to ensure that AI remains a tool for educational equity rather than a mechanism that perpetuates bias. Continuous AI audits and stakeholder engagement in policy implementation are crucial strategies for maintaining fairness in AI-powered education systems (Zhang, 2024).

Algorithmic Bias in Education: Case Studies on AI-Induced Disparities in Learning Outcomes

AI-driven decision-making in education, while promising, has the potential to reinforce historical biases, leading to disparities in student outcomes. A case study on AI-powered grade prediction revealed that models systematically favored students from specific racial and economic backgrounds, raising concerns about fairness in algorithmic assessments (Mangal & Pardos, 2024). These biases often stem from historical data embedded in AI models, resulting in ethical dilemmas in automated decision-making processes (Nazeer, 2024). Addressing such biases requires targeted bias mitigation techniques, including data curation and algorithmic transparency, to ensure equitable AI-driven decisions (Nazeer, 2024).

Bias in AI-driven educational tools extends beyond race and economic background, affecting gender and socioeconomic status as well. Studies indicate that AI models used in student assessments and admissions may perpetuate existing disparities, necessitating fairness-aware algorithms and diverse training datasets to counteract such effects (Chinta et al., 2024). The disproportionate impact of algorithmic bias on marginalized student groups is particularly concerning, as predictive modeling in admissions and assessments often disfavors these

populations. To address this, bias detection frameworks have been suggested as a means of evaluating fairness in AI-powered learning applications (Baker & Hawn, 2021).

Ensuring fairness in AI systems remains complex, requiring an intersectional approach that considers legal, social, and technical aspects. A proposed framework attempts to balance algorithmic fairness with predictive accuracy, aiming to reduce discriminatory outcomes in student evaluations while maintaining the reliability of AI-generated assessments (Udoh et al., 2024).

AI in Decision-Making: Policy Frameworks Ensuring Fairness and Equity in Education Policies

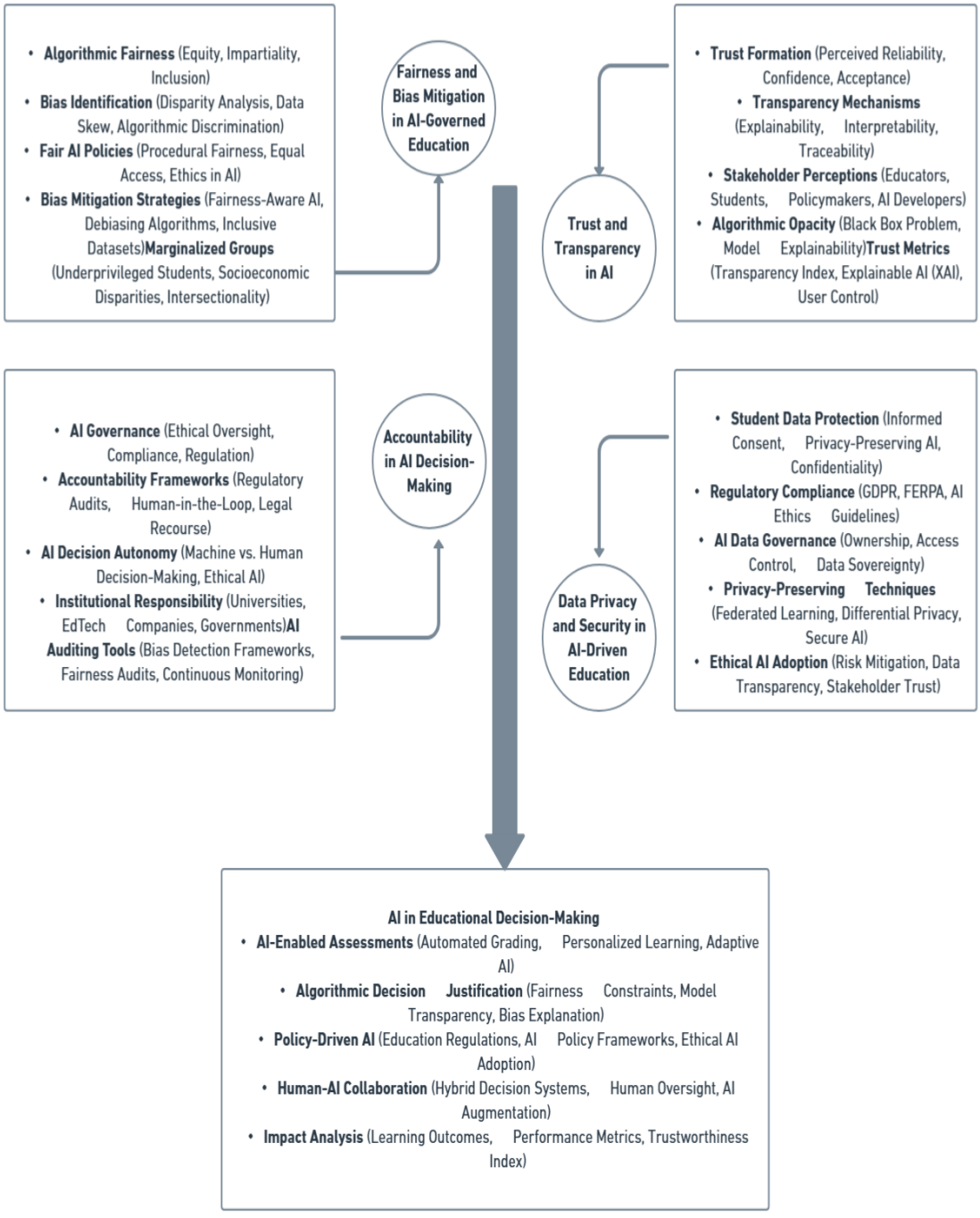
AI's increasing role in educational policy-making necessitates careful scrutiny, as it can inadvertently reinforce pre-existing inequities if fairness is not explicitly programmed into its decision processes. Policymakers must integrate fairness-aware AI models to ensure that all students, regardless of background, receive equitable opportunities in education (Wong, 2019). The role of AI-driven learning analytics further complicates fairness considerations, as biased evaluation metrics can lead to skewed interpretations of student progress. To mitigate these risks, fairness constraints must be embedded within AI-based educational analytics frameworks, ensuring that evaluations remain impartial and promote equity (Fu et al., 2020).

