Explainable AI:
Bridging Transparency and Trust in AI

Michael Oyedele Oyenuga
http://orcid.org/0000-0001-9921-5711
Woxsen University, Hyderabad, India

Thomas Oyetunde Oladele
Woxsen University, Hyderabad, India

ABSTRACT

With the rapid development of Artificial Intelligence (AI) in many areas, demand for explainability and reliability of systems has also increased. AI Explainability (XAI) addresses the study of AI that can be understood by humans, since interpretability is essential for user trust, ethical, and legal issues. This article explores the intricate relations between explainability, transparency, and trust of AI systems and investigates how, on the one hand, transparency can generate trust, support ethical decision making, and engage users and, from the other, inform them about how AI systems work, in the era of XAI, drawing on extant literature and practice in the area of XAI. For this reason, the article considers the relevance of reliable AI for efficient human interaction with AI and the emergence of evidence-based perspectives for users and relevant stakeholders, respectively.

INTRODUCTION

Artificial intelligence (AI) is being adopted increasingly in diverse settings, from marketing to medical imaging systems, and its dependability is critical for the public (Alam et al, 2024). Making benefits ranging from ethics, law and society towards calling to account AI highlighted several ethical, legal and societal considerations (Liu, 2023; Thiebes et al., 2020). Explainable AI (XAI) is at the heart of a disciplined response to these concerns, as it provides users visibility into AI systems' decision-making and thereby addresses the issues of bias and accountability in systems (Esmaeili et al., 2021; Miller, 2019; Antoniadi et al., 2021).

The AI explainability is not simply to peel the layers of the machine capability but is more to better one must comprehend how machines operate in the broader-spectrum of the world and hence it needs to be trusted and understood by humans for effective enabling of the machine (Amann et al., 2020; Balasubramaniam et al., 2020). Users are more likely to accept and trust decisions made by AI when they know the rationale behind them (Kumar & Bargavi, 2024; Dorton et al., 2022). Additionally, the ethical

DOI: 10.4018/406019

considerations of AI decision-making suggest that transparency can help mitigate potential harms and enhance user autonomy (Amann et al., 2020; Stawicka & Anderson, 2022). Explainability is not only about functionality, it is about ensuring that AI systems reflect human values and ethical paradigms, thereby fostering a healthier AI culture (Balasubramaniam et al., 2020; Stawicka & Anderson, 2022).

Further, researchers have also reported that trust and explainability are mutually related. Giving users information about AI processes improves their understanding of them, thus leading to greater trust in, and predictable use of, such processes (Villegas-Ch et al., 2023; Stanton & Jensen, 2021; Balasubramaniam et al., 2022). Organisations that develop AI systems have to ensure that XAI principles are statistically and dynamically followed during the deployment. Thiebes et al., 2020) The necessary focus on users is reinforced by the legal responsibility of companies to be able to justify when their systems fail.

Hence, the road to forming a basis for explainable and trustworthy AI is ultimately no longer a technical, pragmatic issue only, but a matter of accountability and the future practice of making use of AI for human augmentation, well (Liu, 2023; Thiebes et al., 2020; Kwon et al., 2020). In this chapter, we review the various facets of explainability in XAI and examine how these have been mined from the literature through a broad section of innovators looking to quell user trust with translucency in AI. By teasing out the (inter-)relationship between ethical problems, applied suggestions and theoretical concepts, we hope to stimulate a conversation on what kind of governance might be appropriate for emergent AI power.

FOUNDATIONS OF EXPLAINABLE AI

Artificial intelligence (AI) has increased in its pervasiveness in our decision-making processes, as well as the need for transparency into these systems transparency and interpretability. At the heart of Explainable AI (XAI) lies an important mutually intelligible explanation of AI decisions that can foster trust between users and the systems, a prerequisite for ethical AI compliance. Such abstractions necessitate a sophisticated understanding of machine learning, yet gauge the effectiveness of a single, isolated component thereof, rather than the explanatory power of the process, which makes the concept of explainability multidimensional, a confluence of layers relating to computational learning theory, cognitive psychology and ethics, thus emphasizing the need for a holistic implementation and evaluation of explainability.

The core vision of XAI is to address the opaque "black box" characteristics of contemporary advanced AI algorithms, particularly those that use deep learning and complex architectures (Haresamudram et al., 2023; Das & Rad, 2020; Walmsley, 2020; Balasubramaniam et al., 2022; Das & Rad, 2020); this makes it very difficult to probe them for potentially emergent new unintended biases, or inaccuracies. In this regard, researchers and practitioners have come up with various frameworks and methodologies, with the focus being on functional transparency. This kind of transparency enables stakeholders to grasp the functionality of algorithms and demystifies the data and logic supporting AI-originated outputs (HOSAİN et al., 2023; Subías-Beltrán et al., 2024).

Several organisations have proposed principles that encompass the salient attributes of explainability, with specific emphasis on accountability, interpretability, and articulating the rationale behind the decision. For example, the National Institute of Standards and Technology (NIST) have described four principles thought to be vital to the creation of explainable AI systems, and one of these principles is transparency as a means of establishing trust with users (Phillips et al., 2021). In line with these principles, meaningful user interactions with AI systems can offer even more than an explanation (Ehsan et al., 2021).