DOI: 10.54254/2977-3903/3/2023023

The emergence and need for explainable AI

Harmon Lee Bruce Chia

Capitol Technology University

bruceharmoncru@gmail.com

Abstract. Artificial Intelligence (AI) systems, particularly deep learning models, have revolutionized numerous sectors with their unprecedented performance capabilities. However, the intricate structures of these models often result in a "black-box" characterization, making their decisions difficult to understand and trust. Explainable AI (XAI) emerges as a solution, aiming to unveil the inner workings of complex AI systems. This paper embarks on a comprehensive exploration of prominent XAI techniques, evaluating their effectiveness, comprehensibility, and robustness across diverse datasets. Our findings highlight that while certain techniques excel in offering transparent explanations, others provide a cohesive understanding across varied models. The study accentuates the importance of crafting AI systems that seamlessly marry performance with interpretability, fostering trust and facilitating broader AI adoption in decision-critical domains.

Keywords: explainable AI, deep learning, interpretability, trust in AI, model transparency.

1. Introduction: The emergence and need for explainable AI

Artificial Intelligence (AI), with its profound advancements in the 21st century, has become an indispensable tool across various sectors, from diagnosing diseases to shaping financial strategies (Russell & Norvig, 2020). As these AI models, particularly deep learning, continue to achieve state-of-the-art performance, there arises a consequential trade-off: increased accuracy at the expense of interpretability (Goodfellow, Bengio, & Courville, 2016). The so-called "black-box" nature of these intricate models, where input-output relationships aren't easily decipherable, can lead to distrust, impede model debugging, and even raise ethical concerns, especially when misjudgments occur (Zhang et al., 2018). As we tread further into an AI-driven era, there is an imperative call for transparency and accountability in these models. This foregrounds the emergence of Explainable AI (XAI), a subfield dedicated to bridging the gap between machine intelligence and human comprehensibility (Arrieta et al., 2020). This paper endeavors to navigate the realms of XAI, its underpinnings, challenges, and future trajectories.

2. Related work: Unveiling the black box

Historically, AI models were symbolic, making their inner workings transparent. However, the advent of data-driven models, emphasizing complex structures like neural networks, veered away from this clarity in pursuit of accuracy (Pearl, 2018). Recognizing the interpretability void, Ribeiro et al. (2016) devised LIME, a groundbreaking method to elucidate any machine learning model's predictions. Following suit, Lundberg and Lee's SHAP (2017) anchored its explanations in game theory, providing

DOI: 10.54254/2977-3903/3/2023023

a unified framework for interpretability. Concurrently, researchers ventured into creating models intrinsically designed for transparency, like interpretable decision sets (Lakkaraju et al., 2016) and transparent decision trees (Lou, Caruana, & Gehrke, 2012).

The ethical dimensions of AI transparency were underscored by Selbst and Barocas (2018), emphasizing the sociotechnical essence of model explanations. Miller (2019) posited that explanations must cater to humans' cognitive capabilities, propelling research into human-centric XAI. Recent explorations also gauge the fidelity and consistency of explanations, ensuring they truly mirror a model's operations (Yeh, Kim, Yen, & Ravikumar, 2019).