Developing flexible algorithmic fairness frameworks is essential to accommodate the varying legal, technological, and ethical dimensions of AI governance. A proposed adaptive algorithm suggests that fairness constraints should be tailored to different educational contexts, preventing the rigid application of one-size-fits-all policies that may not address specific institutional needs (Hacker et al., 2020). Additionally, regulatory policies should mandate fairness audits and transparency reports for AI-based learning systems, fostering an environment of accountability and trust (Chakraborty & Gummadi, 2020).

AI fairness policies must evolve to address the complexities of real-world educational settings. A participatory approach to AI governance, involving key stakeholders such as educators, students, and policymakers, is necessary to ensure that AI remains a tool for educational equity rather than a mechanism that perpetuates bias. Continuous AI audits and stakeholder engagement in policy implementation are crucial strategies for maintaining fairness in AI-powered education systems (Zhang, 2024).

Conceptual Model Development

The SAFE-T Framework (Stakeholder-Aligned Fairness, Ethics, Transparency in AI-Education) provides a structured approach to addressing trust, fairness, transparency, accountability, and data privacy concerns in AI-driven education. As AI increasingly influences educational decision-making, ensuring that these systems are trustworthy, unbiased, and ethically governed is paramount (Knowles et al., 2022). Trust in AI is deeply intertwined with transparency and accountability, requiring mechanisms such as explainable AI (XAI) and



stakeholder engagement to enhance user confidence in AI-powered education systems (Patidar et al., 2024).

Figure 1: SAFE-T Framework illustration. Source: Authors.

A fundamental component of the SAFE-T Framework is algorithmic fairness, ensuring that AI-driven educational tools do not reinforce existing biases in student evaluations, admissions, or personalized learning systems (Bogina et al., 2021). Studies highlight that AI-based systems can perpetuate disparities if fairness is not explicitly programmed, making fairness-aware algorithms and diverse training datasets crucial for mitigating systemic inequalities (Manias et al., 2023). Additionally, ethical oversight and accountability mechanisms, such as fairness audits and human-in-the-loop approaches, are necessary to prevent discriminatory outcomes in AI decision-making (Angerschmid et al., 2022).

