Article

Not peer-reviewed version

"How Good Is Your Explanation?": Towards a Standardised Evaluation Approach for Diverse XAI Methods on Multiple Dimensions of Explainability

Aditya Bhattacharya * and Katrien Verbert

Posted Date: 29 May 2024

doi: 10.20944/preprints202405.1964.v1

Keywords: Explainable AI; XAI; Explainable AI Evaluation

Article

"How Good Is Your Explanation?": Towards a Standardised Evaluation Approach for Diverse XAI Methods on Multiple Dimensions of Explainability †

Aditya Bhattacharya * and Katrien Verbert

KU Leuven, Leuven, Belgium; katrien.verbert@kuleuven.be

* Correspondence: aditya.bhattacharya@kuleuven.be

† Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (UMAP Adjunct'24), July 1–4, 2024, Cagliari, Italy.

Abstract: Artificial Intelligence (AI) systems involve diverse components, such as data, models, users and predicted outcomes. To elucidate these different aspects of AI systems, multifaceted explanations that combine diverse explainable AI (XAI) methods are beneficial. However, popularly adopted user-centric XAI evaluation methods do not measure these explanations across the different components of the system. In this position paper, we advocate for an approach tailored to evaluate XAI methods considering the diverse dimensions of explainability within AI systems using a normalised scale. We argue that the prevalent user-centric evaluation methods fall short of facilitating meaningful comparisons across different types of XAI methodologies. Moreover, we discuss the potential advantages of adopting a standardised approach, which would enable comprehensive evaluations of explainability across systems. By considering various dimensions of explainability, such as data, model, predictions, and target users, a standardised evaluation approach promises to facilitate both inter-system and intra-system comparisons for user-centric AI systems.

Keywords: explainable AI; XAI; explainable AI evaluation

1. Introduction

Artificial Intelligence (AI) systems constitute multiple components, such as the data, models, predictions and the users who will use such systems. Explainable AI (XAI) methods are designed to elucidate these different components of "black-box" AI systems [1,2]. The act of offering explanations throughout the distinct components of an AI system is also referred to as *dimensions of explainability* [3, 4]. Recent works have demonstrated the benefits of combining different types of XAI methods in multifaceted explanations, particularly for non-expert users in AI [5–7].

Additionally, it has been argued that the *No Free Lunch* theorem in Machine Learning [8] (i.e., there is no one-size fits all algorithm for all tasks or datasets) also applies to XAI, as specific XAI methods can only explain certain specific dimensions of explainability [7]. Thus, it is essential to include multiple types of XAI methods for a holistic explanation of AI systems. For example, an XAI system with both feature importance explanations and counterfactual explanations can provide more holistic explanations rather than a system using only one of these methods. However, these methods elucidate completely different dimensions of explainability. The former tries to explain the importance of certain factors considered by the model for generating predictions (i.e., model-centric explanations), whereas the latter provides various conditions for obtaining a different prediction (i.e., outcome-centric explanations). Therefore, a specific XAI method can elucidate a specific dimension of explainability.

The process of user-centric evaluation for XAI predominantly involves user-reported scores for metrics such as *understandability*, *trust*, *actionability*, *stability*, *usefulness*, and etc., from the participants involved in user studies [9–12]. Despite the significance of such methods, their implications can be limited for evaluating an XAI system as they do not consider different explainability dimensions other than the user perspective. Moreover, it could be difficult to compare diverse XAI methods only based on user perspectives due to certain limitations, such as users agreeing to misleading explanations even when the system makes incorrect predictions [13–15].

 doi:10.20944/preprints202405.1964.v1

To overcome these challenges, researchers have also adopted other approaches that provide an objective evaluation of XAI methods, such as task-driven approaches [16,17], algorithmic evaluation metrics [18], and model-specific measurement metrics similar to Quantus [19]. However, estimating the individual impact of a particular XAI method becomes challenging in the presence of multiple combined XAI methods within a system. Thus, despite various objective and subjective metrics for measuring XAI methods [20], comparing diverse methods considering the different dimensions of explainability remains an unsolved problem [5]. To enable comparative evaluation studies, there is an evident need for a normalised XAI evaluation approach that considers both subjective and objective metrics across the various dimensions of explainability. This necessity raises an essential open research question: "*How can we compare XAI methods elucidating different dimensions of explainability when used in XAI systems?*"

