ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor = 8.35, January-June 2026, Published on 28th February 2026

Artificial Intelligence in Education: Ethical Frameworks for Fairness, Safety, and Accountability

Dr. Monika, Assistant Professor of English, Shah Satnam Ji Girls' College, Sirsa Email: Monasarora1@gmail.com

Abstract

The accelerated integration of Artificial Intelligence (AI) into contemporary educational systems has reshaped pedagogical methodologies, institutional governance, and assessment cultures worldwide. While AI-powered adaptive platforms, predictive analytics, generative systems, and automated evaluation tools promise efficiency and personalized learning, their implementation raises urgent ethical concerns related to algorithmic bias, surveillance, accountability, transparency, and psychological well-being. Drawing upon critical theory, the capabilities approach, surveillance studies, democratic theory, and recent global AI regulatory frameworks including the UNESCO 2021 Recommendation on the Ethics of Artificial Intelligence and the 2024 European Union AI Act, this study proposes a comprehensive human-centred governance model for educational AI. It argues that without ethical safeguards, AI risks institutionalizing structural inequalities and weakening human creativity and agency. Responsible AI integration must prioritize fairness auditing, explainability, regulatory oversight, inclusive data practices, and AI literacy to ensure educational justice in digitally mediated learning environment.

Introduction: The world of education is changing fast, driven by the rapid growth of Artificial Intelligence (AI). AI is no longer a distant idea from the future; it is already shaping how teaching happens, how schools are managed, and how students are assessed. Adaptive platforms, prediction tools, generative AI, and automated grading are now part of everyday classroom life. Supporters argue that AI can make learning more efficient and highly personalized, tailoring lessons to each student's needs. Yet this optimistic, technology-first view hides serious ethical problems. The use of AI raises urgent questions about bias in algorithms, the normalizing of surveillance, the weakening of accountability, the hidden "black-box" nature of decisions, and the impact on students' mental health.

This chapter argues that unless strong ethical safeguards are put in place and human values are kept at the centre, the uncritical use of AI in education will deepen existing inequalities and weaken the very creativity and agency it claims to support. Drawing on critical theory, the capabilities approach, surveillance studies, and democratic theory—framed within global policy tools such as the UNESCO 2021 Recommendation on the Ethics of Artificial Intelligence and the 2024 European Union AI Act—this study proposes a comprehensive human-centred governance model. It holds that responsible AI integration must be an ongoing, deliberate process built on fairness auditing, explainability, strong regulation, inclusive data practices, and a deep commitment to AI literacy, so that education can serve justice in our increasingly digital classrooms.

The Double-Edged Sword: AI's Transformative Promise and Peril in Education

The main attraction of AI in education is its promise to solve long-standing problems. Adaptive learning platforms can change the difficulty and type of content based on a student's performance, creating a customized learning path that aims to improve understanding and retention (Kaplan and Haenlein 33). Predictive analytics can examine large amounts of student data to identify those likely to fall behind, allowing schools to intervene early before a student drops out or disengages (Baker and Inventado 1). Generative AI tools, such as large language models, can act as creative partners, helping students brainstorm ideas, draft essays, or practice conversation in a virtual tutoring style. For school leaders, automated grading systems can save many hours of manual marking, freeing teachers to focus on richer, more interactive teaching. This picture of an efficient, responsive, and personalized education system is powerful and, in many ways, genuinely beneficial.

But this promise comes with serious risks. The most worrying is **algorithmic bias**. AI learns from data, and if that data contains patterns of racial, gender, class, or disability-based bias, the AI will copy and often worsen those patterns at scale (O'Neil 53). Imagine an algorithm trained on past student records: it may learn that students from low-income backgrounds, certain language groups, or particular communities have historically done worse and therefore label new students with similar traits as "low potential." This can create a digital tracking system that pushes such students into easier or less ambitious courses, turning social inequality into a technical decision.

Linked to this is the problem of **surveillance**. Educational AI platforms constantly collect fine-grained data—every click, every login, every second spent on a page, every keystroke. As scholars of surveillance note, this turns the classroom into a kind of "panopticon," where students know they are being watched and respond by self-censoring, not by exploring freely (Lyon 85). In this setting, anxiety and performance replace curiosity and experimentation. The companies that sell these tools often treat student data as a commercial product, which raises questions about **accountability and transparency**. When an AI decision harms a student—such as a wrong placement or a misleading warning—who is responsible? The software developer, the school, or the tech company? Because many AI systems are "black boxes," it is hard to know *why* a decision was made, let alone who should be held accountable. This climate of constant monitoring and algorithmic judgment can also hurt students' mental health, creating a culture where learners focus more on pleasing the system than on exploring ideas in a genuine, even messy, way.