Transparency plays a central role in the SAFE-T Framework, as the black-box nature of AI systems can create uncertainty among educators, students, and policymakers (Chaudhry et al., 2022). The integration of a Transparency Index for AI-powered education tools enables stakeholders to assess the explainability, ethical considerations, and continuous system improvements of AI applications (Chaudhry et al., 2022). Explainable AI methodologies have been shown to increase trust by allowing users to understand and validate AI-generated recommendations, ultimately fostering greater acceptance and adoption of AI-driven education platforms (Peney et al., 2024).

Accountability within the SAFE-T Framework emphasizes the need for regulatory oversight and compliance with international data protection policies, such as the GDPR, to safeguard student data privacy (Akinrinola et al., 2024). Universities and policymakers must implement AI governance frameworks that uphold ethical AI principles, including human oversight, redress mechanisms, and data privacy protections (Westover, 2024). Additionally, AI literacy initiatives that educate students and educators about algorithmic risks and data privacy concerns can further reinforce trust and accountability in AI-powered learning environments (Bendechache et al., 2021).

The SAFE-T Framework ensures that AI technologies serve as tools for equitable and responsible learning by integrating fairness, ethics, and transparency into AI-driven education. Future research should focus on refining governance structures and bias detection frameworks to uphold the ethical and regulatory standards necessary for trustworthy AI adoption in education (Zhang, 2024).

Methodology

This study employs a qualitative research design that focuses on secondary data analysis to explore trust, fairness, transparency, accountability, and privacy concerns in AI-driven education. Instead of conducting interviews or primary data collection, the study relies on analyzing existing literature, policy documents, and government reports to assess AI governance in education. The research draws from multiple sources, including AI education policies, institutional equity reports, and governmental frameworks that guide AI adoption in educational settings. Additionally, peer-reviewed academic literature on AI ethics, fairness in education, and policy frameworks is examined to provide a comprehensive understanding of how AI is shaping decision-making in educational institutions.

The data collection process involved document analysis, which entailed reviewing AI education policies, government regulations, institutional reports, and scholarly publications to extract relevant insights on AI governance and fairness. This method allows for a broad examination of existing regulatory frameworks and AI implementation strategies across different educational contexts. The study also evaluates international policies such as the GDPR and other AI governance models that influence transparency, accountability, and fairness in AI-powered education. Thematic analysis was used to analyze the collected data, allowing for the identification of patterns related to AI policy effectiveness, ethical considerations, and access challenges. This approach facilitated a structured examination of recurring themes across different policy documents and scholarly discussions, enabling a deeper understanding of how AI-driven education can be designed to be fair, transparent, and accountable.

This study highlights both the successes and challenges of AI governance in education, by systematically reviewing secondary data sources. It presents a comparative analysis of policy interventions and best practices, assessing their alignment with the SAFE-T Framework. The study's findings contribute to the ongoing discourse on AI ethics and policy-making, offering insights that can inform future governance strategies to ensure AI-driven education remains equitable, trustworthy, and aligned with ethical principles.

Case Studies on AI and Trust in Education: Applications of the SAFE-T Framework

Algorithmic Bias in AI-Powered Student Assessments

In a study on AI-driven grade prediction, researchers found that machine learning models exhibited biases favoring students from certain racial and socioeconomic backgrounds (Mangal & Pardos, 2024). The biased training data led to disparities in predicted grades, which could impact student confidence, placement, and future opportunities. Applying the SAFE-T Framework, an intervention incorporating fairness-aware AI models and diverse training datasets was proposed to mitigate bias and improve equity in predictive assessments. This case highlights the necessity of algorithmic fairness and transparency to ensure AI-driven assessments do not perpetuate systemic inequalities.

Transparency in AI-Based Admissions Systems

A university implementing AI-driven admissions processes faced backlash after students and policymakers raised concerns over opaque decision-making criteria (Chaudhry et al., 2022). The lack of transparency led to distrust among applicants, particularly those from underrepresented groups. Adopting the SAFE-T Framework, the institution introduced an explainability feature, allowing applicants to understand how their applications were evaluated. Additionally, a Transparency Index was developed to provide insights into AI-driven selection criteria. As a result, trust in the admissions process increased, demonstrating the importance of explainability in AI-driven education policies.