In this position paper, we discuss the necessity of a standardised approach for evaluating diverse explainability methods used within XAI systems, considering the diverse explainability dimensions. We also discuss the potential advantages of adopting a standardised approach, which would enable comprehensive evaluations of explainability across systems. We believe that existence of such an approach can be extremely beneficial for the Human Centred XAI (HCXAI) community for evaluating XAI systems.

2. The Needs for a Standardised Evaluation Approach for Explainable AI Systems

A standardised evaluation methodology for assessing the explainability of diverse XAI methods within AI systems could effectively mitigate challenges associated with cross-dimensional comparisons. Such an approach should integrate both objective and subjective evaluation metrics, standardised to a consistent scale. By normalising evaluation scores across dimensions, such an approach could facilitate meaningful comparisons of the efficacy of various XAI methods in elucidating the multifaceted components of AI systems.

We highlight the following needs that could be fulfilled by a standardised evaluation approach for XAI systems:

- *Holistic Measurement*: This metric should provide a standardised and holistic measure of the effectiveness of different XAI methods in elucidating the multiple components of AI systems. This addresses the need for a comprehensive evaluation that goes beyond individual metrics.
- *Flexibility*: The standardised approach should offer flexibility in incorporating existing evaluation metrics across individual dimensions of explainability. This should encompass objective measures assessing training data quality, model performance, and considerations for prediction uncertainty alongside user-centric evaluations such as trustworthiness, understandability, and others.
- *Model-Agnostic Property*: This normalised evaluation approach should be model-agnostic, i.e., it could be applied to evaluate any XAI method or AI models. A model-agnostic evaluation approach could broaden its applicability to different application domains and diverse AI systems.
- *Intra-System and Inter-System Comparison*: The normalised approach should be used for comparing two or more XAI systems. It should also allow individual XAI methods across the different explainability dimensions. However, the main goal of such an approach is to compare different XAI methods within a system, using both subjective and objective measures.

3. Summary

To summarise, in this position paper, we discuss the needs for an approach designed to assess different dimensions of explainability of diverse XAI methods within XAI systems. We also discuss the benefits and opportunities for leveraging this approach for evaluating XAI systems. We eagerly seek feedback from the workshop participants to refine and operationalise these ideas into a robust evaluation framework.

 doi:10.20944/preprints202405.1964.v1

Acknowledgments: We thank Maxwell Szymanski and Robin De Croon for their valuable feedback on this research. This research was supported by Research Foundation–Flanders (FWO grants G0A3319N, G0A4923N and G067721N) and KU Leuven Internal Funds (grant C14/21/072) [21].

