

Transparency and Accountability in AI: Exploring the Role of Explainable AI (XAI)

Godavari Basavaraj, Research scholar, Department of Computer Science, OPJS University, Churu (Raj.)

Abstract

Artificial Intelligence (AI) has rapidly transformed decision-making across sectors such as healthcare, finance, education, employment, governance, and criminal justice. However, the increasing use of AI systems has raised significant concerns regarding transparency, accountability, fairness, and trust. Many advanced AI models operate as complex “black boxes,” making it difficult for users, organizations, and regulators to understand how particular decisions or predictions are generated. Explainable Artificial Intelligence (XAI) has emerged as an important approach to addressing these challenges by providing interpretable and understandable explanations of AI-driven outcomes. This paper explores the role of XAI in promoting transparency and accountability in AI-based decision-making. It examines how explanation methods can help stakeholders understand model behaviour, identify potential biases, assess errors, and challenge inappropriate or harmful decisions. The study also considers the relationship between explainability and responsible AI governance, emphasizing that transparency should extend beyond technical explanations to include information about data sources, model limitations, system objectives, and decision-making processes. At the same time, XAI faces challenges related to the accuracy, consistency, complexity, and meaningfulness of explanations, particularly for highly sophisticated machine-learning models. The paper argues that explainability can contribute significantly to responsible and trustworthy AI, but it should be integrated with broader mechanisms of human oversight, auditing, documentation, and regulatory accountability. Thus, XAI represents an important component of a comprehensive framework for transparent and accountable AI systems.

Keywords: Artificial Intelligence, Explainable AI, XAI, Transparency, Accountability, Responsible AI, Algorithmic Bias, AI Governance, Interpretability, Trustworthy AI.