Data Privacy Challenges in AI-Driven Learning Analytics

A secondary school piloted AI-powered learning analytics to track student progress, but concerns emerged regarding data privacy and informed consent (Huang, 2023). Parents and educators feared unauthorized access to sensitive student data and potential misuse of personal information. Implementing the SAFE-T Framework, the school introduced GDPR-aligned data protection policies, ensuring secure data handling and informed consent mechanisms. AI literacy programs were also introduced to educate stakeholders on data security best practices. This case study underscores the significance of accountability and regulatory compliance in AI-driven educational tools.

Fairness in AI-Powered Scholarship Allocation

An AI system used to allocate merit-based scholarships was found to disproportionately disadvantage students from rural and lower-income backgrounds due to biased training data (Bogina et al., 2021). The institution responded by incorporating fairness audits and bias detection tools, following the SAFE-T Framework's emphasis on ethical oversight and human-in-the-loop decision-making. As a result, the AI model was restructured to prioritize fairness without compromising merit-based evaluation. This case demonstrates the impact of fairness-aware AI policies in ensuring equitable access to educational opportunities.

Ethical AI Implementation in Personalized Learning Systems

An adaptive learning platform designed to personalize coursework for students faced ethical scrutiny when it was revealed that the AI recommended different curricula based on students' demographic backgrounds (Westover, 2024). The lack of transparency in the recommendation algorithms raised concerns about reinforcing educational inequalities. Applying the SAFE-T Framework, the platform's developers introduced explainable AI features, fairness constraints, and real-time monitoring to ensure equitable content distribution. This intervention enhanced student trust and improved the ethical integrity of the AI-powered platform. These case studies illustrate the practical applications of the SAFE-T Framework in addressing challenges related to fairness, transparency, accountability, and data privacy in AI-driven education.

Results and Discussion

The SAFE-T Framework highlights key barriers and solutions in AI-driven education, particularly concerning trust, transparency, fairness, and accountability. Scholars largely agree that trust in AI-driven education is compromised by privacy risks, lack of transparency, and ethical concerns (Knowles et al., 2022; Huang, 2023). The opacity of AI systems, often referred to as the "black-box problem," undermines confidence in algorithmic decision-making, particularly in student assessments and admissions (Chaudhry et al., 2022). Some researchers emphasize that transparency mechanisms, such as explainable AI (XAI), are necessary to foster trust, while others argue that transparency alone is insufficient if fairness and accountability measures are not simultaneously implemented (Patidar et al., 2024; Peney et al., 2024).

Privacy concerns are central to trust barriers, particularly regarding student data protection. While GDPR-style regulations offer a foundation for safeguarding student data (Akinrinola et al., 2024), studies indicate that many educational institutions lack comprehensive AI governance frameworks to ensure compliance (Westover, 2024). Scholars such as Bendechache et al. (2021) advocate for AI literacy programs to mitigate privacy risks by empowering students and educators to understand how their data is used. However, others, including Mutuku (2024), argue that legal protections must precede AI literacy efforts, as awareness alone does not prevent data misuse. The SAFE-T Framework integrates both perspectives, emphasizing regulatory compliance alongside education initiatives as complementary strategies for fostering trust.

Policy interventions addressing AI-driven education governance have shown varying degrees of success. Case studies indicate that AI-based admissions processes can be enhanced by integrating fairness metrics and explainability features, as demonstrated in models that prioritize diverse training datasets to reduce biases (Bogina et al., 2021; Manias et al., 2023). However, AI-driven grading and scholarship allocation systems continue to face scrutiny due to algorithmic bias, as seen in Mangal and Pardos' (2024) study, where AI grade prediction models disproportionately disadvantaged marginalized groups. Zhang (2024) suggests that fairness-aware algorithms must be dynamically updated to reflect evolving educational contexts, a position that aligns with the SAFE-T Framework's emphasis on continuous AI audits and monitoring.

Accountability remains a contentious issue in AI-powered education. While some researchers advocate for strict regulatory oversight and transparency reports to ensure ethical AI deployment (Chakraborty & Gummadi, 2020; Wong, 2019), others contend that participatory policy-making, involving key stakeholders such as educators, students, and policymakers, is more effective (Fu et al., 2020; Hacker et al., 2020). The SAFE-T Framework reconciles these viewpoints by advocating a hybrid model that integrates regulatory mandates with participatory governance, ensuring that AI policies are both enforceable and context-sensitive.

