International Journal of Research and Review
Vol.7; Issue: 11; November 2020
Website: www.ijrrjournal.com
Review Paper *E-ISSN: 2349-9788; P-ISSN: 2454-2237*

Shedding Light on AI: Exploring Explainable AI Techniques

Deekshitha Kosaraju

Independent Researcher, Texas, USA

ABSTRACT

The emergence of Artificial Intelligence (AI) tools has influenced sectors creating a need, for reliable and understandable systems that stakeholders can rely on. Explainable AI (XAI) strives to AI operations enhance transparency and strengthen the trustworthiness of AI systems. This study explores the significance of XAI in fields, such as healthcare and finance where AI driven decisions can carry substantial consequences. Through the integration of XAI these sectors can ensure that AI decisions are reasonable and understandable to those they affect. We delve into XAI approaches, their practical uses, and the inherent challenges in creating AI systems that are both efficient and interpretable. Moreover, we discuss future paths for XAI while highlighting the importance of ongoing innovation to keep up with advancements in AI technology and its applications. This summary lays the groundwork for a conversation on how XAI not only enhances the performance of AI systems but also aligns them with ethical norms and regulatory standards ultimately fostering deeper trust, between humans and machines.

Keywords: Explainable AI (XAI), AI Transparency, AI Ethics, Model Interpretability, AI Regulation, Decision Making in AI, Machine Learning Algorithms, AI in Healthcare, AI in Finance, Real-time AI Systems

1. INTRODUCTION

In times the rapid advancement of Artificial Intelligence (AI) technology has brought significant progress in various sectors. However, there are increasing concerns regarding the transparency and dependability of these AI systems. As AI becomes more complex understanding how they reach decisions is crucial in areas where these decisions hold significant consequences. Explainable AI (XAI) is an area of study focused on making AI systems more transparent and understandable not just for developers but also for end users and stakeholders affected by AI choices [1]. XAI aims to guarantee that as AI becomes more prevalent, in society it respects established norms and societal principles. This overview lays the groundwork for an examination of XAI methods delving into their significance implementation challenges and broader implications when deploying such systems, in real world scenarios [5].

Technique Type	Advantages	Disadvantages	Typical Use Cases
Model-specific	High interpretability tailored to specific models	Limited to specific model types; less flexible	Deep learning models, neural networks
Model-Agnostic	Flexible, applicable to any model	May offer less insight into complex model behaviors	Any AI model, broad applicability