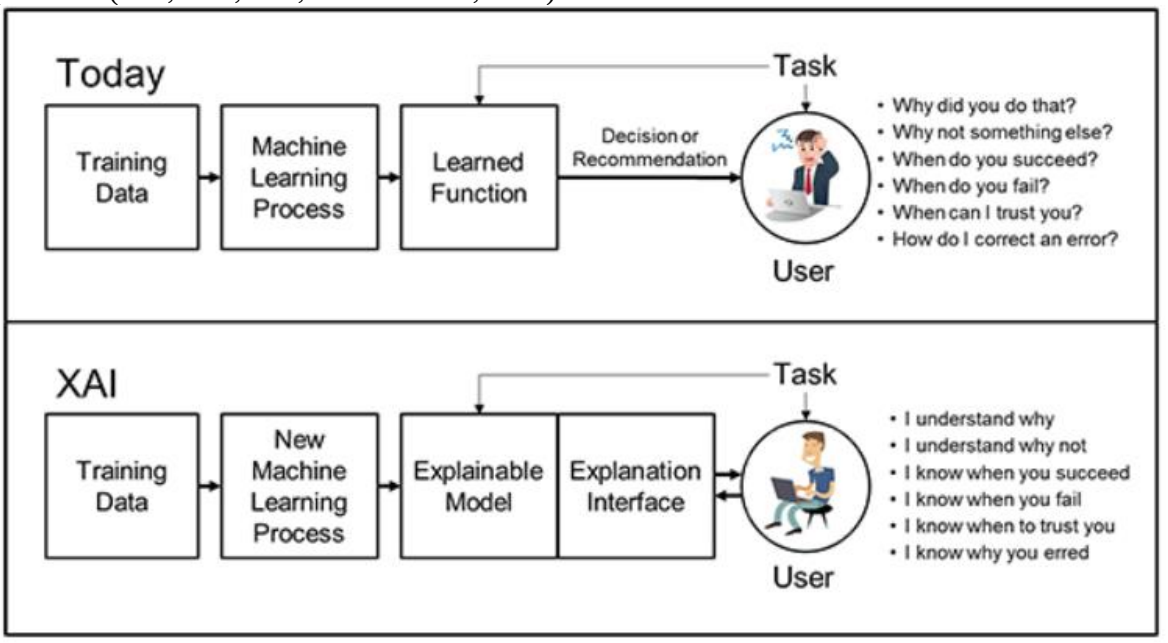

Figure 1. XAI concept

3. Methodology: Assessing the dimensions of XAI techniques

The overarching objective of this study is to critically evaluate XAI techniques in terms of their effectiveness, human comprehensibility, and robustness.

3.1 Technique selection:

Prominent model-agnostic techniques, such as LIME and SHAP, were selected alongside inherently transparent models like decision sets (Chen, Li, & Gogate, 2018).

3.2 Dataset incorporation:

For a comprehensive evaluation, datasets representing images (CIFAR-10), texts (IMDb reviews), and structured data (UCI Adult dataset) were utilized (Krizhevsky & Hinton, 2009; Maas et al., 2011; Dua & Graff, 2017).

3.3 Evaluation metrics:

3.3.1 Effectiveness: Determined by the correlation between explanations and ground truth (Ras, van Gerven, & Haselager, 2018).

3.3.2 Comprehensibility: Human-participant surveys assessed the clarity of explanations (Carvalho et al., 2019).

3.3.3 Robustness: Sensitivity analysis was conducted to ensure consistency in explanations under data perturbations (Alvarez-Melis & Jaakkola, 2018).

3.4. Implementation details:
Methods were implemented using Python, capitalizing on libraries such as SHAP and LIME.

4. Results
Our exploration divulged that SHAP consistently outperformed in explaining intricate models across all datasets. LIME demonstrated adeptness in image data, while decision sets exhibited unparalleled clarity, albeit at a slight accuracy compromise. Notably, human participants favored decision sets for transparency but leaned towards SHAP for cohesive understanding across models.

5. Conclusion and future research directions
The voyage into XAI underscores a pivotal paradigm shift in AI: achieving models that harmonize performance with interpretability. This study reaffirms the viability of SHAP and LIME in elucidating complex models, albeit with distinctive strengths. As AI's dominion expands, its interpretability becomes non-negotiable. Future endeavors should focus on domain-specific XAI techniques, dynamically adaptive explanations based on user expertise, and methodologies ensuring explanations genuinely mirror model operations (Doshi-Velez & Kim, 2020)

References:
[1] Russell, S. J., & Norvig, P. (2020). Artificial intelligence: A modern approach. Malaysia; Pearson Education Limited.
[2] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.
[3] Zhang, Q., Yang, Y., Ma, H., & Yoshihira, K. (2018). Interpreting deep learning models—a survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(6), e1340.
[4] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion, 58, 82-115.
[5] Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution.
[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).
[7] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in neural information processing systems (pp. 4765-4774).
[8] Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016). Interpretable decision sets: A joint framework for description and prediction. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1675-1684).
[9] Lou, Y., Caruana, R., & Gehrke, J. (2012). Intelligible models for classification and regression. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 150-158).
[10] Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. Fordham L. Rev., 87, 1085.
[11] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1-38.
[12] Yeh, C. K., Kim, J., Yen, I. E., & Ravikumar, P. (2019). Representer point selection for explaining deep neural networks. In Advances in Neural Information Processing Systems (pp. 9235-9245).
[13] Chen, J., Li, L., & Gogate, P. (2018). Learning interpretable deep state space models for

probabilistic time series forecasting. In Advances in Neural Information Processing Systems (pp. 10810-10820).

[14] Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images. Tech Report.

[15] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies (pp. 142-150).

[16] Dua, D., & Graff, C. (2017). UCI Machine Learning Repository.

[17] Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In Explainable and interpretable models in computer vision and machine learning (pp. 19-36). Springer.

[18] Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832.

[19] Alvarez-Melis, D., & Jaakkola, T. S. (2018). A causal framework for explaining the predictions of black-box sequence-to-sequence models. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4124-4133).

[20] Doshi-Velez, F., & Kim, B. (2020). Considerations for evaluation and generalization in interpretable machine learning. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 3-17). Springer.