Theoretical Lenses for Critique and Construction

To understand these challenges and build better solutions, we need more than just technical knowledge. We must use strong theoretical frameworks that help us see power, justice, and human development in new ways.

Critical theory, especially from thinkers like Jürgen Habermas and the Frankfurt School, reminds us that technology is never neutral (Habermas 36). From this view, AI is not simply a tool; it is a form of power that can support existing social hierarchies and technocratic control. An education system that relies heavily on AI for speed and efficiency may favor instrumental rationality—doing things the fastest and cheapest—over communicative rationality, which focuses on dialogue, participation, and shared understanding. For education to be truly democratic, it must encourage students to think critically, to question authority, and to imagine different futures. Purely algorithmic teaching cannot do this; it can only optimize what is already given.

The **capabilities approach**, developed by Amartya Sen and Martha Nussbaum, offers a moral standard for judging whether AI helps or harms students (Nussbaum 14). Instead of asking only whether students get higher scores, we must ask whether they gain real freedoms or "capabilities" to live the kind of life they value. Does an adaptive platform support a student's ability to think critically and reason for themselves, or does it lock them into a narrow, pre-set path? Does automated grading strengthen a sense of respect and belonging, or does it turn human relationships into cold numbers? An ethical AI system in education must clearly expand core human capabilities such as agency, critical thinking, and emotional well-being, not shrink them in the name of efficiency.

Surveillance studies, drawing on Michel Foucault's work on disciplinary power, help us see how education is becoming "datafied" (Foucault 200). Foucault's panopticon—a prison design where inmates feel they are always watched, even when they are not—maps eerily well onto today's AI-driven classrooms. Continuous data collection creates a system of invisible control, where students adjust their behaviour to avoid negative labels or flags. This is made worse by the logic of **surveillance capitalism**, where companies profit from student data, turning learners into data sources (Zuboff 8). When the commercial goals of EdTech firms clash with

the educational goals of schools, students' interests can easily be pushed aside.

Finally, **democratic theory**, inspired by thinkers like John Dewey and John Rawls, reminds us that schools are not just places for individual learning but the foundation of a fair society (Dewey 87; Rawls 72). Dewey saw schools as small democratic communities, where students learn through collaboration and shared problem-solving. Over-reliance on isolated, algorithm-driven learning can break this sense of community, turning education into a private, data-tracked experience. Rawls's "difference principle" says that social inequalities are only acceptable if they help the least advantaged first. AI systems that widen the digital divide or give extra advantages to already privileged students clearly fail this test and are therefore unjust in a democratic context.

Towards a Human-Centred Governance Model

Given these challenges, a patchy or reactive approach will not be enough. We need a comprehensive, human-centred governance model that is proactive, multi-layered, and clearly oriented toward justice and human flourishing. This model rests on several core pillars.

First, **Fairness Auditing and Algorithmic Impact Assessments (AIAs)** must be mandatory, not optional. Before any AI system is used in education, it must be carefully checked for bias in its data and algorithms. These checks must happen again and again as the system learns and changes over time. The teams doing these audits should be diverse: data scientists, educators, sociologists, ethicists, and community members who represent the groups most affected by the technology.

Second, **Explainability and Transparency** are non-negotiable. Hiding behind a "black box" is an ethical failure. When an AI decision has a big impact on a student—such as placing them in a particular learning track or flagging them as "at risk"—students, parents, and teachers must be able to understand the basic reasons for that decision in language they can follow. This "meaningful transparency" must come with clear ways to appeal or correct errors. The EU AI Act's treatment of certain educational AI systems as "high-risk" and its requirement for high transparency is a key step in this direction (European Parliament).

Third, we need strong **Regulatory Oversight and Multi-Stakeholder Governance**. The UNESCO 2021 Recommendation on the Ethics of Artificial Intelligence offers a global framework that ties AI to human rights, fairness, and sustainability (UNESCO 8). National laws, such as the EU AI Act, must turn these ideals into binding rules. But top-down rules alone are not enough. Schools and districts should create AI ethics committees with teachers, students, parents, and community representatives to oversee how AI tools are chosen, used, and monitored. This kind of participatory governance keeps technology aligned with local educational values, not just corporate or technical interests.

