Explainable and Ethical AI in Education and Healthcare:
Bridging Trust, Transparency, and Innovation

Hewa Majeed Zangana
http://orcid.org/0000-0001-7909-254X
Duhok Polytechnic University, Iraq

Hakem Beitollahi
Soran University, Iraq

Marwan Omar
http://orcid.org/0009-0000-7534-5490
Illinois Institute of Technology, USA

Sabat Salih Muhamad
Soran University, Iraq

Firas Mahmood Mustafa
http://orcid.org/0000-0002-8757-5303
Duhok Polytechnic University, Iraq

Aquil Mirza Mohammed
http://orcid.org/0000-0001-7756-4363
Hong Kong Polytechnic University, Hong Kong

Sharyar Wani
http://orcid.org/0000-0001-6812-0066
International Islamic University Malaysia, Malaysia

ABSTRACT

Artificial Intelligence (AI) is increasingly integrated into education and healthcare systems, offering transformative benefits such as personalized learning and precision medicine. However, the adoption of AI in these critical domains necessitates a strong emphasis on explainability and ethical considerations to ensure trust, accountability, and societal acceptance. This chapter explores the intersection of explainable AI (XAI) and ethical AI in education and healthcare, emphasizing how transparent and interpretable models can foster informed decision-making and equitable outcomes. It highlights current challenges, regulatory landscapes, and the role of interdisciplinary collaboration in developing AI systems that are both innovative and aligned with human values. By bridging trust, transparency, and innovation, the chapter provides a comprehensive framework for deploying responsible AI that benefits both learners and patients alike.

DOI: 10.4018/406021

INTRODUCTION

Artificial Intelligence (AI) is rapidly transforming the foundational structures of education and healthcare—two of the most sensitive and impactful sectors in society. As AI systems increasingly influence decision-making, diagnosis, assessment, treatment plans, and personalized learning, concerns around ethics, accountability, and transparency have intensified (Akhtar, Kumar, & Nayyar, 2024). Explainable AI (XAI) has emerged as a response to these concerns, aiming to make algorithmic decision-making interpretable and transparent, particularly when AI outcomes significantly affect human lives (Barnes & Hutson, 2024; Alam, Kaur, & Kabir, 2023).

The concept of explainability refers to the capacity of an AI system to provide understandable justifications for its predictions or decisions (Chamola et al., 2023). Foundational scholarship in AI ethics and interpretability has long emphasized that transparency is not merely a technical requirement but a normative obligation, particularly in high-stakes domains (Albahri et al., 2023; Kiseleva, Kotzinos, & De Hert, 2022; Moorthy et al., 2025). Seminal ethical frameworks argue that AI systems must be accountable, contestable, and aligned with human values, especially when they influence rights, opportunities, or wellbeing. In parallel, early explainability research introduced model-agnostic explanation techniques such as Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), which remain cornerstones of modern XAI research. These foundational perspectives provide the theoretical basis upon which contemporary explainable and ethical AI frameworks in education and healthcare continue to evolve (Fatima et al., 2025; Khosravi et al., 2022).

However, technical explainability alone is insufficient. Ethical AI frameworks are essential to ensure that the deployment of AI technologies aligns with principles such as fairness, accountability, inclusiveness, and respect for human autonomy (Díaz-Rodríguez et al., 2023; Nasir, Khan, & Bai, 2024). This dual emphasis on explainability and ethics has led to the conceptualization of trustworthy AI—systems that are not only transparent but also governed by ethical and legal standards (Herzog, Blank, & Stahl, 2024; Khan et al., 2024). The need for such systems is particularly acute in global healthcare and educational ecosystems, where sociocultural disparities, digital divides, and data sensitivity add layers of complexity (Kong et al., 2023; Zangana & Omar, 2025).

Explainable and ethical AI offers pathways to increase stakeholder trust and encourage responsible innovation. As highlighted by Haque (2024), fostering human collaboration and value-centered design is critical in leveraging AI for the public good. In education, this means providing transparency in AI-driven tutoring systems, exam grading algorithms, or admission recommendations (Salloum, 2024). In healthcare, it means ensuring that AI-assisted diagnostics, triage systems, and risk prediction models are free from biases and can be scrutinized by both clinicians and patients (Chattopadhyay, Barman, & Lakshmi, 2025; Metta et al., 2024).

Several studies emphasize the integration of interdisciplinary knowledge and stakeholder involvement to achieve trustworthy AI ecosystems (Patel et al., 2024; Singh, 2024; Akhtar, Kumar, & Nayyar, 2024). By incorporating expertise from law, medicine, education, ethics, and computer science, AI developers can create systems that are not only high-performing but also societally responsible (Alahmed et al., 2023; Eke & Shuib, 2024). Furthermore, leadership plays a crucial role in ensuring inclusivity and driving responsible technological adoption across sectors (Zangana & Omar, 2025).

Despite progress, numerous challenges persist. These include the black-box nature of complex machine learning models (Erkiniy, 2025; Marey et al., 2024), lack of standardized regulatory compliance (Moorthy et al., 2025), and insufficient guidelines for ethical governance in