References

1. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.
2. Cai, C.J.; Reif, E.; Hegde, N.; Hipp, J.; Kim, B.; Smilkov, D.; Wattenberg, M.; Viegas, F.; Corrado, G.S.; Stumpe, M.C.; Terry, M. Human-Centered Tools for Coping with Imperfect Algorithms during Medical Decision-Making, 2019. https://doi.org/10.48550/ARXIV.1902.02960.
3. Bhattacharya, A. Applied Machine Learning Explainability Techniques. In *Applied Machine Learning Explainability Techniques*; Packt Publishing: Birmingham, UK, 2022.
4. Sokol, K.; Flach, P. Explainability fact sheets: a framework for systematic assessment of explainable approaches. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA, 2020; FAT* '20, p. 56–67. https://doi.org/10.1145/3351095.3372870.
5. Bhattacharya, A.; Ooge, J.; Stiglic, G.; Verbert, K. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. Proceedings of the 28th International Conference on Intelligent User Interfaces; Association for Computing Machinery: New York, NY, USA, 2023; IUI '23, p. 204–219. https://doi.org/10.1145/3581641.3584075.
6. Bhattacharya, A.; Stumpf, S.; Gosak, L.; Stiglic, G.; Verbert, K. Lessons Learned from EXMOS User Studies: A Technical Report Summarizing Key Takeaways from User Studies Conducted to Evaluate The EXMOS Platform, 2023, [arXiv:cs.LG/2310.02063].
7. Bhattacharya, A.; Stumpf, S.; Gosak, L.; Stiglic, G.; Verbert, K. EXMOS: Explanatory Model Steering Through Multifaceted Explanations and Data Configurations. Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24); ACM: New York, NY, USA, 2024; CHI '24, p. 27. https://doi.org/10.1145/3613904.3642106.
8. Goldblum, M.; Finzi, M.; Rowan, K.; Wilson, A.G. The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning, 2023, [arXiv:cs.LG/2304.05366].
9. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects, 2019, [arXiv:cs.AI/1812.04608].
10. Liao, Q.V.; Zhang, Y.; Luss, R.; Doshi-Velez, F.; Dhurandhar, A. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI, 2022, [arXiv:cs.AI/2206.10847].
11. Nauta, M.; Trienes, J.; Pathak, S.; Nguyen, E.; Peters, M.; Schmitt, Y.; Schlötterer, J.; van Keulen, M.; Seifert, C. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Comput. Surv.* **2023**, *55*. https://doi.org/10.1145/3583558.
12. Tintarev, N.; Masthoff, J., Designing and Evaluating Explanations for Recommender Systems; 2011; pp. 479–510. https://doi.org/10.1007/978-0-387-85820-3_15.
13. Chromik, M.; Eiband, M.; Buchner, F.; Krüger, A.; Butz, A. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. 26th International Conference on Intelligent User Interfaces; Association for Computing Machinery: New York, NY, USA, 2021; IUI '21, p. 307–317. https://doi.org/10.1145/3397481.3450644.
14. Poursabzi-Sangdeh, F.; Goldstein, D.G.; Hofman, J.M.; Wortman Vaughan, J.W.; Wallach, H. Manipulating and Measuring Model Interpretability. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; Association for Computing Machinery: New York, NY, USA, 2021; CHI '21. https://doi.org/10.1145/3411764.3445315.
15. Schneider, J.; Meske, C.; Vlachos, M. Deceptive XAI: Typology, Creation and Detection. *SN COMPUT. SCI.* **2024**, *5*, 81. https://doi.org/10.1007/s42979-023-02401-z.
16. Doshi-Velez, F.; Kim, B. Towards A Rigorous Science of Interpretable Machine Learning, 2017, [arXiv:stat.ML/1702.08608].
17. Morrison, K.; Jain, M.; Hammer, J.; Perer, A. Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game with a Purpose. *Proc. ACM Hum.-Comput. Interact.* **2023**, *7*. https://doi.org/10.1145/3610064.

 doi:10.20944/preprints202405.1964.v1

18. Agarwal, C.; Krishna, S.; Saxena, E.; Pawelczyk, M.; Johnson, N.; Puri, I.; Zitnik, M.; Lakkaraju, H. OpenXAI: Towards a Transparent Evaluation of Model Explanations. Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2022.
19. Hedström, A.; Weber, L.; Krakowczyk, D.; Bareeva, D.; Motzkus, F.; Samek, W.; Lapuschkin, S.; Höhne, M.M.M. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond. *Journal of Machine Learning Research* **2023**, *24*, 1–11.
20. Mohseni, S.; Zarei, N.; Ragan, E.D. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Trans. Interact. Intell. Syst.* **2021**, *11*. https://doi.org/10.1145/3387166.
21. Bhattacharya, A. Towards Directive Explanations: Crafting Explainable AI Systems for Actionable Human-AI Interactions. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24); ACM: New York, NY, USA, 2024; CHI EA '24, p. 6. https://doi.org/10.1145/3613905.3638177.