Table 1: Model-specific and Model-Agnostic XAI techniques

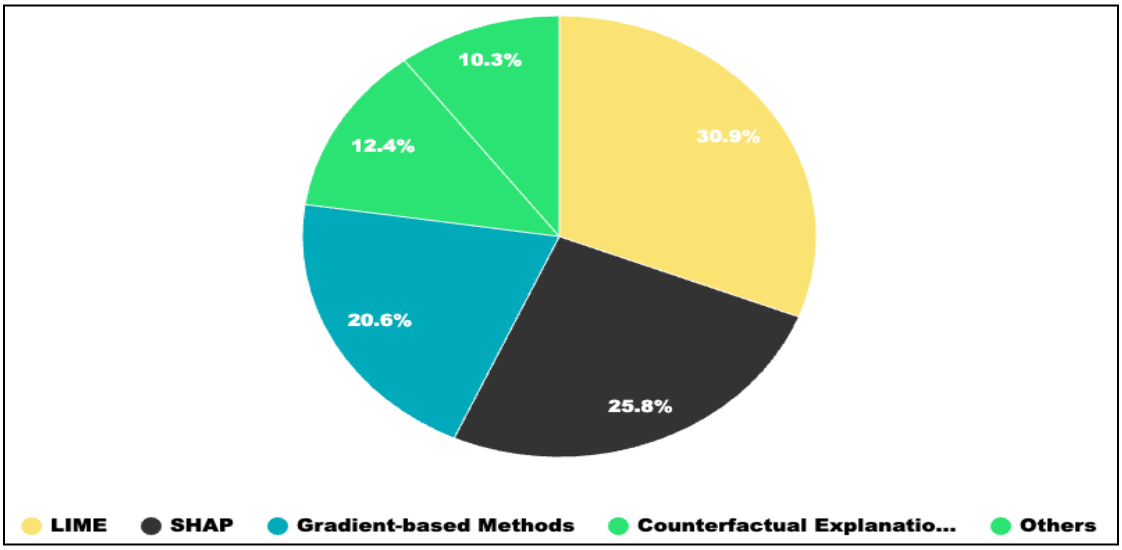
2. Main Body

2.1. Problem Statement

The growing reliance on AI systems in critical sectors has brought attention to a major issue, the lack of transparency within these systems. Often referred to as " box" models these AI systems function without revealing their internal workings making it challenging for users and regulators to comprehend how decisions are reached. This lack of clarity can give rise to ethical and legal concerns especially when these decisions have an impact on human health, legal proceedings, or financial opportunities. For instance, in the healthcare sector an incomprehensible AI decision could lead to treatment plans while within the judicial system it could result in biased sentencing influenced by undisclosed algorithms [7]. The importance of transparency becomes more evident as AI systems continue to influence aspects of human life emphasizing the necessity, for the development of systems that stakeholders can not only utilize but also trust and understand.

2.2. Solution

To tackle the risks associated with AI systems Explainable AI (XAI) techniques have been developed to illuminate the decision-making process of AI. These techniques range from methods that offer insights, into decisions to more complex frameworks that enable a thorough analysis of the entire decision-making process. For example, techniques like SHapley Additive exPlanations (SHAP) help clarify the role of each variable in a decision enabling stakeholders to grasp the importance and impact of factors. Another method, Layer Relevance Propagation (LRP) traces how each input of neural networks influences the final decision unveiling the inner workings of deep learning models. These techniques do not increase transparency in AI systems but also aid in ensuring they adhere to ethical standards by identifying and rectifying any biases that might be present within the models [1].



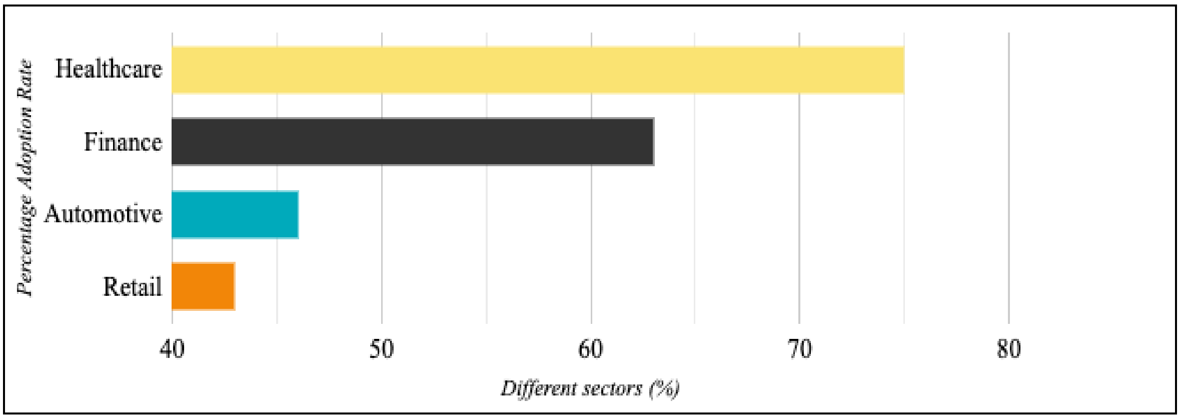
Contribution of Various XAI Techniques to Model Transparency

XAI Technique	User Trust Level (Scale of 1-10)	Feedback Summary
LIME	8	Highly valued for local explanations
SHAP	7	Praised for detailed contribution analysis
Gradient-based Methods	6	Trusted for visual clarity, less intuitive
Counterfactual Explanations	9	Favored for intuitive, scenario-based explanations

Table 2: User Trust Levels Across Different XAI Techniques

2.3. Uses

The practical uses of XAI are extensive and diverse across industries. In the healthcare field XAI assists healthcare providers in making informed decisions by offering explanations for AI generated diagnoses potentially enhancing patient outcomes. In the sector XAI promotes transparency in automated decision-making processes like credit scoring, where it can clarify the reasons behind loan approvals or rejections fostering fairness and reducing biases. This transparency is essential for upholding accountability and trust especially when decisions have financial impacts, on the individuals involved [8].



