IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

Explainable AI (XAI): History, Basic Ideas and Methods

Prachi Zodage, Hussain Harianawala, Hafsa Shaikh, Asad Kharodia
M.H. Saboo Siddik College of Engineering, Mumbai, Maharashtra, India

Abstract: *Explainable Artificial Intelligence (XAI) is a field that aims to make artificial intelligence (AI) processes more transparent, explainable, and understandable. As AI processes become more complex, the need to reveal the "black box" nature of these models and provide explanations for their results and decisions is increasing. XAI aims to bridge the gap between the opaque inner workings of AI and human understanding by creating AI that is accurate, useful, and can explain reasoning and decision-making processes in ways that humans can understand. XAI's importance stems from several factors. First, it addresses trust and accountability issues in AI systems, particularly in high-risk sectors like healthcare, finance, and technology. By providing explanations for AI decisions, stakeholders can better understand the logic behind them, detect inconsistencies, and ensure moral and administrative compliance. Second, XAI encourages collaboration and decision-making between people and intelligence, allowing experts and decision-makers to use their knowledge and experience to make better decisions. Thirdly, XAI plays a crucial role in modeling, debugging, and continuous improvement by identifying flaws, biases, or inconsistencies and working to improve performance standards and reliability. Various methods and techniques are used in XAI, each with their own advantages and limitations. Model-free explanations such as LIME, Anchor and SHAP are particularly important because they can be applied to any AI model, regardless of its design or complexity.*

Keywords: Explainable Artificial Intelligence, intelligible machine learning, Interpretability

REFERENCES

[1] Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P. & Holzinger, A. (2018). “Explainable AI: The new 42?”. Paper presented at the CD-MAKE 2018, 27-30 Aug 2018, Hamburg, Germany. doi: 10.1007/978-3-319-99740-7_21.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. “ Deep learning. ” Nature, 521:436 EP –, 05 2015.

[3] Andreas Holzinger, Anna Saranti, Christoph Molnar, Przemyslaw Biecek, and Wojciech Samek, “ Explainable AI Methods - A Brief Overview ” Springer Link, 17 April 2022.

[4] Robert R. Hoffman, Shane T. Mueller, Gary Klein, Jordan Litman, “Metrics for Explainable AI: Challenges and Prospects”, Presented with arXivLabs ,https://doi.org/10.48550/arXiv.1812.04608 ,(11 Dec 2018).

[5] Feiyu Xu, Hans Uszkoreit , Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu, “ Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges ”, DOI:10.1007/978-3-030-32236-6_51 In book: Natural Language Processing and Chinese Computing (pp.563-574) (September 2019).

[6] Dwivedi Rudresh, Dave Devam, Naik Het, Singhal Smiti, Rana Omer, Patel Pankesh, Qian, Bin, Wen Zhenyu, Shah Tejal, Morgan Graham and Ranjan Rajiv. “Explainable AI (XAI): core ideas, techniques and solutions. ACM Computing Surveys, Publishers page: http://dx.doi.org/10.1145/3561048 (2023)

[7] Alexey Ignatiev Monash University, Australia alexey.ignatiev@monash.edu , “ Towards Trustable Explainable AI ”, Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) Early Career Track.

[8]GiorgioVisani, “LIME: explain Machine Learning predictions” , Published in Towards Data Science on Dec 18, 2020.

[9] “What is Local Interpretable Model-Agnostic Explanations (LIME)?” , blog on C3.ai What is Local Interpretable Model-Agnostic Explanations (LIME)?.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, April 2024

[10] Tobias Goerke & Magdalena Lang , "Scoped Rules (Anchors)" from book "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable ", Christoph Molnar (2023-08-21).

[11] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems (2017).

[12] Alexey Ignatiev, Nina Narodytska, and Joao Marques-Silva. On validating, repairing and refining heuristic ML explanations. CoRR, abs/1907.02509, 2019.