Explainable AI in healthcare: Fundamentals challenge

While research of explainable artificial intelligence (XAI) has increased significantly since the European Union (EU) General Data Protection Regulation (GDPR) went into effect in 2018 and demand for XAI in medicine is high, defining what constitutes an acceptable explanation remains a challenge. Therefore, we have commenced a study with the intent to synthesise for the first time, published data and expert opinions from a diverse research background regarding what it means for AI to be *explainable*. We welcome expressions of interest from experts who wish to contribute.

The International Data Corporation (IDC) predicts AI spending will exceed $300bn by 2026 as more companies integrate AI technologies into their product and service offerings.[1] When they ratified General Data Protection Regulation (GDPR) in 2016, the European Union (EU) granted European citizens a *right to explanation* if they are affected by algorithmic decision.[2] For example, if an AI system rejects an individual's application for a loan, the applicant is entitled to request the justifications that led to the decision so that they may ensure consistency with other laws, regulations and rights. In 2019 the EU published the *Ethics Guidelines for Trustworthy AI* which includes a general framework where explainability is an integral component.[3] These guidelines formed the basis for several sections of the *Artificial Intelligence Act* proposed by the European Commission (EC) in April 2021. Finally, the 2023 UK white paper '*A pro-innovation approach to AI regulation*' identifies *appropriate transparency and explainability* as one of the key principles for development of responsible AI. The need for transparent AI led to a significant increase in the size of the XAI community over the last few years.[4]

In healthcare, XAI is essential for many purposes including medical education, research and clinical decision making.[5,6] If medical professionals are complemented by sophisticated AI technologies and in some cases even overruled, the human experts must, on demand, still have a chance to understand and to retrace the machine decision process.[7] A key requirement for adoption of these technologies is that users must feel confident in their recommendations.[8] For high stakes decisions, like those faced in healthcare, XAI becomes more urgent.[7] Without a solution to the problem of AI trustworthiness and user acceptance of healthcare technologies generally, the undeniable benefits of these systems will never be realised and all our efforts to develop accurate health-AI will be in vain.

There is a growing demand for medical AI technologies that can not only perform well, but that are also trustworthy, transparent, interpretable, and explainable.[9] Medical AI are considered to be a high-risk AI application in the proposal by EC legislation, which strengthens the urgency for XAI. Yet, and despite significant research interest on XAI, papers on foundational aspects of the structure and composition of such explanation are insufficient. XAI lack a formal and universally agreed definition for what constitutes an explanation. Papers on the attributes for a good explanation in healthcare are also scarce. There remains no clear consensus on the necessary form a good explanation should take.[10]

We believe that investigating and developing the definition and attributes of XAI in healthcare is important and will significantly benefit the XAI research community and ultimately AI users, including healthcare professionals and patients. More specifically, by understanding the desiderata for a good explanation, new algorithms for developing improved and more holistic explanations can be developed. However, generating explanations is not enough as it is also crucial to evaluate the quality of those explanations. Thus, the proposed list of attributes will also serve as the basis to develop a formalised evaluation process, which is currently lacking. In addition, having better explanations that suit users’ needs will bring XAI close to adoption in key domains like healthcare. Clinicians will benefit directly from this as they will be able to understand how and why AI technologies made a particular recommendation, which increases the ability for healthcare professionals to better understand the day-to-day patterns and needs of their patients and provide more personalised care and support. Adoption of AI technologies can also reduce cost and waste and resolve issues of resource contention for consultations, surgeries, and medicines. Finally, the most important impact will be on patients’ lives as augmented clinical decision making – when using the explainable health-AI, can significantly improve care.

For these reasons we have commenced a Delphi study to inform our understanding of what it means for AI to be explainable in the context of healthcare. The study will support development of a broad and nuanced definition and global list of characteristics for construction of a context-appropriate and robust explanation for health-AI. We welcome expressions of interest and participation from experts from a diverse research background who wish to contribute. In this study, you will be asked to score and comment on explainable AI definitions and attributes identified from an extensive literature review. The questionnaire contains 18 questions, many of which will take less than a minute to answer. Please email the lead author to express your interest to participate.

Steering Group: Evangelia Kyrimi[1] 🖂, David Lagnado[2], Zane Perkins[3], William Marsh[1], Scott McLachlan[4], Kudakwashe Dube[5]

[1] School of Electronic Engineering and Computer Science, Queen Mary University of London
[2] Experimental Psychology, University College London
[3] Centre for Trauma Sciences, Blizard Institute, Queen Mary University of London
[4] Division of Applied Technologies for Clinical Care with the Faculty of Nursing, Midwifery & Palliative Care, King’s College London
[5] Research Fellow in Digital Technologies for Health, King's College London

🖂e-mail: e.kyrimi@qmul.ac.uk

References

1. Mc Gowran, L. *Silicon Republic* (2023).

2. Goodman, B. & Flaxman, S. *AI Mag.* **38**, 50–57 (2017).

3. The High-Level Expert Group on Artificial Intelligence. Set up by the European Commission (2019).

4. Schwalbe, G. & Finzel, B. *Data Min. Knowl. Discov.* (2023).

5. Chaddad, A., Peng, J., Xu, J. & Bouridane, A. *Sensors* **23**, 1–19 (2023).

6. Ridley, M. *Inf. Technol. Libr.* **41**, (2022).

7. Sheu, R. K. & Pardeshi, M. S. *Sensors* **22**, (2022).

8. Kyrimi, E. *et al. Artif. Intell. Med.* **116**, 102079 (2021).

9. Holzinger, A. *et al.* WIREs: Data Mining and Knowledge Discovery **9**, 1–13 (2019).

10. Confalonieri, R., Coba, L., Wagner, B. & Besold, T. R. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **11**, 1–21 (2021).

Acknowledgement

This research was supported by the Royal Academy of Engineering under the Research Fellowship programme.

Competing interests

The authors declare no competing interests.