Best practices for enhancing trust in AI-powered education focus on transparency, fairness, and user control. Explainable AI is widely regarded as a crucial mechanism for increasing trust, as demonstrated in studies highlighting the positive impact of transparency indices on stakeholder confidence (Chaudhry et al., 2022; Patidar et al., 2024). However, transparency alone does not eliminate bias, as highlighted by Chinta et al. (2024), who stress the need for fairness-aware AI models that incorporate diverse and representative datasets. This is particularly relevant in scholarship allocation, where algorithmic biases have been shown to disadvantage rural and lower-income students (Bogina et al., 2021). The SAFE-T Framework aligns with these findings by promoting a holistic approach that integrates transparency with fairness-aware AI and human oversight to ensure ethical decision-making.

The role of stakeholder-centered AI policies is crucial for ethical AI implementation in education. While some scholars emphasize regulatory compliance as the primary safeguard against AI-related risks (Akinrinola et al., 2024; Westover, 2024), others argue that AI policies must be adaptive and co-created with input from students, educators, and technology providers

(Hong et al., 2022; Peney et al., 2024). The SAFE-T Framework supports this multi-stakeholder approach, advocating for ethical AI policies that incorporate diverse perspectives to address concerns about fairness, transparency, and privacy. This perspective is reinforced by studies showing that participatory AI policy design leads to higher adoption rates and greater trust in AI-driven education systems (Bendechache et al., 2021; Elantheraiyan et al., 2024).

The policy and practice implications of this study are significant. The SAFE-T Framework provides a structured approach to addressing AI-related challenges in education by integrating fairness, transparency, and accountability mechanisms. From a policy standpoint, educational institutions should implement AI governance frameworks that align with international data protection standards while incorporating fairness audits and explainability mandates. This would help mitigate algorithmic bias and reinforce trust in AI-powered learning environments (Chakraborty & Gummadi, 2020; Zhang, 2024). Additionally, policymakers must recognize the importance of AI literacy initiatives, ensuring that students and educators understand AI decision-making processes and data privacy risks (Bendechache et al., 2021; Mirishli, 2024).

In practice, AI developers must prioritize fairness-aware AI systems that incorporate diverse training datasets and bias detection mechanisms to prevent discriminatory outcomes (Manias et al., 2023; Chinta et al., 2024). AI decision-making in education should not be fully autonomous but should instead follow a human-in-the-loop approach to ensure accountability and fairness (Angerschmid et al., 2022). Institutions must also establish transparency indices that allow stakeholders to assess AI-driven recommendations and interventions (Chaudhry et al., 2022). By integrating these strategies, AI-driven education can transition from a model of algorithmic opacity to one of ethical and equitable decision-making.

Ultimately, the SAFE-T Framework serves as a foundational model for guiding the ethical and effective deployment of AI in education. Future research should explore the long-term impact of AI governance models and develop adaptive fairness constraints that evolve with changing educational landscapes. Through addressing the limitations of existing AI policies and advocating for a balanced approach between regulation and participatory governance, this study contributes to the broader discourse on AI trust, fairness, and accountability in education.

Conclusion

The SAFE-T Framework provides a comprehensive model for addressing trust, fairness, transparency, accountability, and privacy concerns in AI-driven education. The increasing reliance on AI in education necessitates a structured approach to ensure that algorithmic decision-making remains ethical, explainable, and fair. Scholars widely agree that trust in AI is contingent upon transparency and fairness, but the literature also reveals tensions regarding the best approaches to mitigating bias and ensuring accountability. While some emphasize the need for regulatory oversight, others advocate participatory policy-making to ensure that AI policies reflect the diverse needs of stakeholders. The SAFE-T Framework reconciles these perspectives by integrating fairness-aware AI, transparency mechanisms, regulatory compliance, and stakeholder involvement. Case studies highlight the risks of biased AI systems, opaque decision-making processes, and weak data privacy protections, reinforcing the

necessity of adopting a structured framework that balances technological innovation with ethical safeguards. The implications of this study suggest that AI adoption in education must be guided by principles of equity, accountability, and human-centered governance to ensure that students, educators, and policymakers can trust AI-driven systems to enhance learning outcomes rather than perpetuate existing disparities.

