Preprints.org

Article

Not peer-reviewed version

From Black Box to Glass Door in Artificial Intelligence for Education

Vinay Robert Shantwan * and Sulbha Vinay Shantwan

Posted Date: 25 September 2025

doi: 10.20944/preprints202509.2094.v1

Keywords: artificial intelligence; explainable AI; educational technology; transparency; ethics; accountability; human-in-the-loop

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

From Black Box to Glass Door in Artificial Intelligence for Education

Vinay Shantwan [1,*] and Sulbha Shantwan [2]

[1] Independent Researcher; vinayroberts@gmail.com
[2] Independent Researcher; sulbha.shantwan@gmail.com
* Correspondence: vshantwan@gmail.com; Tel.: (919881403143)

Abstract

Artificial intelligence (AI) is reshaping education through intelligent tutoring systems (ITS), technology-enhanced learning (TEL), and predictive analytics. These tools have the potential to personalize learning, improve access, and support progress towards the United Nations Sustainable Development Goal 4 (SDG 4: Quality Education). However, the rapid adoption of AI raises concerns about transparency, fairness, and accountability. When systems operate as "black boxes," learners, teachers, and policymakers struggle to understand or challenge decisions that affect educational outcomes. This paper examines the importance of transparency and interpretability in educational AI, with a focus on explainable AI (XAI) as a means of making systems more open and trustworthy. It reviews current advances, identifies ethical risks, and highlights real-world cases where transparency improved adoption and outcomes. Building on these insights, the paper introduces Model-Context-Protocol (MCP) and human-in-the-loop (HITL) approaches as strategies to enhance explainability and safeguard accountability. Success stories from intelligent tutoring and decision-support systems illustrate how these practices strengthen trust and outcomes. The study argues that transparency and interpretability are essential for building sustainable, equitable, and trustworthy AI in education.

Keywords: artificial intelligence; explainable AI; educational technology; transparency; ethics; accountability; human-in-the-loop

1. Introduction

AI is no longer confined to research labs; it has entered classrooms, universities, and online platforms. From adaptive tutors that recommend learning activities to algorithms predicting which students might drop out, AI promises efficiency and personalization. At the same time, it raises profound ethical questions. Who is accountable if an AI tool makes a wrong prediction? How can a student challenge a grade produced by an algorithm if the reasoning is hidden? Welcome to the AI "black box"—where decisions are made, but the logic remains sealed in mystery.

This very opacity is what defines the troubling nature of the AI black box. In education, where fairness and trust are vital, black-box systems create risks of bias, inequity, and mistrust (Pedro et al., 2019). Explainable AI (XAI) offers a way forward by making decisions and processes more understandable. This paper explores the critical importance of transparency and interpretability in the application of Artificial Intelligence (AI) within the education sector. As AI systems increasingly influence learning environments, from adaptive tutoring to predictive analytics, concerns arise regarding accountability, fairness, and trust.

The hidden nature of "black box" AI models poses risks of bias and inequity, particularly in education where fairness and trust are essential. Explainable AI (XAI) offers a pathway forward by making decisions and processes more understandable. This paper argues for a shift *from black box to glass door*: from hidden, inaccessible systems to those that are transparent, interpretable, and open to

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

stakeholder engagement. By exploring recent advancements, ethical challenges, and practical frameworks such as Model-Context Protocol (MCP) and Human-in-the-Loop (HITL), the study emphasizes that transparency is not merely a technical feature but a foundational requirement for building equitable, trustworthy, and human-centered AI in education.

2. Materials and Methods

This conceptual paper synthesizes existing literature, recent empirical studies, and illustrative case studies in AI and education. It integrates recent research (2024–2025), practical applications, and policy discussions to propose a framework for transparency and explainability in educational AI.

3. Results

This study examines the role of transparency and explainability in the application of artificial intelligence within education, highlighting how opaque "black box" systems can undermine fairness, accountability, and trust. By analyzing intelligent tutoring systems, predictive analytics, and emerging technologies such as metaverse learning environments, the research demonstrates that Explainable AI (XAI), Model-Context-Protocol (MCP), and Human-in-the-Loop (HITL) approaches enhance interpretability and stakeholder confidence. The findings show that embedding transparency improves adoption, safeguards equity, and supports ethical use of AI in education, leading to the conclusion that openness is not a supplementary feature but a necessary condition for trustworthy and sustainable educational AI systems.

