Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240477

Demystifying XAI: Requirements for Understandable XAI Explanations

Jan STODT[a,1], Christoph REICH[a] and Martin KNAHL[a]
[a] *Institute for Data Science, Cloud Computing, and IT Security, Furtwangen University, Furtwangen, Germany*

Abstract. This paper establishes requirements for assessing the usability of Explainable Artificial Intelligence (XAI) methods, focusing on non-AI experts like healthcare professionals. Through a synthesis of literature and empirical findings, it emphasizes achieving optimal cognitive load, task performance, and task time in XAI explanations. Key components include tailoring explanations to user expertise, integrating domain knowledge, and using non-propositional representations for comprehension. The paper highlights the critical role of relevance, accuracy, and truthfulness in fostering user trust. Practical guidelines are provided for designing transparent and user-friendly XAI explanations, especially in high-stakes contexts like healthcare. Overall, the paper's primary contribution lies in delineating clear requirements for effective XAI explanations, facilitating human-AI collaboration across diverse domains.

Keywords. Understandability Requirements, XAI, Explanations, Non-AI Experts

1. Introduction

In the realm of Artificial Intelligence (AI), explanations of decisions are often tailored by and for AI experts, a phenomenon coined by Miller [1] as "inmates running the asylum". However, in critical domains like healthcare, understanding AI decisions is crucial for informed decision-making and patient safety. Healthcare professionals rely on AI systems for diagnosis and treatment planning, but the complexity of AI algorithms can hinder comprehension for those without AI expertise. This conceptual paper addresses the need for clear and interpretable AI explanations, particularly in healthcare, to improve decision-making and patient outcomes.

2. Methods

The literature review methodology follows the Systematic Literature Review method, following the guidelines of the PRISMA framework. In the following, the 6 steps performed are presented.

1. ***Define Literature Research Questions [RQ] and Research Objectives [RO]:***
   RQ: What criteria contribute to the understandability of explanations?
   RO: The literature review aims to contribute to the advancement of knowledge in the field of understandable XAI.
2. ***Inclusion and Exclusion Criteria:***

[1] Corresponding Author: Jan Stodt; E-mail: jan.stodt@hs-furtwangen.de.

Inclusion: Domain of Explainable AI, General definition of explanation
Exclusion: Other Domains

3. ***Selection of Databases:***
Google Scholar is chosen as the primary search engine due to its extensive indexing of high-quality journals (such as IEEE, ACM, Elsevier, Springer, etc.) and its inclusion of the preprint service arXiv, which hosts many papers that are later published in reputable journals.

4. ***Definition of Search Components:***

Keyword Component 1	Keyword Component 2	Keyword Component 3
Understandable Comprehensible Pragmatic	Understandability Comprehensibility Pragmatism	Explainable AI XAI

5. ***Development of Search Strings:***

Search Strings
(Understandable or Understandability) or (Comprehensible or Comprehensibility) or (Pragmatic or Pragmatism) and (Explainable AI or XAI)
(Understandable or Understandability) or (Comprehensible or Comprehensibility) or (Explainable AI or XAI)

6. ***Conducting the Research:***
The literature resulted initially in 27 papers, of which 8 papers are not related, and 5 which were not available in full text. 14 papers are included in the final review.

3. Results: Requirements for Understandable XAI Explanations

Local Explanation: Local explanations, such as "why" and "why-not" explanations, are especially effective for non-AI experts. Herm [2] found them beneficial for audiences like prospective physicians, as they focus on specific decisions to improve comprehension. Mohseni et al. [3] also emphasize the preference for why explanations, aligning with users unfamiliar with AI concepts. Miller et al. [4] liken local explanations to everyday explanations, as they offer insight into specific events.
Avoidance of Excessive Detail: Miller [1] emphasizes the importance of avoiding excessive detail, advocating for simplicity and coherence in explanations. Users generally prefer concise, general explanations aligned with their prior knowledge. Research by Lombrozo [5] and Read et al. [6] indicates that too much detail can overwhelm users, impeding understanding and acceptance. Miller [1] compares explanations to conversations, recommending they be accurate, concise, and tailored to the user's expertise. Duán [7] and Kim et al. [8] highlight that users generally favor simple and clear explanations, which enhance comprehension, usability, and trust in AI systems, fostering better human-AI collaboration.
Accuracy and Truthfulness: Rong et al. [9]. emphasize that accuracy and truthfulness are crucial in XAI explanations, enhancing user comprehension and trust in AI systems. Accurate explanations provide reliable insights, fostering effective understanding and communication, while misleading explanations can cause misunderstandings. Additionally, accurate and truthful explanations have educational value, enabling users to better understand AI systems, fostering engagement, and supporting informed decision-making. Trust in AI is essential for positive user experiences and widespread adoption of AI solutions.