Recommendations

Policymakers should implement AI governance frameworks that emphasize transparency, fairness, and accountability, ensuring compliance with data protection regulations such as the GDPR. Regulatory bodies must establish explainability mandates and fairness audits to assess the impact of AI-driven educational tools on diverse student populations. Institutions should integrate human oversight into AI decision-making, ensuring that automated systems support rather than replace educators in critical decision-making processes. AI developers must prioritize the design of fairness-aware algorithms by utilizing diverse datasets and bias detection mechanisms to prevent discriminatory outcomes. Universities and schools should introduce AI literacy programs to educate students and educators about algorithmic risks, data privacy, and ethical AI use. Continuous AI audits and monitoring should be incorporated into AI deployment strategies to assess long-term impacts on educational equity. Future research should focus on refining adaptive fairness constraints that evolve with changing educational environments and developing participatory governance models that include input from students, educators, and policymakers. AI in education should be seen as a tool that enhances learning and decision-making while adhering to ethical principles that promote trust and equity across all levels of education.

Conflict of Interest: the authors declare no conflict of interest

References

Abimbola, C., Eden, C. A., Chisom, O. N., & Adeniyi, I. S. (2024). Integrating AI in education: Opportunities, challenges, and ethical considerations. *Magna Scientia Advanced Research and Reviews*. 10(02), 006–013. https://doi.org/10.30574/msarr.2024.10.2.0039

Akinrinola, O., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024). Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews*. 18(03), 050–058. https://doi.org/10.30574/gscarr.2024.18.3.0088

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and explanation in AI-informed decision-making. *Machine Learning and Knowledge Extraction, 4*(2), 556-579. https://doi.org/10.3390/make4020026

Baker, R., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education, 32,* 1052-1092. https://doi.org/10.1007/s40593-021-00285-9

Bendechache, M., Tal, I., Wall, P., Grehan, L., Clarke, E., O'Driscoll, A., Van Der Haegen, L., Leong, B., Kearns, A. M., & Brennan, R. (2021). AI in My Life: AI, ethics & privacy workshops for 15-16-year-olds. *13th ACM Web Science Conference*. https://doi.org/10.1145/3462741.3466664

Bogina, V., Hartman, A., Kuflik, T., & Tal, A. S. (2021). Educating software and AI stakeholders about algorithmic fairness, accountability, transparency, and ethics. *International Journal of Artificial Intelligence in Education, 32*, 808-833. https://doi.org/10.1007/s40593-021-00248-0

Chakraborty, A., & Gummadi, K. (2020). Fairness in algorithmic decision making. *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. https://doi.org/10.1145/3371158.3371234

Chaudhary, G. (2024). Unveiling the black box: Bringing algorithmic transparency to AI. *Masaryk University Journal of Law and Technology*. 93–122 https://doi.org/10.5817/mujlt2024-1-4

Chaudhry, M. A., Cukurova, M., & Luckin, R. (2022). A transparency index framework for AI in education. *ArXiv*. https://doi.org/10.48550/arXiv.2206.03220

Chinta, S. V., Wang, Z., Yin, Z., Hoang, N., Gonzalez, M., Quy, T. L., & Zhang, W. (2024). FairAIED: Navigating fairness, bias, and ethics in educational AI applications. *ArXiv*. https://doi.org/10.48550/arXiv.2407.18745

Elantheraiyan, P., Priya, K. M., Gamadia, R., Abdulhasan, M. M., Abood, B. S. Z., & Al-khalidi, A. (2024). Ethical design and implementation of AI in the field of learning and education: Symmetry learning technique. *2024 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 1144-1148. https://doi.org/10.1109/ICACITE60783.2024.10616584

Farooqi, M. T. K., Amanat, I., & Awan, S. M. (2024). Ethical considerations and challenges in the integration of artificial intelligence in education: A systematic review. *Journal of Excellence in Management Sciences*. https://doi.org/10.69565/jems.v3i4.314

Fu, R., Huang, Y., & Singh, P. V. (2020). AI and algorithmic bias: Source, detection, mitigation, and implications. *InfoSciRN: Machine Learning (Sub-Topic)*. https://doi.org/10.2139/ssrn.3681517