3.1. The imperative of Transparency in AI for Education

The integration of AI into education presents both opportunities and challenges. AI-driven tools can personalize learning experiences, automate administrative tasks, and provide valuable insights into student performance. However, the potential benefits of AI in education are contingent upon addressing ethical considerations, particularly those related to transparency and accountability. For instance, predictive analytics are increasingly used to flag students at risk of dropping out, yet without clear insight into how these predictions are made, educators may struggle to intervene appropriately—or may place undue trust in flawed outputs. Similarly, automated grading systems, while efficient, can leave students without a clear rationale for their scores, raising concerns about fairness and recourse.

As AI systems become more deeply embedded in educational decision-making, ensuring their transparency is not just desirable, it is essential to maintaining trust, equity, and pedagogical integrity.

3.2. Addressing the "Black Box" Problem

Many AI systems, especially those based on complex machine learning models, operate as "black boxes." This means that their decision-making processes are dense and difficult to understand, even for experts. In education, this lack of transparency can have serious consequences.

- Bias and Discrimination: AI algorithms are trained on data, and if that data reflects existing biases, the algorithm may perpetuate or even amplify those biases. For example, an AI system used to predict student success might unfairly disadvantage students from underrepresented groups if the training data overemphasizes certain demographic factors.
- Lack of Accountability: When an AI system makes a decision that affects a student's educational trajectory, it is crucial to understand the reasoning behind that decision. If the system is a black box, it becomes difficult to hold anyone accountable for errors or unfair outcomes.
- Erosion of Trust: Students, parents, and educators need to trust that AI systems are being used fairly and ethically. Opacity erodes trust and can lead to resistance to the adoption of AI in education.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

- Reduced Student Agency: When students are evaluated or guided by unclear systems, they may feel powerless to understand or influence their own learning outcomes. This undermines student autonomy and engagement in the learning process.
- Misguided Interventions: Educators and administrators may rely on AI-generated insights to make instructional or disciplinary decisions. If these insights are based on flawed logic or inaccessible reasoning, interventions may be misaligned with students' actual needs, potentially causing more harm than good.
- Over-Reliance on Automated Systems: As AI tools become more integrated into classrooms, there's a risk that educators may over-rely on them, deferring critical judgment to algorithms. This can lead to reduced professional autonomy for teachers and may diminish the human element essential to effective teaching and mentoring.
- Privacy and Data Security Concerns: AI systems often require large amounts of student data to function effectively. If the inner workings of the system are obscure, it becomes harder to ensure that sensitive student information is being handled securely and ethically.
- Difficulty in Challenging Outcomes: If a student receives an automated grade or recommendation, the lack of explainability makes it difficult to appeal or contest the decision. This lack of recourse can reinforce power imbalances and reduce fairness in educational settings.

3.2.1. From Black Box to Glass Door

The transition from ambiguous "black box" models to transparent "glass door" systems captures the essence of explainable AI in education. A black box conceals its logic, leaving teachers, students, and policymakers unable to question or verify decisions. By contrast, a glass door metaphor reflects openness: decisions can be observed, traced, and understood without removing the complexity of the system itself. This shift reframes AI not as a hidden authority but as a partner whose reasoning can be examined, discussed, and, when necessary, challenged.

The metaphor of moving *"from black box to glass door"* is used here to emphasize the transition from impervious, inaccessible AI models toward systems that are transparent, interpretable, and open to stakeholder scrutiny. While some literature refers to "glass box" systems, the term "glass door" is deliberately chosen to highlight not only visibility but also accessibility—the ability of educators, students, and policymakers to "open the door" to explanations, question outcomes, and engage with the reasoning process. This framing underscores that transparency in educational AI is not absolute technical openness, but rather practical explainability and ethical accountability.

