Explainable AI (XAI) Explanation Techniques

Subjects: Computer Science, Artificial Intelligence | Education & Educational Research | Computer Science, Interdisciplinary Applications

Contributor: Saša Brdnik , Vili Podgorelec , Boštjan Šumak

Interest in artificial intelligence (AI) has been increasing rapidly over the past decade and has expanded to essentially all domains. Along with it grew the need to understand the predictions and suggestions provided by machine learning. Explanation techniques have been researched intensively in the context of explainable AI (XAI), with the goal of boosting confidence, trust, user satisfaction, and transparency.

Explainable Artificial Intelligence | learning analytics | XAI | XAI techniques

1. Explainable Artificial Intelligence

The term XAI is best described as “AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future” [1]. Research on XAI shows that introducing explanations to AI systems to illustrate their reasoning to end users can improve transparency, interpretability, understanding, satisfaction, and trust [2][3][4][5]. Observing the explainability techniques with relation to the machine learning models, Barredo et al. [6] presented a taxonomy that separates transparent models (such as decision trees, logistic regression, linear regression, and K-nearest neighbor) that are de facto explainable from models where post-hoc explainability has to be utilized (e.g., support vector machines, convolutional neural networks) to generate their explanations. Post-hoc explanations can be model-agnostic or model-specific. The former can be applied to any machine learning model with no regard to its inner process or representation, while the latter is related to the interpretation and understanding of a specific machine learning model. Various classifications exist for explanations in AI. They can be categorized mainly as global approaches, explaining the entire model, versus local approaches explaining an individual prediction; or as self-explainable models with a single structure versus post-hoc approaches explaining how a model produces its predictions without clarifying the structure of the model [6][7].

Common explainability approaches [6][7] include *global explanations*, which explain how different features/variables affect predictions within the model in question; *feature relevance*, which presents the computed relevance of each feature in the prediction process (simplified displays with a selection of the most important features are often used); and *example-based explanations*, which select a particular instance to explain the model, offering a more model-agnostic approach, which can be local or global. Additionally, *local explanations* are often used in systems for students and focus on a particular instance, independent of the higher-level general model. *Comparison* uses a selection of instances to explain the outcome of other instances on a local level. *Counterfactual explanations* describe a causal situation (i.e., formulated as “If X had not occurred, Y would not have occurred”) and explain and demonstrate the effects of small changes of feature values on the predicted output. *Explanations by simplification* use mentioned techniques to build a new similar yet simplified system (with reduced complexity but similar performance) based on the trained model to be explained. The aforementioned techniques for post-hoc explanations can include visualizations and text explanations. Their selection is conditioned by the type of machine learning model used for prediction.

Lim [8] presented a slightly different classification of ten explanation types, dividing them into model-independent and model-dependent explanation types. Model-independent explanations include *input explanations*, which inform users about the used input sensors and data sources, to ensure understanding of the explanation scope; *output explanations* inform users about all the possible outputs a system can produce; *what explanations* inform users of the system state in terms of output value; and *what if explanations* allow users to speculate about different outcomes by changing the set of user-set inputs. Model-dependent explanations, on the other hand, include *why* explanations, informing users why the output is derived from input values, possibly returning used conditions (rules); *why not* explanations, presenting users with information about why the alternative output was not produced based on the input; *how to* explanations, which provide explanation as to how the desired outcome is generally produced; and *certainty* explanations, which inform users about the certainty of the produced outcome.

Explanations within XAI lack standardization for their design, as well as their evaluation, as confirmed by literature reviews of the field [6][9]. Haque et al. [9] conducted a literature review of the XAI field and extracted major research themes as future research directions: XAI standardization (which includes developing comprehensive guidelines or standards for developing an XAI system), XAI visualization (focus on empirically measuring the explanation quality dimensions), and XAI effects (measuring user perceptions of the transparency, understandability, and usability of XAI systems). Additionally, Mohseni [10] recognized that the XAI design and evaluation methods should be adjusted based on the set goals of XAI research.

2. XAI in Education

AI systems are complex and, by default, suffer from bias and fairness issues. Explanations of AI were introduced in the field of human–computer interaction as a way to allow users to interact with systems that might be faulty in unexpected ways [11]. Explanations allow users to engage with AI systems in an informed manner and adapt their reliance based on the provided explanations [1]. Multiple studies have shown that introducing explanations in tutoring and e-learning systems increases students’ trust. Ooge et al. [5] observed changes in trust after introducing explanations in an e-learning platform for mathematics exercise recommendations. Explanations increased initial trust significantly when measured as a multidimensional construct (consisting of competence, benevolence, integrity, intention to return, and perceived transparency), while no changes were observed with one-dimensional measures. Conati et al. [12] presented students with personalized XAI hints within an intelligent tutoring system, evaluating their usefulness, intrusiveness, understanding, and trust. Providing students with explanations led to higher reported trust, while personalization improved their effectiveness further. The improvement in understanding of the explanations was related to students’ reading proficiency; students with high levels of reading proficiency benefited from explanations, while students with low levels did not. A study of XAI in education [7] analyzed the concepts of fairness, accountability, transparency, and ethics and proposed a framework for studying educational AI tools, including analysis of stakeholders, benefits, approaches, models, designs, and pitfalls.

