Explainability and artificial intelligence in medicine

In recent years, improved artificial intelligence (AI) algorithms and access to training data have led to the possibility of AI augmenting or replacing some of the current functions of physicians.[1] However, interest from various stakeholders in the use of AI in medicine has not translated to widespread adoption.[2] As many experts have stated, one of the key reasons for this restricted uptake is the scarce transparency associated with specific AI algorithms, especially black-box algorithms.[3] Clinical medicine, primarily evidence-based medical practice, relies on transparency in decision making.[3-5] If there is no medically explainable AI and the physician cannot reasonably explain the decision-making process, the patient's trust in them will erode. To address the transparency issue with certain AI models, explainable AI has emerged.[3]

In *The Lancet Digital Health*, Marzyeh Ghassemi and colleagues[6] have argued that explainable AI applications that are currently available are imperfect and provide only a partial explanation of the inner workings of AI algorithms. They have called for stakeholders to move away from insisting on explainability and to seek other measures, like validation, to enable trust and confidence in black-box models. There is some validity in their criticism of certain explainable frameworks, like post-hoc explainers. These explainers mostly approximate the underlying machine learning mechanisms to explain the decision making. However, based on the limitations of certain explainable AI methods, the argument to restrict explainable AI and prioritise other validation approaches, like randomised controlled trials, is specious.

Models or systems whose decisions cannot be well interpreted can be hard to accept[7], especially in fields like medicine.[4] Reliance on the logic of black-box models violates medical ethics. Black-box medical practice hinders clinicians from assessing the quality of model inputs and parameters. If clinicians cannot understand the decision making, they might be violating patients' rights to informed consent and autonomy.[4,5] When clinicians cannot decipher how the results were arrived at, it is unlikely that they will be able to communicate and disclose with the patient appropriately, thus affecting the patient's autonomy and ability to engage in informed consent. Increasingly, there have been examples of high performing black-box models that have been caught using wrong or confounding variables to achieve their results. For example, patients with asthma were found by a deep learning model to be at low risk of death by pneumonia because the model learnt from a training dataset that included a group of patients with asthma who had active intervention from clinicians.[8] In another example, a deep learning model developed to screen x-rays for pneumonia used confounding information like the scanner's location to detect pneumonia.[8] In a third example, a deep learning model developed to distinguish high-risk patients from lower-risk patients, based on x-rays, used hardware-related metadata to predict the risk.[3] These cases suggest that reliance on the accuracy of the models is insufficient. Additional trust enhancing frameworks, like explainable AI, are required.

Although criticism of explainable AI methods has been growing in recent years, there seems to be astonishingly little scrutiny of what led to the need for explainable AI: deep learning models. Such models have no explicit declarative knowledge representation, which poses a challenge in deriving an explanatory narrative. Many high performing deep learning models have millions or even billions of parameters that are only identifiable by their location in a complex network, not as human interpretable labels, leading to the black-box situation.[9] Also, many deep learning models that do well on training datasets do not do well on independent datasets. Further, deep learning algorithms require a large amount of data to be trained for both interpolation and extrapolation. These issues with deep learning models have yet to be meaningfully resolved and persist in various applications, including in medicine.

Critics of explainable AI have argued for the prioritisation of validity measures over explainability frameworks.[6,8] The rationale is that, currently, many

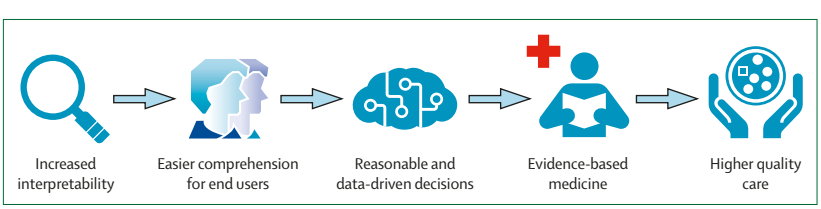

Figure: **How does explainable artificial intelligence drive better medical care?**

drugs and medical devices adopt validation processes (such as randomised controlled trials) to indicate efficacy, and so should AI-enabled medical devices or software. However, we believe this argument is inappropriate. Generally, the performance of AI systems is assessed on prediction accuracy measures.[10] Even with the best efforts, AI systems are unlikely to achieve perfect accuracy due to different sources of errors.[3] If perfect accuracy was achieved theoretically, there is no guarantee that the AI system is still free of biases—especially when the systems have been trained with heterogeneous and complex data, as occurs in medicine.

Ignoring or restricting explainable AI is detrimental to the adoption of AI in medicine as few alternatives exist that can comprehensively respond to accountability, trust, and regulatory concerns while engendering confidence and transparency in the AI technology. The use of explainable frameworks could help to align model performance with clinical guidelines objectives.[3] Therefore, enabling better adoption of AI models in clinical practice. Transparent algorithms or explanatory approaches can also make the adoption of AI systems less risky for clinical practitioners.[2,3] There are already an increasing number of examples of how explainable frameworks in various medical specialities enhance transparency and insight.[8] These case studies can guide the integration of explainable AI with AI medical systems. Through this integration, a second level of explainability and multiple benefits can be achieved, including higher interpretability, better comprehension for clinicians leading to evidence-based practice, and improved clinical outcomes (figure).

I hold directorship and shares in Medical Artificial Intelligence Pty Ltd.

Copyright © 2022 The Author(s). Published by Elsevier Ltd. This is an Open Access article under the CC BY-NC-ND 4.0 license.

*Sandeep Reddy
sandeep.reddy@deakin.edu.au

School of Medicine, Deakin University, Geelong, VIC 3216, Australia

1 Reddy S, Fox J, Purohit MP. Artificial intelligence-enabled healthcare delivery. *J R Soc Med* 2019; **112:** 22–28.
2 Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019; **17:** 195.
3 Amann J, Blasimme A, Vayena E, Frey D, Madai VI, Precise Q. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inform Decis Mak* 2020; **20:** 310.
4 Kundu S. AI in medicine must be explainable. *Nat Med* 2021; **27:** 1328.
5 Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *J Med Ethics* 2021; medethics-2020-107102.
6 Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3:** e745–50.
7 Vellido A. Societal issues concerning the application of artificial intelligence in medicine. *Kidney Dis* 2019; **5:** 11–17.
8 Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare-perspectives on trustworthiness, explainability, usability, and transparency. *NPJ Digit Med* 2020; **3:** 47.
9 Marcus G. Deep learning: a critical appraisal. *arXiv* 2018; published online January 2. https://arxiv.org/abs/1801.00631 (preprint).
10 Desai AN. Artificial intelligence: promise, pitfalls, and perspective. *JAMA* 2020; **323:** 2448–49.