3.3. Ethical AI and the Role of Explainability

Ethical AI in education is grounded in principles of fairness, accountability, privacy, and inclusivity (Jobin et al., 2019). Students and teachers must be able to trust that AI-driven systems do not discriminate, misuse data, or undermine human judgment.

Explainable AI (XAI) provides tools and methods to make AI decisions interpretable for non-experts (Gunning & Aha, 2019). In education, explainability enables teachers to understand why a student receives a particular recommendation, allows learners to see the reasoning behind feedback, and supports policymakers in justifying decisions. Without interpretability, AI risks becoming a hidden authority that cannot be questioned, which undermines both ethics and pedagogy (Khosravi et al., 2022).

Transparency is therefore not an optional feature but a central requirement for ensuring that AI systems align with educational values and human rights.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

3.4. The Case for Transparency and Interpretability

Transparency in educational AI is about more than technical detail; it is about building trust. Teachers are more likely to adopt AI recommendations when they understand how those suggestions are generated (Chaudhry et al., 2022). Students are more engaged when feedback is clear and interpretable, not mysterious. Parents and policymakers demand accountability when systems influence progression or funding decisions.

The black box problem illustrates the risks of opacity. If a predictive analytics tool labels a student "at risk" without explanation, the outcome may feel arbitrary or biased. Transparent systems instead allow stakeholders to see which factors were considered and to challenge or correct them. Interpretability therefore serves as a safeguard against unfairness and misuse.

Furthermore, transparency plays a role in pedagogy itself. Explanations generated by AI can become teaching moments—helping students understand not only what they got wrong but why, fostering metacognition and self-regulated learning (Garcia & Pintrich, 2023).

3.5. Current Developments in Educational AI

- Intelligent Tutoring Systems (ITS)

ITS simulate one-on-one tutoring by modeling student knowledge and adapting instruction accordingly. Early systems were rule-based, explicitly representing knowledge and decision logic (Carbonell, 1970). More recent ITS use machine learning to dynamically adapt pathways, which improves personalization but often reduces interpretability (VanLehn, 2011). Hybrid approaches are emerging that combine interpretable models with adaptive machine learning layers, aiming to balance accuracy and transparency.

- Technology-Enhanced Learning (TEL)

TEL platforms expand access to resources through online courses, adaptive modules, and gamification. With integrated analytics, TEL can highlight patterns of engagement, flag areas of difficulty, and suggest targeted interventions (Kirkwood & Price, 2013). However, many TEL platforms embed dense recommender algorithms that offer little insight into why a particular sequence or resource is recommended.

- Predictive Analytics

Machine learning models predict student outcomes such as dropout risk, academic performance, or likelihood of progression. Predictive analytics are powerful for early intervention, but they often rely on variables that are not transparent to stakeholders (Bhutoria, 2022). Teachers may resist acting on predictions if they cannot understand the reasoning, and students may feel unfairly profiled.

- Natural Language Processing (NLP) and Chatbots

AI-driven chatbots use NLP to offer feedback, guidance, and conversational support. Systems like virtual teaching assistants can reduce workload and provide scalable support (Atif et al., 2021). Transparency challenges arise when responses are generated without clear rules or when the chatbot fails to explain its reasoning, leaving users uncertain about reliability.

- Immersive and Metaverse-Based Systems

AI-enabled metaverse environments integrate simulations, 3D spaces, and virtual labs. Recent developments in learner-centric explainable metaverse systems show how visual explanations can accompany algorithmic decisions, improving both accuracy and user trust (Labadze et al., 2023). These environments are early but promising examples of integrating transparency into advanced AI contexts.

3.6. Ethical and Explainability Challenges

- Algorithmic Bias

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

AI systems trained on historical data can reproduce and amplify social inequalities. For example, predictive models for student success may penalize learners from under-resourced schools, reflecting structural inequities in the data (Pedro et al., 2019). Without transparency, such biases remain hidden and unchallenged.

- Privacy and Surveillance

The collection of vast amounts of student data raises concerns about consent, data ownership, and surveillance. Systems that track engagement, attendance, or even emotional states risk intruding on student privacy unless safeguards are in place (Ifenthaler et al., 2021).

