A Path toward Trustworthiness: Possible Solutions of XAI Challenges

Haochen Zhao*
Yuanpei College
Peking University
2100017417@stu.pku.edu.cn

Abstract

Explainable Artificial Intelligence (XAI) plays a pivotal role in ensuring the responsible and ethical deployment of advanced machine learning models. A non-explainable system will never be fully trusted and developed, since we cannot be persuaded by what we do not understand. As the demand for transparency and interpretability continues to grow, future XAI research faces several challenges that must be addressed to enhance the effectiveness and trustworthiness of AI systems. This essay summaries current solutions, explores the major challenges in XAI research and tries to proposes potential avenues for their resolution.

1 Introduction

Understanding the importance of XAI is crucial in navigating the intricate landscape of modern artificial intelligence applications. As AI systems become increasingly integrated into various aspects of our lives as applications like ChatGPT has been more and more popular day after day, from healthcare and finance to autonomous vehicles, the need for transparency and interpretability in decision-making processes cannot be overstated.

XAI serves as a critical bridge between the highly complex inner workings of advanced machine learning models and the human stakeholders who rely on these systems. Trust is fundamental to the successful integration of AI technologies, and XAI plays a pivotal role in building and maintaining that trust. When individuals, businesses, and societies can comprehend and validate the reasoning behind AI-driven decisions, confidence in these technologies is bolstered. Moreover, the importance of XAI extends beyond mere transparency. It is a key enabler of accountability and ethical AI practices. In applications where AI impacts human lives, such as healthcare diagnostics or loan approvals, the ability to explain decisions is not just a preference but a necessity. XAI empowers users to question, understand, and, if necessary, challenge the outcomes of AI models. This transparency fosters a responsible and ethical approach to AI deployment, preventing unintended biases, discrimination, or other adverse consequences.

In the following sections, several major challenges of future XAI research will be introduced. Also, we try to identify essential components in the communicative XAI framework, then talk about how to improve based on the limitation of the components.

2 Model Intrinsic

There is an easy way to separate different components of XAI framework, model intrinsic or Post-hoc.[2] This classification is based on whether the XAI method is integrated to a specific model or can be applied in general. In this part, model intrinsic will be introduced.

*Thanks to course instructor Yixin Zhu, TA Guangyuan Jiang and Yuyang Li for their helpful suggestions.

The reason we call it model intrinsic is that the ability to be explainable is fused into the structure of the model already. There is no need to add any outside modification to show its explainablity. Thus, it depends on the model specifics and can't be generalized to other models.

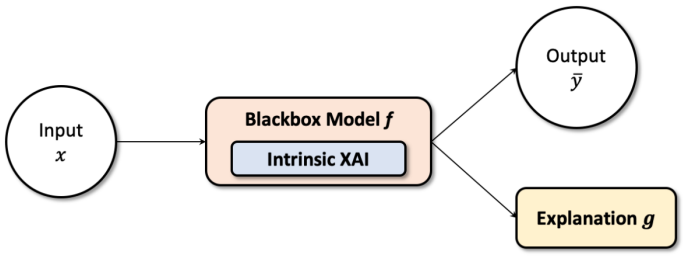

Figure 1: Illustration of model intrinsic in terms of XAI[2]

Typical architecture with model intrinsic include trees and rule-based models[4], LDA[3] and Generalized additive models (GAM)[1] among others. Trees and rule-based models above realized explainablity because we designed the rules it used to make sure every IF/ELSE-THEN reduction it made is under the rules we agreed. But as the models in early stages of AI, these rule-based methods are not flexible enough to handle complex and unseen mission, and designing self-consistent rules to comprehensive scenarios is of great difficulty. LDA was a kind of textual hierarchical labeling. The domain it can be used is limited. GAMs adopted very large decision trees (sometimes millions) to make explainable decisions. But it requires considerable space and time to compute, and is not flexible as well.

Obviously, these methods seemed primitive pre-deep-learning-era solutions. Their performance is poorer than advanced deep learning models. But model intrinsic can't be generalized to different model. As a result, these hand-crafted designs are hard to apply to existing high-accuracy models. So the solution is to develop a new powerful model with intrinsic XAI.

3 Post-Hoc

Since model intrinsic is hard to apply, we need model agnostic methods to take advantage of the already developed powerful models. Post-hoc explanation methodology, as another thinking, is deemed useful as existing accurate models can be improved by adding extra interpretability.[2]

Shapley sampling methods (SHAP)[5] is a typical work. This model agnostic method assigns each feature an importance value for a particular prediction. It achieved better consistency with human intuition compared with previous methods, because it contained different estimation methods. But it was not fast enough. So the improvements can be done in this way.

As large language models (LLMs) like ChatGPT and Llama becomes more and more popular, new methods are emerging as well. A natural thought is that since the model gets bigger and bigger, it is harder and harder to look into the layers of the complex networks for explainabilty. We can just take the LLM as a blackbox and ask it to explain its decision. Following this idea, Chain of Thought (CoT) arrived.[6]

Nowadays, CoT has been widely used as prompting methods are adopted[7]. It is a very simple method: just adding "Let's think step by step" to the prompt. Though the explanation LLM provides, it can achieve higher performance and become more explainable as well. Thus this simple approach is commonly adopted.

4 Conclusion

In summary, the importance of XAI lies at the intersection of technical advancement, ethical considerations, and societal trust. By addressing the current solutions anf challenges in XAI research and prioritizing its development, we pave the way for a future where artificial intelligence is not only powerful but also accountable, transparent, and aligned with human values. Embracing XAI is not

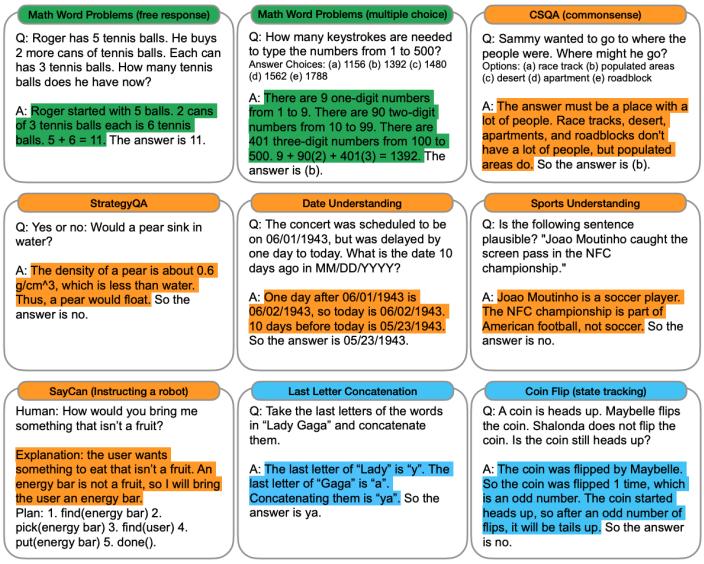

Figure 2: Examples of CoT usage[6]

just a technological choice; it is a commitment to a future where AI augments human capabilities while respecting individual rights and societal well-being.

References

[1] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015. 2

[2] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020. 1, 2

[3] Finale Doshi-Velez, Byron Wallace, and Ryan Adams. Graph-sparse lda: a topic model with structured sparsity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 2

[4] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. 2015. 2

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 2

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 2, 3

[7] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 2