Adoption Rates of XAI in Various Sectors

2.4. Impact

The use of XAI brings about changes in how AI is implemented improving both the functionality and perceived reliability of AI systems. By increasing transparency in the decision-making processes of AI XAI builds trust with users and stakeholders which's vital for the successful integration of AI technologies into everyday operations. Moreover, XAIs capability to ensure adherence to standards plays a key role, in upholding ethical practices across various industries reducing risks and promoting responsible deployment of AI technologies [5].

2.5. Scope

The future prospects of XAI are closely linked to the evolving landscape of AI technologies. As AI systems become more intricate and their uses broaden the methods for providing explanations must also progress. Future advancements in XAI are expected to concentrate on improving the effectiveness of explanation techniques extending their usability to intricate models and making it easier for users to comprehend explanations. This continual progress in XAI plays a role in ensuring that the advantages of AI benefit society all while upholding ethical standards and responsibility, in AI operations [7].

3. CONCLUSION

With the rise of Artificial Intelligence (AI) in facets of our daily lives there is a growing acknowledgment of the significance of Explainable AI (XAI). XAI not clarifies how AI reaches decisions but also introduces an ethical dimension to ensure that these systems are in line, with societal values and moral principles [1]. By making AI operations transparent and decisions understandable XAI helps build trust and acceptance of AI technologies. This trust is essential for the use of AI in crucial areas like healthcare, law enforcement and finance where decisions can have significant consequences, on people’s lives [7]. Furthermore, XAI plays a role in supporting efforts for regulatory compliance by providing regulators and policymakers with the necessary tools to effectively oversee AI technologies [5]. As regulations adapt to keep up with progress XAI will play a key role in bridging the gap between rapid AI advancements and public policy. The interaction between the advancement of XAI methods and the development of frameworks emphasizes the ongoing necessity for flexible solutions capable of addressing the intricacies of modern AI systems.

Looking ahead the expansion of XAI is anticipated to spur innovation in AI propelling technologies that are not just powerful and efficient but also understandable and manageable for everyday users. With AI systems gaining autonomy the importance of XAI in ensuring these systems operate in a manner that is comprehensible and acceptable to humans cannot be overstated. Ultimately the future progression of AI relies on our capacity to demystify algorithms making the digital realm accessible and secure, for all individuals.

REFERENCES

1. A. Barredo Arrieta et al., “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, vol. 58, no. 1, pp. 82–115, Jun. 2020, Available: https://arxiv.org/pdf/1910.10045.pdf
2. A. Rai, “Explainable AI: from black box to glass box,” Journal of the Academy of Marketing Science, vol. 48, no. 1, pp. 137–141, Dec. 2019, doi: 10.1007/s11747-019-00710-5.
3. Augusto Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, C. M. Aguilera, and S. García, “eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research,” PLOS Computational Biology, vol. 16, no. 4, pp. e1007792–e1007792, Apr. 2020, doi: https://doi.org/10.1371/journal.pcbi.1007792.
4. E. @ T. TRN, “Explainable AI - what is it and why do we need it?,” Medium, https://medium.com/the-research-nest/explainable-ai-what-is-it-and-why-do-we-need-it-261509e48cc
5. “Explainable AI: The basics”, https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf
6. G. Vilone and L. Longo, “Explainable Artificial Intelligence: a Systematic Review.” Available: https://arxiv.org/pdf/2006.00093
7. Payrovnaziri SN;Chen Z;Rengifo-Moreno P;Miller T;Bian J;Chen JH;Liu X;He Z;, “Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review,” Journal of the American Medical Informatics Association : JAMIA, https://pubmed.ncbi.nlm.nih.gov/32417928/.
8. S. Chandler, “How explainable AI is helping algorithms avoid bias,” Forbes, Feb. 18, 2020. [Online]. Available: https://www.forbes.com/sites/simonchandler/2020/02/18/how-explainable-ai-is-helping-algorithms-avoid-bias/?sh=7e0fbfee5ed3

How to cite this article: Deekshitha Kosaraju. Shedding light on AI: exploring explainable AI techniques. International Journal of Research and Review. 2020; 7(11): 608-611.

******