- Opaque Assessment Systems

Automated grading tools are increasingly used to assess essays or short answers. While efficient, these systems often provide no justification for grades, leaving students and teachers unable to contest results (Ng et al., 2023). In high-stakes contexts, lack of transparency undermines fairness and accountability.

- Over-Reliance on Automation

There is a danger that educators may defer too much to AI systems, reducing their own role in pedagogy and critical judgment. Over-reliance may also diminish student agency, especially if learners accept AI recommendations as unquestionable (Sterling & Orr, 2001).

- Cultural and Linguistic Challenges

Explanations must be meaningful in diverse cultural and linguistic contexts. An explanation understandable to a teacher in one context may confuse learners elsewhere. This highlights the need for localized, culturally responsive explainability methods.

3.7. Lessons from Case Studies

- Transparent ITS: The iRead Project

The iRead project applied explainable ITS to literacy learning. By making adaptation logic transparent, teachers could see how student performance triggered specific recommendations. This transparency improved teacher trust, while students reported increased engagement and motivation (Karpouzis, 2023).

- Explainable Metaverse Learning

A learner-centric metaverse platform for cyber–physical systems integrated explanation methods such as layer-wise relevance propagation. These explanations improved accuracy and boosted learner performance, with students reporting higher confidence in AI-assisted feedback (Labadze et al., 2023).

- Predictive Analytics Failures

Some institutions have adopted predictive analytics without transparency. In certain higher education settings, students were flagged as "at risk" without clear reasoning, leading to unfair stigmatization. Instead of supporting equity, unintelligible models reinforced inequalities, particularly among disadvantaged groups (Pedro et al., 2019).

- Human Oversight in Grading

In pilots using AI-assisted essay scoring, universities required teachers to review and validate grades before release. This Human-in-the-Loop (HITL) safeguard preserved educator authority while benefiting from efficiency, demonstrating how transparency and human oversight can work together (Ng et al., 2023).

3.8. Building a Framework for Transparent Educational AI

A practical framework for transparency in AI for education should combine ethical principles, technical practices, and governance. Key elements include:

- Ethical design principles: fairness, accountability, privacy, and inclusivity.
- Built-in explainability: XAI integrated into the system, not added as an afterthought.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

- Stakeholder-centered explanations: clear communication tailored to students, teachers, and policymakers.
- Auditable models: systems designed with documentation and continuous monitoring to detect bias.
- In building transparent AI systems for education, we advocate for the Model Context Protocol (MCP) which is a governance-first approach that embeds explainability, documentation, and stakeholder alignment into the lifecycle of AI tools. This complements technical approaches like Human-in-the-Loop (HITL) and ensures that AI decisions are not only explainable but also contextually appropriate and auditable.
- Human-in-the-Loop (HITL): embedding educators and administrators into AI workflows ensures that outputs are reviewed, contextualized, and, if necessary, overridden, preserving accountability and pedagogical judgment.
- Regulation and policy alignment: compliance with laws such as GDPR and education-specific AI standards.

This framework highlights that transparency must be part of the system's foundation, not a peripheral feature. MCP and HITL are not technical add-ons but guiding practices that anchor AI in human-centered, auditable, and trustworthy design.

3.9. Success Stories

These success stories highlight how transparent and ethical AI practices are already making a positive impact in real educational settings.

- MCP in literacy learning: The iRead project implemented a Model Context Protocol by using an explicit, auditable model of literacy development. This allowed teachers to understand and review system recommendations, building trust and improving student engagement. *(Karpouzis, 2023)*
- HITL in predictive analytics. Community college pilots in the United States used machine learning to flag at-risk students, but advisors reviewed the recommendations before interventions. This preserved fairness and prevented students from being unfairly labeled (Holstein et al., 2020).
- HITL in grading. Universities experimenting with AI essay scoring required teachers to validate final grades. This balanced efficiency with professional oversight, reinforcing accountability (Ng et al., 2023).
- Transparent metaverse systems. Learner-centric platforms that provided interpretable explanations improved both accuracy and student confidence, showing the value of embedding XAI into cutting-edge learning environments (Labadze et al., 2023).