Displays that aggregate different indicators about learners, learning processes, and/or learning context into visualizations can be categorized as learning analytics (LA) [13]. A systematic review of LA dashboard creation [14] showed that most dashboards (75%) are developed for teachers and that less focus is put on solutions targeted at learners. Additionally, only two observed propositions provided feedback or warnings to users, and only four papers used multiple data sources, indicating that this is an opportunity for future research. It is important to note that LA does not necessarily include AI. In the core literature [15], LA is defined as the “analysis and representation of data about learners in order to improve learning”. It can be conducted using traditional statistical methods or other data analysis approaches without the involvement of AI. Predictive modeling, the base functionality of many LA systems, is not that different from a traditional teacher recognizing which students are struggling in their class and providing them extra help or direction during the semester. The cost of LA utilization is derived from its functionalities; firstly, the predictions and analyses displayed in LA systems are based on estimations and probabilities, which many users fail to understand correctly [5][14][15]. Making decisions based on wrongly understood probabilities is problematic, especially if the output triggers other actions, or self-regulated learning, without the teacher’s involvement [15]. Additionally, there are challenges with privacy, data quality, availability, and fitness of data used in LA solutions in education [16]. On the other hand, there are many benefits of utilizing LA, mainly the improvement of the learning process based on the data available. Furthermore, students can improve their perceptions of the activity and have their personalized analyses available in more depth than a teacher could provide to each student during their limited time [15]. Overview of the trends in education systems [17] has shown that AI has been recognized as a trend in the educational setting, as more and more AI systems are used in LA, learning management systems, and educational data mining [16]. Some of the most common uses of AI [18] include use cases for profiling and prediction, assessment and evaluation, adaptive systems and personalization, and intelligent tutoring systems. Along with AI models, interpretable machine learning and XAI have been gaining interest in LA systems, as they offer a better understanding of the predictive modeling [16]. The trend of including AI in education has resulted in the development of the term artificial intelligence in education (AIEd). This field overlaps with LA. The main benefits of introducing AI in education and in the LA field [19] can be summarized with the development of intelligent agents, personalized learning systems, or environments and visualizations that offer deeper understanding than the classic non-AI analyses.

Related work on predicting students’ course achievement used logs from virtual learning environments [20] along with demographic data [21] and grades [22] in their prediction models. The need for the interpretability of the complex models used in education mining data techniques has been highlighted [23], and explanations of the model’s predictions have been introduced slowly, by [24] offering verbal explanations (i.e., “Evaluation is Pass because the number of assessments is high”), and by [5] offering verbal and visual explanations to students. In a related study, Conijn et al. [25] analyzed the effects of explanations of an automated essay scoring system on students’ trust and motivation in the context of higher education. The results indicated there is no one-size-fits-all explanation for different stakeholders and in different contexts.

3. Measuring Trust and Satisfaction

Various elements can be observed for measuring the effectiveness of an explanation; namely, user satisfaction, trust assessment, mental models, task performance, correctability [1], and fairness [26]. Researchers focused on the first two measures. Researchers followed the definition of trust as provided by Lee [27], defining it as "an attitude that an agent will achieve an individual's goal in a situation characterised by uncertainty and vulnerability". Many scales for assessing trust are presented in the scientific literature, and many of them were created with interpersonal (human-to-human) trust in mind. A considerable research gap is still reported in the studies, focusing on human–AI trust [4][28]. Vereschak et al. [28] surveyed existing methods to empirically investigate trust in AI-assisted decision-making systems. This overview of 83 papers shows a lack of standardization in measuring trust and considerable variability in the study designs and the measures used for their assessment. Most of the observed studies used questionnaires designed to assess trust in automation (i.e., [29][30][31][32]). Numerous factors have been shown to increase users' trust [33]. Transparency has gained much attention, highlighting the need for explanations that make the systems' reasoning clear to humans. However, trust has been found to increase when the reasoning for the AI system's decision is provided and to decrease when information on sources of uncertainty is shared with the user [4].

