Why Explain a decision made by AI? Explainable AI (XAI)

Rodrigo Almeida

Abstract—In this paper, I will review the literature relating to transparency for decisions made by artificial intelligence (AI). As AI advances, not only our dependence on AI models are increasing but also a need for understanding how such results and decisions were made by the model. Due to the complexity of AI, Explainable Artificial Intelligence (XAI) has become a great choice, bringing more clarity, where outputs can be explained reducing bias and bringing transparency to the decision-making process, as opposed to ”Black Box” algorithms that we cannot explain why or how the AI arrived in such a conclusion/classification/decision, it outputs directly from data, and not even who developed such algorithms are able to explain how the model arrived in such result.

1 Introduction

Explainable AI is emerging, and with it, the concerns from scientists, the general public and policymakers are also on the rise. As artificial intelligence, machine learning and deep learning are becoming integrated into applications, decisions from such models can carry consequences for a person and society, changing the focus from accuracy to the necessity of being able to explain and clarify how these decisions were made.[2] As we are becoming more dependent and relying hugely on intelligent machines, whether it is email filters, AI helping with medical diagnoses or self-driving cars, machine learning is being used across almost all sectors. Such growth in these technologies only highlights the importance of understanding the interpretability of the outcome generated by it. This is the main motivation for explainable artificial intelligence.[4] The impact of the outcome will vary based on the field. A wrong prediction for computing and business might lead to misleading recommendations, potentially affecting the revenue of such businesses, however wrong predictions in critical sectors like healthcare could put human life at risk. Opening the ”black box” is incredibly important, not just for social acceptance but also for regulatory purposes. Understanding how AI systems reach their conclusions is crucial in ensuring accountability, transparency, and ethical use of these technologies, making it essential for both public trust and regulatory compliance.[2]

In Figure 1 we can easily understand the main differences between the ”Black Box” and XAI.[6]

2 Definitions

Explainable AI (XAI) is a field focused on increasing human understanding of AI systems. Although terms like transparency, interpretability, and explainability are frequently used together, they can have different meanings, there are differences between these theories.[3]

- Transparency: a model deemed transparent is one that can be understood easily on its own. It is the opposite of a ”black box”, indicating that the inner workings of the model are understandable[1].

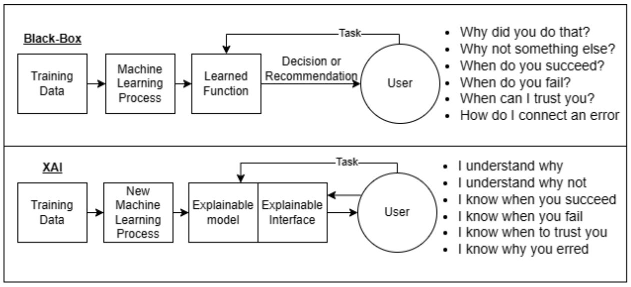

Fig. 1: XAI Concept

- Interpretability: It is the ability to explain concepts that are complex or result in a way that humans will easily understand[5].
- Explainability: it is the concept that provides explanations as a bridge between AI systems and humans. It involves the ability to create AI systems that are not only accurate but also people can understand.[5].

3 Overview of Explainable AI Algorithms

Researchers have developed many algorithms to explain AI systems. These explanations could be categorized into two primary groups: self-interpretable models refer to the algorithm model itself or a representation of the algorithm that can be directly read and comprehended by a human. In this case, the model itself serves as the explanation. On the other hand, post-hoc explanations are descriptions, explanations, or models of the algorithm often produced by separate software tools. These tools aim to provide an understanding of how the algorithm operates. Post-hoc explanations are particularly useful for algorithms for which the inner workings are not fully transparent, as they can be employed to generate insights without requiring deep knowledge of the algorithm’s internal mechanisms. Instead, they rely on querying the algorithm for outputs based on selected inputs.[7].

4 Future Work

5 Conclusion

Appendix A
Proof of the First Zonklar Equation

Appendix B

References

[1] Amina Adadi and Mohammed Berrada. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)". In: *IEEE Access* 6 (2018), pp. 52138–52160. DOI: 10.1109/ACCESS.2018.2870052.

[2] P. Angelov et al. "Explainable artificial intelligence: an analytical review". In: *WIREs Data Mining and Knowledge Discovery* 11 (5 2021). DOI: 10.1002/widm.1424.

[3] Anna Markella Antoniadi et al. "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review". In: *Applied Sciences* 11.11 (2021), p. 5088.

[4] Finale Doshi-Velez and Been Kim. *Towards A Rigorous Science of Interpretable Machine Learning*. 2017. arXiv: 1702.08608 [stat.ML].

[5] Leilani H. Gilpin et al. *Explaining Explanations: An Overview of Interpretability of Machine Learning*. 2019. arXiv: 1806.00069 [cs.AI].

[6] AI concept vs. XAI Concept.

[7] P Jonathon Phillips et al. "Four principles of explainable artificial intelligence". In: *Gaithersburg, Maryland* 18 (2020).