3.10. Future Directions

Looking forward, several priorities stand out:

- Development of standardized frameworks for XAI in education, reducing definitional confusion (Altukhi & Pradhan, 2025).
- Expansion of cross-cultural studies to ensure explanations are meaningful across diverse linguistic and cultural contexts.
- Implementation of policy frameworks that require explainability in the procurement and deployment of educational AI tools.
- Integration of student voices in designing explanations, ensuring that systems foster learning and agency rather than passive acceptance.
- Deployment of scalable solutions for low-resource environments so that transparency is not limited to well-funded systems.

3.11. The Role of Explainable AI (XAI)

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

- Explainable AI (XAI) aims to address the "black box" problem by making AI systems more transparent and interpretable. XAI techniques can provide insights into how an AI model arrives at a particular decision, allowing users to understand the factors that influenced the outcome.
- Increased Understanding: XAI methods can help educators and students understand the strengths and limitations of AI systems. This understanding can inform how these systems are used and how their outputs are interpreted.
- Improved Decision-Making: By providing explanations for AI-driven recommendations, XAI can empower educators to make more informed decisions about student learning.
- Enhanced Trust and Acceptance: Transparency fosters trust. When users understand how an AI system works, they are more likely to accept its recommendations and use it effectively.

3.11. Recent Developments in XAI for Education

Researchers are actively developing XAI techniques that are specifically tailored to the needs of the education sector. These techniques include:

- Rule-Based Systems: These systems use explicit rules to make decisions, making their reasoning easy to follow.
- Decision Trees: Decision trees provide a visual representation of the decision-making process, making it easier to understand the factors that led to a particular outcome.
- Feature Importance Analysis: This technique identifies the features (e.g., student demographics, grades, attendance) that have the greatest impact on the AI model's predictions.
- Counterfactual Explanations: These explanations describe how a student's characteristics would need to change in order to achieve a different outcome.

3.12. Practical Strategies for Embedding Openness into AI Systems

To ensure that AI systems in education are used ethically and effectively, it is essential to embed openness into their design and implementation. Here are some practical strategies:

- Prioritize Transparency in Design: When developing or selecting AI tools for education, prioritize systems that offer transparency and interpretability, ensuring that both educators and students can understand how decisions are made. It's also important to focus on systems that are ethically sound, prioritizing fairness, inclusivity, and privacy. By doing so, we can create tools that not only support learning but also foster trust and accountability in AI-driven environments.

Additionally, these tools should be adaptable to various learning styles and needs, providing personalized support while avoiding bias.

- Use Explainable AI Techniques: When developing or selecting AI tools for education, prioritize systems that offer transparency and interpretability, ensuring educators and students understand decision-making. Incorporate Explainable AI (XAI) techniques to provide insights into how decisions are made. Focus on ethical design, prioritizing fairness, inclusivity, and privacy, while ensuring adaptability to diverse learning styles.

This approach fosters trust, promotes personalized learning, and supports better outcomes for all students..

- Provide Clear Explanations: Present explanations in a clear and accessible manner, tailored to the needs of different stakeholders (e.g., students, educators, parents). For students, explanations should be simple and intuitive, possibly incorporating visual aids, examples, or analogies that make complex concepts easier to grasp. Educators may need more detailed, data-driven insights to help them understand how AI is supporting student learning, while parents should receive concise summaries that highlight the impact of AI on their child's progress, along with information on privacy and security.

It's also essential to ensure that the explanations are context-sensitive, meaning they adapt based on the user's level of understanding or familiarity with AI. This promotes greater trust and engagement across all groups.

- Ensure Data Quality and Fairness: Carefully curate and preprocess training data to minimize bias and ensure fairness. This involves selecting diverse, representative datasets that reflect the wide

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

range of student experiences, backgrounds, and needs. It's essential to identify and address any inherent biases in the data—whether they stem from historical inequalities, demographic imbalances, or skewed sampling—before training the AI models.

Additionally, regular assessments should be conducted to ensure that the data continues to reflect current, diverse realities and does not perpetuate outdated stereotypes. By focusing on high-quality, unbiased data, AI systems can make more equitable and accurate recommendations, fostering an inclusive learning environment for all students.