Explanations cannot be evaluated without measuring the user's satisfaction with the provided explanation, which Hoffman [34] defines as "the degree to which users feel that they understand the AI system or process being explained to them. It is a contextualised, a posteriori judgment of explanations". A similar study measuring trust, explanation satisfaction, and mental models with different types of explanations has been conducted in the case of self-driving cars [35]. The study reported the lowest user satisfaction with causal explanations and the highest levels of trust with intentional explanations, while mixed explanations led to the best functional understanding of the system. Related evaluation of understandability, usefulness, trustworthiness, informativeness, and satisfaction with explanations, generated with popular XAI methods (LIME [36], SHAP [37], and Partial Dependence Plots or PDP [38]) was conducted by [39], reporting higher satisfaction with global explanations with novice users compared to local feature explanations. Comparing the popular methods, PDP performed best on all evaluated criteria.

Comparing levels of explanation satisfaction and trust between different groups of users can be conducted based on various user characteristics. Level of experience and age are (along with personality traits) two of the major user characteristics recognized to affect user performance and preferences in general human–computer interaction. Although the scale from novice to expert is continuous, there is no universally accepted classification and definition of users' level of experience and/or knowledge [40]. Level of experience is recognized as "the relative amount of experience of user segments of the user population" [41]. In higher education, groups of students can be distinguished based on the amount of ECTS (European Credit Transfer and Accumulation System) points they acquired during their studies. ECTS credits express the volume of learning based on the defined learning outcomes and their associated workload [42].

References

1. Gunning, D.; Aha, D. DARPA’s explainable artificial intelligence (XAI) program. AI Mag. 2019, 40, 44–58.
2. Kulesza, T.; Burnett, M.; Wong, W.K.; Stumpf, S. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In Proceedings of the IUI ’15: 20th International Conference on Intelligent User Interfaces, New York, NY, USA, 29 March–1 April 2015; Association for Computing Machinery: New York, NY, USA; pp. 126–137.
3. Kraus, S.; Azaria, A.; Fiosina, J.; Greve, M.; Hazon, N.; Kolbe, L.; Lembcke, T.B.; Muller, J.P.; Schleibaum, S.; Vollrath, M. AI for explaining decisions in multi-agent environments. Proc. AAAI Conf. Artif. Intell. 2020, 34, 13534–13538.
4. Vössing, M.; Kühl, N.; Lind, M.; Satzger, G. Designing Transparency for Effective Human-AI Collaboration. Inf. Syst. Front. 2022, 24, 877–895.
5. Ooge, J.; Kato, S.; Verbert, K. Explaining Recommendations in E-Learning: Effects on Adolescents’ Trust. In Proceedings of the IUI ’22: 27th International Conference on Intelligent User Interfaces, Helsinki, Finland, 22–25 March 2022; Association for Computing Machinery: New York, NY, USA, 2022; pp. 93–105.
6. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf. Fusion 2020, 58, 82–115.
7. Khosravi, H.; Shum, S.B.; Chen, G.; Conati, C.; Tsai, Y.S.; Kay, J.; Knight, S.; Martinez-Maldonado, R.; Sadiq, S.; Gašević, D. Explainable Artificial Intelligence in education. Comput. Educ. Artif. Intell. 2022, 3, 100074.
8. Lim, B.Y.; Dey, A.K. Toolkit to Support Intelligibility in Context-Aware Applications. In Proceedings of the UbiComp ’10: 12th ACM International Conference on Ubiquitous Computing, Copenhagen, Denmark, 26–29 September 2010; Association for Computing Machinery: New York, NY, USA, 2010; pp. 13–22.
9. Bahalul Haque, A.K.M.; Najmul Islam, A.K.M.; Patrick Mikalef, P. Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. Technol. Forecast. Soc. Chang. 2023, 186, 122120.
10. 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, 24.
11. Liao, Q.V.; Varshney, K.R. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv 2022, arXiv:2110.10790.

12. Conati, C.; Barral, O.; Putnam, V.; Rieger, L. Toward personalized XAI: A case study in intelligent tutoring systems. Artif. Intell. 2021, 298, 103503.

13. Schwendimann, B.A.; Rodríguez-Triana, M.J.; Vozniuk, A.; Prieto, L.P.; Boroujeni, M.S.; Holzer, A.; Gillet, D.; Dillenbourg, P. Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research. IEEE Trans. Learn. Technol. 2017, 10, 30–41.

14. Jivet, I.; Scheffel, M.; Specht, M.; Drachsler, H. License to Evaluate: Preparing Learning Analytics Dashboards for Educational Practice. In Proceedings of the LAK ’18: 8th International Conference on Learning Analytics and Knowledge, Sydney, Australia, 7–9 March 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 31–40.

15. Clow, D. An overview of learning analytics. Teach. High. Educ. 2013, 18, 683–695.

16. Mathrani, A.; Susnjak, T.; Ramaswami, G.; Barczak, A. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Comput. Educ. Open 2021, 2, 100060.