Relevance: Relevance in XAI explanations, as highlighted by Miller [1], Grice's maxims [10], and psychology research *[11], [12], [13]*, is crucial. Explanations should exclude irrelevant details for clarity and understanding. Miller [1] emphasizes the importance of relevance to the question and the listener's mental model. Duràn [7] stresses the need for relevance to meet the specific needs of recipients, improving engagement and applicability. Kim et al. [8] assert that relevance is essential for users to find information useful, while Nyrup and Robinson [14] highlight its role in enhancing inferential abilities to improve understanding in XAI explanations.

Tailoring to User Expertise: Tailoring explanations to user expertise, as discussed by Kim et al. [15] Miller [1], Langer et al. [16], and Nyrup and Robinson [14], is vital. Different expertise levels require different approaches, with simplified explanations helping those with less knowledge, and detailed ones benefiting experts. Mohseni et al. [3] emphasize that tailoring improves accessibility and comprehension. Durán [7] highlights its importance in preventing confusion, and Cabitza et al. [17] suggest assessing user comprehension across expertise levels. Chromik and Schuessler [18] and Fok and Weld [19] underscore its influence on study design and verification of AI recommendations.

Inclusion of Domain Knowledge: Panigutti et al. [20] stress that incorporating domain knowledge is crucial. Contextual relevance clarifies why black-box models made decisions, enhancing user understanding and trust. Domain knowledge simplifies complex concepts for non-experts, improving trust and acceptance. It ensures explanations accurately reflect the model's decision-making process, aligning them with human expertise for better intuitiveness. Adding domain knowledge improves decision support, particularly in complex fields like healthcare.

Focus on Understanding: Wang et al. [21] and Arrieta et al. [22] emphasize that prioritizing comprehension in XAI explanations is vital. Transparent explanations promote trust by allowing users to verify AI credibility and fairness. Understanding AI reasoning enhances user agency and involvement, helping with error detection and correction. Clear explanations also serve as educational tools for complex decision-making. In healthcare, understanding AI recommendations fosters trust and patient involvement in care decisions, as LaRosa and Danks [23] highlight, which is crucial for effective collaboration.

Non-Propositional Representation: Páez [24] and Mittelstadt et al. [25] advocate for using visual aids like diagrams, graphs, and maps in explanations, as they are more intuitive than text, reducing cognitive load and improving comprehension. Kim et al. [8] highlight their effectiveness in simplifying complex information and aiding communication. Visualizations increase user engagement and interest, and interactive features offer a personalized experience. They also transcend language barriers and cater to diverse backgrounds, making them widely accessible. Páez [24] emphasizes that this pragmatic approach enhances the usability and accessibility of XAI explanations.

Low Cognitive Load, High Task Performance, and Low Task Time: Hoffman et al. [26] and Herm [2] emphasize the importance of low cognitive load, high task performance, and short task times. Herm's study with 271 prospective physicians showed that XAI explanations significantly influence cognitive load and task performance. Balancing cognitive load is crucial for accurately representing decisions and facilitating understanding. Guidelines are needed in high-stakes situations to avoid over-reliance on certain types of explanations, ensuring that XAI effectively supports user interaction with AI systems.

Table 1. Requirements for Understandable Explanations

Requirements	
R1 – Local Explanation	Focus on individual decisions.
R2 – Avoidance of Excessive Detail	Strike a balance to avoid overwhelming users.
R3 – Accuracy and Truthfulness	Ensure explanations are accurate and truthful.
R4 – Relevance	Provide only pertinent details.
R5 – Tailoring to User Expertise	Tailor explanations to user's expertise level.
R6 – Inclusion of Domain Knowledge	Incorporate relevant domain knowledge.
R7 – Focus on Understanding	Aim to foster understanding.
R8 – Non-Propositional Representation	Use visual aids for enhanced comprehension.
R9 – Low Cognitive Load, High Task Performance, and Low Task Time	Minimize cognitive load and maximize task performance.

4. Discussion

The requirements for effective Explainable Artificial Intelligence (XAI) emphasize the importance of user-centric explanations in high-stakes environments like healthcare, where incorporating domain knowledge and non-propositional representations enhances comprehension and aligns explanations with user expertise. These principles are crucial for improving decision-making, enhancing user trust, and fostering collaboration between humans and AI systems. However, balancing clarity and detail remains challenging.

5. Conclusions

In conclusion, this paper emphasizes the critical need for clear and interpretable explanations in Artificial Intelligence (AI), especially in domains like healthcare. By establishing essential requirements for Explainable AI (XAI) methods, we provide a roadmap for developing transparent and user-friendly AI explanations. Moving forward, prioritizing user-centric design principles is crucial to foster trust and collaboration between AI systems and users. Embracing transparency in AI explanations will ultimately enhance decision-making and improve outcomes across various domains.

Acknowledgement

We would like to express our gratitude to the funding organization German Federal Ministry of Education and Research with their funding program "Forschung an Fachhochschulen" contract number 13FH5I09IA for their support in accordance with the regulations associated with this project.