Hacker, P., Wiedemann, E., & Zehlike, M. (2021). Towards a flexible framework for algorithmic fairness. *Proceedings of the 16th International Conference on Artificial Intelligence and Law*. 99-108. https://doi.org/10.18420/inf2020_09

Hajoary, D., Narzary, R., & Basumatary, R. (2023). Exploring the evolving dynamics of data privacy, ethical considerations, and data protection in the digital era. *International Journal on Recent and Innovation Trends in Computing and Communication, 11*(9), 2760–2771. https://doi.org/10.17762/ijritcc.v11i9.9363

Hong, Y., Nguyen, A., Dang, B., & Nguyen, B. T. (2022). Data ethics framework for artificial intelligence in education (AIED). *2022 International Conference on Advanced Learning Technologies (ICALT)*, 297-301. https://doi.org/10.1109/ICALT55010.2022.00095

Huang, L. (2023). Ethics of artificial intelligence in education: Student privacy and data protection. *Science Insights Education Frontiers*. 16 (2), 2577-2587 https://doi.org/10.15354/sief.23.re202

Knowles, B., Richards, J. T., & Kroeger, F. (2022). The many facets of trust in AI: Formalizing the relation between trust and fairness, accountability, and transparency. *ArXiv*. https://doi.org/10.48550/arXiv.2208.00681

Mangal, M., & Pardos, Z. (2024). Implementing equitable and intersectionality-aware ML in education: A practical guide. *British Journal of Educational Technology, 55*(5), 2003-2038. https://doi.org/10.1111/bjet.13484

Manias, G., Apostolopoulos, D., Athanassopoulos, S., Borotis, S. A., Chatzimallis, C., & Draksler, T. Z. (2023). AI4Gov: Trusted AI for transparent public governance fostering democratic values. *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, 548-555. https://doi.org/10.1109/DCOSS-IoT58021.2023.00090

Mirishli, S. (2024). Ethical implications of AI in data collection: Balancing innovation with privacy. *ANCIENT LAND*. 6(8), 40-55. https://doi.org/10.36719/2706-6185/38/40-55

Mutuku, M. (2024). Legal and ethical implications of data privacy in artificial intelligence: A review of data privacy among learners in Kenyan secondary schools. *International Journal of Innovative Science and Research Technology (IJISRT)*. 9(9), https://doi.org/10.38124/ijisrt/ijisrt24sep208

Nazeer, M. Y. (2024). Algorithmic conscience: An in-depth inquiry into ethical dilemmas in artificial intelligence. *International Journal of Research and Innovation in Social Science*. 725-732. https://doi.org/10.47772/ijriss.2024.805052

Patidar, N., Mishra, S., Jain, R., Prajapati, D., Solanki, A., Suthar, R., Patel, K., & Patel, H. (2024). Transparency in AI decision-making: A survey of explainable AI methods and applications. *Artificial Intelligence Research and Technology, 2*(1). https://doi.org/10.23880/art-16000110

Peney, L., Dernee, R., & Alers, H. (2024). Encouraging trust in AI-powered teaching tools: Ranking design principles. *2024 IEEE Conference on Artificial Intelligence (CAI)*, 476-479. https://doi.org/10.1109/CAI59869.2024.00096

Udoh, E., Yuan, X., & Rorissa, A. (2024). A framework for defining algorithmic fairness in the context of information access. *Proceedings of the Association for Information Science and Technology*. 61(1), 667-672. https://doi.org/10.1002/pra2.1077

Westover, J. (2024). AI and trust in organizations. *Human Capital Leadership Review*. ***13(3)***. https://doi.org/10.70175/hclreview.2020.13.3.7

Wong, P.-H. (2019). Democratizing algorithmic fairness. *Philosophy & Technology, 33*, 225-244. https://doi.org/10.1007/S13347-019-00355-W

Yang, E., & Beil, C. (2024). Ensuring data privacy in AI/ML implementation. *New Directions for Higher Education*. 207, 63-78. https://doi.org/10.1002/he.20509

Zhang, W. (2024). AI fairness in practice: Paradigm, challenges, and prospects. *AI Magazine*. 45(3), 386-395. https://doi.org/10.1002/aaai.12189