17. Rachha, A.; Seyam, M. Explainable AI In Education: Current Trends, Challenges, And Opportunities. In Proceedings of the SoutheastCon 2023, Orlando, FL, USA, 13–16 April 2023; pp. 232–239.

18. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education—Where are the educators? Int. J. Educ. Technol. High. Educ. 2019, 16, 39.

19. Zhang, K.; Aslan, A.B. AI technologies for education: Recent research & future directions. Comput. Educ. Artif. Intell. 2021, 2, 100025.

20. Wang, X.; Guo, B.; Shen, Y. Predicting the At-Risk Online Students Based on the Click Data Distribution Characteristics. Sci. Program. 2022, 2022, 9938260.

21. Kuzilek, J.; Hlosta, M.; Herrmannova, D.; Zdráhal, Z.; Wolff, A. OU Analyse: Analysing at-risk students at The Open University. Learn. Anal. Rev. 2015, LAK15-1, 1–16.

22. Al-Azawei, A.; Al-Masoudy, M. Predicting Learners’ Performance in Virtual Learning Environment (VLE) based on Demographic, Behavioral and Engagement Antecedents. Int. J. Emerg. Technol. Learn. 2020, 15, 60–75.

23. Chitti, M.; Chitti, P.; Jayabalan, M. Need for Interpretable Student Performance Prediction. In Proceedings of the 2020 13th International Conference on Developments in eSystems Engineering (DeSE), Liverpool, UK, 14–17 December 2020; pp. 269–272.

24. Alonso, J.M.; Casalino, G. Explainable Artificial Intelligence for Human-Centric Data Analysis in Virtual Learning Environments. In Higher Education Learning Methodologies and Technologies

Online; Burgos, D., Cimitile, M., Ducange, P., Pecori, R., Picerno, P., Raviolo, P., Stracke, C.M., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 125–138.

25. Conijn, R.; Kahr, P.; Snijders, C. The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. J. Learn. Anal. 2023, 10, 37–53.
26. Shulner-Tal, A.; Kuflik, T.; Kliger, D. Fairness, Explainability and in-between: Understanding the Impact of Different Explanation Methods on Non-Expert Users’ Perceptions of Fairness toward an Algorithmic System. Ethics Inf. Technol. 2022, 24, 2.
27. Lee, J.D.; See, K.A. Trust in Automation: Designing for Appropriate Reliance. Hum. Factors 2004, 46, 50–80.
28. Vereschak, O.; Bailly, G.; Caramiaux, B. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. Proc. ACM Hum.-Comput. Interact. 2021, 5, 1–39.
29. Jian, J.Y.; Bisantz, A.M.; Drury, C.G. Foundations for an Empirically Determined Scale of Trust in Automated Systems. Int. J. Cogn. Ergon. 2000, 4, 53–71.
30. Chien, S.Y.; Lewis, M.; Sycara, K.; Liu, J.S.; Kumru, A. The Effect of Culture on Trust in Automation: Reliability and Workload. Acm Trans. Interact. Intell. Syst. 2018, 8, 1–31.
31. Merritt, S.M. Affective Processes in Human–Automation Interactions. Hum. Factors 2011, 53, 356–370.
32. Muir, B. Operators’ Trust in and Use of Automatic Controllers in a Supervisory Process Control Task. Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 1989.
33. Benbasat, I.; Wang, W. Trust in and adoption of online recommendation agents. J. Assoc. Inf. Syst. 2005, 6, 4.
34. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects. arXiv 2018, arXiv:1812.04608.
35. Schraagen, J.M.; Elsasser, P.; Fricke, H.; Hof, M.; Ragalmuto, F. Trusting the X in XAI: Effects of different types of explanations by a self-driving car on trust, explanation satisfaction and mental models. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 2020, 64, 339–343.
36. Ribeiro, M.T.; Singh, S.; Guestrin, C. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. arXiv 2016, arXiv:1602.04938.
37. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; pp. 4765–4774.

38. Belle, V.; Papantonis, I. Principles and practice of explainable machine learning. Front. Big Data 2021, 4, 39.
39. Aechtner, J.; Cabrera, L.; Katwal, D.; Onghena, P.; Valenzuela, D.P.; Wilbik, A. Comparing User Perception of Explanations Developed with XAI Methods. In Proceedings of the 2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Padua, Italy, 18–23 July 2022; pp. 1–7.
40. Aykin, N.M.; Aykin, T. Individual differences in human-computer interaction. Comput. Ind. Eng. 1991, 20, 373–379.
41. ISO 9241-1:1997; Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs). International Organization for Standardization: Geneva, Switzerland, 1997.
42. European Commission, Directorate-General for Education, Youth, Sport and Culture. ECTS Users' Guide 2015; Publications Office of the European Union: Luxembourg, 2017.

Retrieved from https://encyclopedia.pub/entry/history/show/103426