TYPE Editorial
PUBLISHED 10 October 2023
DOI 10.3389/fcomp.2023.1291752

OPEN ACCESS

EDITED AND REVIEWED BY
Andrea Turrini,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Chathurika S. Wickramasinghe
✉ chathuriwicky@gmail.com

RECEIVED 10 September 2023
ACCEPTED 20 September 2023
PUBLISHED 10 October 2023

CITATION
Wickramasinghe CS, Marino D and Amarasinghe K (2023) Editorial: Explainable artificial intelligence.
Front. Comput. Sci. 5:1291752.
doi: 10.3389/fcomp.2023.1291752

Editorial: Explainable artificial intelligence

Chathurika S. Wickramasinghe[1]*, Daniel Marino[2] and Kasun Amarasinghe[3]

[1]Capital One, Richmond, VA, United States, [2]Amazon, Seattle, WA, United States, [3]Machine Learning Department and Heinz School of Public Policy, Carnegie Mellon University, Pittsburgh, PA, United States

KEYWORDS

explainable artificial intelligence (XAI), interpretable machine learning, human-system interaction (HSI), trustworthy AI, model governance, artificial intelligence

Editorial on the Research Topic
Explainable artificial intelligence

Despite rapid advancements in the predictive performance of Machine Learning (ML) algorithms across different domains, there is a hesitancy to adopt complex black-box models in human-ML collaborative systems, especially in mission-critical domains (Adadi and Berrada, 2018; Gunning et al., 2019). Appropriately understanding the ML model's decision-making process, adequately trusting its predictions, and understanding its weaknesses to provide corrective actions enable more effective human-ML collaboration. To address this gap, the field of Explainable/Interpretable AI (XAI) has received increased attention in recent years. In this growing body of work, the main goal is to develop ML systems that can explain the rationale behind predictions, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future without compromising predictive performance.

In this backdrop, our goal was to provide researchers a venue to publish their work in XAI methodological research in four overlapping areas: Explaining to justify, Explaining to improve, Explaining to control, and Explaining to discover. Further, we emphasize evaluation strategies for XAI methods that connect with specific use cases, propose evaluation metrics, and assess real-world impact. We believe these areas represent crucial areas in XAI literature that require further attention. Our Research Topic provided a platform for interdisciplinary research that combines computer science, machine learning, and social science methods to design, develop, and evaluate XAI systems. We accepted five articles with novel ideas and techniques, with a focus on human-in-the-loop XAI systems. This editorial summarizes their contributions and provides the editors' points of view on future research directions for XAI.

One of the main focus areas of this topic is application-grounded and human-grounded methods and metrics for evaluating the effectiveness of explainable ML methods in different settings. These methods enable measuring how effective an XAI method is at informing human-ML collaboration and are necessary for benchmarking the large body of methodological work in the field (Silva et al., 2022). Hoffman, Mueller et al. present a framework for achieving a pragmatic understanding of AI systems. The proposed framework enables developers and researchers to (1) assess the a priori goodness of explanations, (2) assess users' satisfaction with explanations, (3) reveal users' mental model of an AI system, (4) assess user's curiosity or need for explanations, (5) assess whether the user's trust and reliance on the AI are appropriate, and (6) assess how the human-XAI system performs at a given task. This work further contributes through an extensive literature survey on the topic and psychometric evaluations of these approaches.

The fundamental question of "Does the outcome of an XAI system provide enough depth for the user's sensemaking?" remains a challenging question. Hoffman, Jalaeian et al. approach this through a cognitive theory perspective and present a framework—that they refer to as an *Explanation Scorecard*—for reflecting the depth of an explanation, i.e., the degree to which an explanation supports the user's sensemaking. Their approach allows users to conceptualize how to extend their machine-generated explanations to support the development of a mental model. This work provides a base for converging AI systems and human cognition to build appropriate trust in the XAI system.

Model-agnostic XAI methods allow the decoupling of explanation generation methods and machine learning models, providing more flexibility over model-specific explainable methods (Molnar, 2023). Björklund et al. present a novel model-agnostic explanation method—*SLISE*—for interpreting individual predictions of black box models. Unlike many popular model-agnostic XAI methods, *SLISE* does not require generating artificial samples. Experimental results show that *SLISE* can generate more generalizable explanations. Further, the authors show that it is usable across different domains as it can handle different input data types.

Trustworthy AI is a widely discussed research area with strong parallels to XAI.[1,2] Although transparency is often regarded as a fundamental stepping stone to achieving trustworthy AI, it has been difficult to measure a direct correlation between transparency and trust. Scharowski et al. present a study of the influence that human-centered explanations have in user's trust in AI systems. In the study, they use reliance on AI recommendations as a measure of trust. They consider two human-centered *post-hoc* explanations: feature importance and counterfactuals (Molnar, 2023). Although they find counterfactuals have an effect on reliance, the type of decision made by the AI has a larger influence. They conclude that trust does not necessarily equate to reliance and emphasize the importance of appropriate, validated, and agreed-upon metrics to design and evaluate human-centered AI.

Among the different explanation types existing in the literature, counterfactual explanations allow users to explore how the outcome of a ML model changes with perturbations to the model inputs (Mothilal et al., 2020; Poyiadzi et al., 2020). While counterfactual explanations promise an avenue for exploring the ML model's decision-making process, their usability is yet to be thoroughly explored. To address this gap, Kuhl et al. present *Alien Zoo,* a game-inspired, web-based experimental framework that allows evaluation of the usability of counterfactual explanations aimed at extracting knowledge from AI systems. Their proof-of-concept result demonstrates the importance of qualitatively and quantitatively measuring the usability of XAI approaches.

In this Research Topic, we found a couple of potentially noteworthy future research directions for advancing XAI:

- Application-grounded and Human-grounded methods and metrics for evaluating the effectiveness of explainable ML methods at improving human-ML collaboration.
- Effective customizations of existing explainable ML methods and their outputs to satisfy requirements of practical use cases.
- Designing and developing novel explainable machine learning methods with targeted use cases.
- Defining metrics and measures for comparing and benchmarking XAI methods.

We hope that readers will find this Research Topic as a useful reference for the emerging field of XAI.

Author contributions

CW: Writing—original draft, Writing—review and editing. DM: Writing—review and editing, Writing—original draft. KA: Writing—review and editing, Writing—original draft.

Conflict of interest

CW was employed by Capital One. DM was employed by Amazon.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

[1] Ensuring Safe, Secure, and Trustworthy AI. The White House, United States of America. Available online at: https://www.whitehouse.gov/wp-content/uploads/2023/07/Ensuring-Safe-Secure-and-Trustworthy-AI.pdf.

[2] Trustworthy AI, IBM Research. Available online at: https://research.ibm.com/topics/trustworthy-ai.

References

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Sci. Robot.* 4, eaay7120. doi: 10.1126/scirobotics.aay7120

Molnar, C. (2023). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, *2nd Edn*. Available online at: https://christophm.github.io/interpretable-ml-book/ (accessed September 6, 2023).

Mothilal, R. K., Sharma, A., and Tan, C. (2020). "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Association for Computing Machinery), 607–617. doi: 10.1145/3351095.3372850

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020). "FACE: feasible and actionable counterfactual explanations," In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)* (New York, NY: Association for Computing Machinery), 344–350.

Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., and Gombolay, M. (2022). Explainable artificial intelligence: evaluating the objective and subjective impacts of xAI on human-agent interaction. *Int. J. Hum. Comp. Interact.* 39, 1390–1404. doi: 10.1080/10447318.2022.2101698