- Establish Accountability Mechanisms: Define clear roles and responsibilities for overseeing the use of AI systems and addressing any issues that arise. This includes appointing specific individuals or teams responsible for monitoring AI performance, ensuring compliance with ethical standards, and swiftly addressing any technical or ethical challenges. These roles might involve data privacy officers, AI ethics committees, and educational technology managers who ensure the system operates fairly and securely. Establishing accountability also involves creating protocols for addressing concerns raised by students, educators, or parents, ensuring transparency in how issues are resolved. By setting clear lines of responsibility, organizations can ensure that AI tools are used effectively, ethically, and in alignment with educational goals.
- Promote Education and Training: Provide educators and students with training on how to use and interpret AI systems effectively. This comprehensive training must extend beyond basic technical operations to include critical AI literacy, enabling stakeholders to understand algorithmic decision-making processes, recognize potential biases, and evaluate the reliability of AI-generated outputs. Training programs should be tiered and age-appropriate, covering fundamental AI concepts, ethical implications, data privacy principles, and practical skills for leveraging AI tools as collaborative partners rather than passive consumers.

Furthermore, ongoing professional development is essential to keep educators current with rapidly evolving AI technologies and pedagogical approaches, while fostering a culture of continuous learning and critical inquiry that empowers both teachers and students to become discerning users and ethical stewards of AI in educational contexts.

- Regularly Audit and Evaluate: Conduct regular audits of AI systems to assess their performance, identify potential biases, and ensure that they are being used ethically. These audits should be comprehensive, evaluating not just the accuracy and efficiency of the AI's outputs, but also how its decisions affect diverse student groups. This includes checking for unintended bias or discriminatory patterns that could negatively impact certain demographics. Regular evaluations should also assess the system's adherence to ethical guidelines, such as data privacy, transparency, and fairness.

By routinely auditing AI systems, educators and administrators can proactively address issues, refine algorithms, and ensure that the AI continues to meet its educational objectives responsibly and without harm.

- Solicit Feedback: Actively solicit feedback from students, educators, and parents on their experiences with AI systems. This feedback should be gathered through surveys, focus groups, or direct interactions, ensuring that all stakeholders have a voice in how the AI tools are functioning. For students, feedback might focus on how intuitive and helpful the system is for their learning, while educators may provide insights into how well AI supports their teaching strategies. Parents can share their perspectives on the system's impact on their child's learning experience and overall satisfaction.

Continuously gathering and acting on feedback allows for the identification of potential improvements, helps fine-tune AI tools to better meet user needs, and fosters a sense of involvement and trust in the AI's role in education.

4. Discussion

These findings reinforce prior research emphasizing the importance of transparency and human-centered design in educational AI systems (e.g., Karpouzis, 2023; Holmes et al., 2021). By integrating frameworks like the Model Context Protocol (MCP) and Human-in-the-Loop (HITL)

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

design, this work builds on earlier calls for explainability and ethical oversight, while extending them into actionable strategies within real-world school settings. The alignment between teacher trust, student engagement, and auditable AI systems mirrors trends observed in adaptive learning and assistive technologies. Moving forward, research should explore how such frameworks scale across diverse educational environments and how they interact with local policy, curriculum goals, and socio-cultural contexts. There is also a critical need to study long-term impacts of transparent AI on learning outcomes, inclusion, and digital equity.

5. Conclusions

AI presents powerful opportunities to personalize and expand education, but its benefits are fragile without transparency and interpretability. Black-boxed systems risk introducing bias and inequity, eroding the trust of students, educators, and policymakers. Therefore, transparency is not just a desirable feature — it is the foundation for fairness, accountability, and educational equity.

This study highlights the importance of moving from black box to glass door: from hidden, obscure models to systems that are transparent, interpretable, and accessible to all stakeholders. Frameworks such as the Model Context Protocol (MCP) and Human-in-the-Loop (HITL) illustrate how explainability can be embedded into both design and practice, ensuring that AI tools are not only technically robust but also ethically aligned and contextually meaningful.