Fourth, we must build **Inclusive Data Practices and Data Dignity**. Bias starts with data. We must make sure that data reflects the full diversity of students, not only a narrow, privileged group. More radically, we should move toward "data dignity," where students and families are treated as co-owners of their own learning data, not just raw material for profit. This means giving people real control over what is collected, how it is used, and who can access it.

Finally, and most importantly, **AI Literacy must become a core part of education**. The best protection against harmful AI use is a population that understands it. AI literacy is not only about knowing how to operate AI tools; it is a form of critical digital literacy that helps students see how AI systems are trained, what biases they may carry, and what limits they have. When students learn to question AI's role in society, they become active, ethical shapers of technology, not passive users.

The Indian Context: A Case for Contextualized Ethics

The need for a human-centred governance model is especially urgent in a country as diverse and unequal as India. The National Education Policy (NEP) 2020 highlights the use of technology to improve learning and equity. But on the ground, huge gaps remain in access to

devices, stable internet, and basic digital skills between urban and rural areas and across social classes. In this context, using one-size-fits-all AI platforms can easily deepen existing inequalities.

An algorithm trained mostly on data from rich urban private schools may misunderstand or misjudge students in rural government schools whose learning environments, languages, and cultural backgrounds are very different. Scholars Nair and Krishnan argue that importing Western EdTech models without careful adaptation can create "a new form of colonialism, one that is algorithmic and data-driven," pushing the already disadvantaged even further to the margins (Nair and Krishnan 51). Therefore, any governance model for AI in Indian education must be deeply local. It requires investment in diverse, multilingual datasets, AI tools that fit local cultures, and strong support for teachers and communities to shape how these tools are used. In India, fairness, transparency, and inclusion are not abstract ideas; they are essential conditions for preventing AI from becoming another engine of social exclusion.

Conclusion

Putting AI into education is not just a technical upgrade; it is a deep ethical and pedagogical change. The "algorithmic pedagogue" offers real benefits in personalization and speed, but it also carries the risk of turning classrooms into more watched, more unequal, and more dehumanized spaces. To handle this duality, we must reject the idea that technology will solve everything on its own and instead choose a deliberate, human-centred path. By grounding our policies in critical theory, the capabilities approach, and democratic ideals—and by building governance inspired by global instruments like the UNESCO Recommendation and the EU AI Act—we can guide AI in education in a positive direction. The pillars of fairness auditing, explainability, multi-stakeholder oversight, data dignity, and widespread AI literacy are more than technical boxes to tick; they are the foundations for ensuring that AI expands human potential and supports true educational justice. Our goal is not to make machines smarter, but to help students become wiser, more creative, and more compassionate human beings. The choices we make today about governing AI in our classrooms will decide whether we succeed at this most important task.

Works Cited

Baker, Ryan Sdz, and Patrick Sebastian Inventado. "The State of Educational Data Mining in 2012: A Review and Future Visions." *Journal of Educational Data Mining*, vol. 4, no. 1, 2012, pp. 1–50.

Dewey, John. *Democracy and Education: An Introduction to the Philosophy of Education*. Macmillan, 1916.

European Parliament. "Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts." 2024.

Foucault, Michel. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan, Vintage Books, 1995.

Habermas, Jürgen. *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society*. Translated by Thomas McCarthy, Beacon Press, 1984.

Kaplan, Andreas, and Michael Haenlein. "Siri, Siri, in My Hand: Who's the Fairest in the Land? On the Interpretations, Illustrations, and Implications of Artificial Intelligence." *Business Horizons*, vol. 62, no. 1, 2019, pp. 15–25.

Lyon, David. *The Culture of Surveillance: Watching as a Way of Life*. Polity Press, 2018.

Nair, R., and S. Krishnan. "Navigating the Digital Divide: Ethical Implications of EdTech in Post-Pandemic India." *Journal of Educational Technology & Society*, vol. 25, no. 3, 2023, pp. 45–58.

Nussbaum, Martha C. *Women and Human Development: The Capabilities Approach*. Cambridge University Press, 2000.

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown, 2016.

Rawls, John. *A Theory of Justice*. Harvard University Press, 1971.

UNESCO. "Recommendation on the Ethics of Artificial Intelligence." 2021, https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, 2019.