References

[1] Miller T, Howe P, Sonenberg L. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences [Internet]. arXiv. 2017 Dec 04 [cited 2024 Feb 05]. Available from: http://arxiv.org/abs/1712.00547

[2] Herm LV. Impact Of Explainable AI On Cognitive Load: Insights From An Empirical Study [Internet]. arXiv. 2023 Apr 18 [cited 2024 Feb 04]. Available from: http://arxiv.org/abs/2304.08861

[3] Mohseni S, Zarei N, Ragan ED. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. ACM Trans Interact Intell Syst. 2021 Dec;11(3-4):1-45. doi: 10.1145/3387166.
[4] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell. 2019 Feb;267:1-38. doi: 10.1016/j.artint.2018.07.007.
[5] Lombrozo T. Simplicity and probability in causal explanation. Cognit Psychol. 2007 Nov;55(3):232-257. doi: 10.1016/j.cogpsych.2006.09.006.
[6] Read SJ, Marcus-Newhall A. Explanatory coherence in social explanations: A parallel distributed processing account. J Pers Soc Psychol. 1993;65(3):429. doi: 10.1037/0022-3514.65.3.429.
[7] Durán JM. Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. Artif Intell. 2021 Aug;297:103498. doi: 10.1016/j.artint.2021.103498.
[8] Kim SSY, Watkins EA, Russakovsky O, Fong R, Monroy-Hernández A. Help Me Help the AI: Understanding How Explainability Can Support Human-AI Interaction. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, in CHI '23. New York, NY, USA: Association for Computing Machinery; 2023 Apr. pp. 1-17. doi: 10.1145/3544548.3581001.
[9] Rong Y, et al. Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations. IEEE Trans Pattern Anal Mach Intell. 2024:1-20. doi: 10.1109/TPAMI.2023.3331846.
[10] Grice HP. Logic and conversation. In: Speech Acts. Brill; 1975. pp. 41-58.
[11] Slugoski BR, Lalljee M, Lamb R, Ginsburg GP. Attribution in conversational context: Effect of mutual knowledge on explanation-giving. Eur J Soc Psychol. 1993;23(3):219-238. doi: 10.1002/ejsp.2420230302.
[12] Jaspars JM, Hilton DJ. Mental models of causal reasoning. 1988.
[13] Byrne RM. The construction of explanations. In: AI and Cognitive Science'90: University of Ulster at Jordanstown 20–21 September 1990. Springer; 1991. pp. 337-351.
[14] Nyrup R, Robinson D. Explanatory pragmatism: a context-sensitive framework for explainable medical AI. Ethics Inf Technol. 2022 Mar;24(1):13. doi: 10.1007/s10676-022-09632-3.
[15] Kim S, et al. Designing an XAI interface for BCI experts: A contextual design for pragmatic explanation interface based on domain knowledge in a specific context. Int J Hum Comput Stud. 2023 Jun;174:103009. doi: 10.1016/j.ijhcs.2023.103009.
[16] Langer M, et al. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artif Intell. 2021 Jul;296:103473. doi: 10.1016/j.artint.2021.103473.
[17] Cabitza F, et al. Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI. Expert Syst Appl. 2023 Mar;213:118888. doi: 10.1016/j.eswa.2022.118888.
[18] Chromik M, Schuessler M. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. 2020.
[19] Fok R, Weld DS. In Search of Verifiability: Explanations Rarely Enable Complementary Performance in AI-Advised Decision Making. arXiv. 2024 Feb 01. Accessed: 2024 Feb 12. Available from: http://arxiv.org/abs/2305.07722.
[20] Panigutti C, Perotti A, Pedreschi D. Doctor XAI: an ontology-based approach to black-box sequential data classification explanations. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, in FAT* '20. New York, NY, USA: Association for Computing Machinery; 2020 Jan. pp. 629-639. doi: 10.1145/3351095.3372855.
[21] Wang K, Oramas J, Tuytelaars T. Towards Human-Understandable Visual Explanations High-frequency Cues Can Better Be Removed. ArXiv210407954 Cs. 2021 Apr. Accessed: 2022 Feb 03. Available from: http://arxiv.org/abs/2104.07954.
[22] Barredo Arrieta A, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020 Jun;58:82-115. doi: 10.1016/j.inffus.2019.12.012.
[23] LaRosa E, Danks D. Impacts on Trust of Healthcare AI. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. New Orleans, LA, USA: ACM; 2018 Dec. pp. 210-215. doi: 10.1145/3278721.3278771.
[24] Páez A. The Pragmatic Turn in Explainable Artificial Intelligence (XAI). Minds Mach. 2019 Sep;29(3):441-459. doi: 10.1007/s11023-019-09502-w.
[25] Mittelstadt B, Russell C, Wachter S. Explaining Explanations in AI. In: Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT 19. 2019. pp. 279-288. doi: 10.1145/3287560.3287574.
[26] Hoffman RR, Mueller ST, Klein G, Litman J. Metrics for Explainable AI: Challenges and Prospects.