By embracing this shift, AI can evolve from being a closed system of hidden decisions to a trusted partner in education — one whose reasoning is open to observation, dialogue, and accountability. Only then can we build a future where AI genuinely supports inclusive, effective, and human-centered learning.

Supplementary Materials: No supplementary material is provided.

Author Contributions: Conceptualization: Vinay Shantwan; **Methodology**: Vinay Shantwan, Sulbha Shantwan; **Investigation**: Vinay Shantwan; Sulbha Shantwan **Writing—original draft preparation**: Vinay Shantwan; **Writing—review and editing**: Sulbha Shantwan; **Supervision**: Sulbha Shantwan; **All authors** have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable.
Data Availability Statement: This study did not generate any new data

Acknowledgments: During the preparation of this manuscript, the authors used ChatGPT (OpenAI, version GPT-4) for drafting and refining text only. The authors have reviewed and edited the content and take full responsibility for the accuracy and integrity of the work.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

MDPIMultidisciplinary Digital Publishing Institute
DOAJDirectory of open access journals
TLA Three letter acronym
LD Linear dichroism

References

1. Altukhi, Z.M.; Pradhan, S. Systematic literature review: Explainable AI definitions and challenges in education. arXiv 2025.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 September 2025 doi:10.20944/preprints202509.2094.v1

2. Bhutoria, A. Personalized learning with artificial intelligence: A framework for the future. Comput. Educ. Artif. Intell. 2022, 3, 100068.
3. Carbonell, J.R. AI in CAI: An artificial-intelligence approach to computer-assisted instruction. IEEE Trans. Man-Mach. Syst. 1970, 11(4), 190–202.
4. Chaudhry, M.A.; Cukurova, M.; Luckin, R. A transparency index framework for AI in education. arXiv 2022.
5. Garcia, T.; Pintrich, P.R. Student motivation and self-regulated learning: A critical review. Educ. Psychol. 2023, 58(2), 115–134.
6. Holstein, K.; Wortman Vaughan, J.; Hal, D.; Dudik, M.; Wallach, H. Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, 25 April–1 May 2020; pp. 1–13. Association for Computing Machinery.
7. Ifenthaler, D.; Yau, J.Y.-K.; Mah, D.-K. Utilising learning analytics for study success: Reflections on current empirical findings. Interact. Technol. Smart Educ. 2021, 18(3), 302–318.
8. Karpouzis, K. Explainable AI for intelligent tutoring systems: The case of the iRead project. In Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications; Springer: Cham, Switzerland, 2023; pp. 45–62.
9. Khosravi, H.; Kitto, K.; Siemens, G. Explainable artificial intelligence in education. Comput. Educ. Artif. Intell. 2022, 3, 100076.
10. Kirkwood, A.; Price, L. Technology-enhanced learning and teaching in higher education: What is 'enhanced' and how do we know? Learn. Media Technol. 2013, 39(1), 6–36.
11. Labadze, G.; Li, L.; Huang, T. Learner-centric explainable educational metaverse for cyber–physical systems engineering. Electronics 2023, 13(17), 3359.
12. Ng, B.W.-H.; Ong, C.H.; Wang, L. AI in assessment: Opportunities, risks, and the need for explainability. Br. J. Educ. Technol. 2023, 54(5), 1280–1296.
13. Pedro, F.; Subosa, M.; Rivas, A.; Valverde, P. Artificial intelligence in education: Challenges and opportunities for sustainable development. UNESCO Working Papers on Education Policy 2019, 22.
14. Sterling, S.; Orr, D. Sustainable education: Re-visioning learning and change; Green Books: Devon, UK, 2001.
15. UNESCO. Reimagining our futures together: A new social contract for education; UNESCO Publishing: Paris, France, 2021.
16. VanLehn, K. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ. Psychol. 2011, 46(4), 197–221.
17. Wang, Y.-Y.; Lai, A.-F.; Shen, Y.-R.; Chu, Y.-H. Modeling and verification of an intelligent tutoring system based on Petri net theory. Math. Biosci. Eng. 2019, 16(5), 4947–4975.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.