Artificial Intelligence Algorithms Need to Be Explainable—or Do They?

Tyler J. Bradshaw[1], Melissa D. McCradden[2], Abhinav K. Jha[3], Joyita Dutta[4], Babak Saboury[5], Eliot L. Siegel[6], and Arman Rahmim[7]

[1]*University of Wisconsin–Madison, Madison, Wisconsin;* [2]*Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada;* [3]*Washington University in St. Louis, St. Louis, Missouri;* [4]*University of Massachusetts Amherst, Amherst, Massachusetts;* [5]*National Institutes of Health, Bethesda, Maryland;* [6]*University of Maryland School of Medicine, Baltimore, Maryland; and* [7]*University of British Columbia, Vancouver, British Columbia, Canada*

With the growing role of artificial intelligence (AI) in radiology, there is concern over the black-box nature of modern AI algorithms. Users of AI often have no way of knowing how or why an algorithm arrived at a prediction, which makes it difficult for a user to appraise or critique the quality of the prediction. The group of methods collectively known as explainable AI (XAI) aims to overcome this limitation by providing human-understandable explanations of the causal relationships between an algorithm's inputs and outputs. XAI's motivations include promoting trust between clinicians and AI systems, enabling detection of errors, and facilitating informed consent. However, it has been argued that XAI may not in fact address the needs of clinicians and may also introduce unintended consequences, potentially compromising the purported value of XAI (*1*). At the 2022 Society of Nuclear Medicine and Molecular Imaging annual meeting, we held a debate over the clinical need for XAI. We summarize that debate here by discussing 5 key arguments. For each argument, we present the case for and against the use of XAI from the perspectives of data science, clinical practice, and bioethics.

First, it should be recognized that the term *XAI* refers to a variety of approaches, most of which were originally developed for uses outside of medicine (*2*). XAI includes interpretable methods in which AI algorithms are designed to be inherently explainable, as well as post hoc methods that are applied to already-trained algorithms. The way that explanations are presented can also vary (*3*). In radiology, XAI is often presented through saliency maps, which highlight the parts of an image that have the most impact on the model's predictions. For example, Miller et al. showed how saliency mapping can highlight the most influential regions of the myocardium for AI-based diagnosis of coronary artery disease in SPECT images (*4*).

IDENTIFICATION OF CONFOUNDING FACTORS

A potential benefit of XAI is that it might help uncover AI biases caused by confounding factors. The tendency for AI to rely on shortcuts—spurious correlations unrelated to biomedical pathology—is well documented. DeGrave et al. demonstrated how an AI system learned to mistakenly rely on laterality markings in radiographs for diagnostic predictions (*5*). XAI could be used as a quality control method, potentially helping users identify these confounding factors. Conversely, correlations uncovered by XAI might turn out to be real but previously unknown biomedical relationships, in which case XAI could be used as a tool for scientific discovery. The counterargument is that the task of identifying confounding factors should not be the responsibility of the users. The workflow of a user, which consists of looking at individual AI predictions during clinical reads, is not well matched to the workflow required to identify confounding factors, which requires inspecting XAI explanations across many samples to discern spurious relationships. In this context, XAI might be better suited as a quality control tool for developers rather than for users. Additionally, biases caused by confounding factors will be better uncovered by comprehensive clinical evaluation of AI algorithms, including learning which patient populations would benefit from the use of an algorithm.

DETECTION OF ALGORITHMIC ERRORS

Another argument in favor of XAI is that it can help users know when an AI algorithm errs. XAI can help users know when to follow and when to reject AI predictions based on the plausibility of the explanation. Also, the additional information provided by XAI explanations could lead to better failure-mode profiling of a system. This could result in a better understanding of the functional dependencies of a system and its vulnerabilities. On the other hand, recent evidence suggests that XAI can actually have the opposite effect and could potentially lead some users to make more judgment errors (*1*). The presence of explanations alongside predictions may further persuade users to follow incorrect outputs by appearing to corroborate erroneous AI predictions (*6*).

RELIABILITY OF EXPLANATIONS

Although XAI is in its early stages, it has already provided meaningful contributions to guiding AI development (*5*). As the field matures, the quality of explanations will continue to improve. Future methods will include more sophisticated explanations beyond just saliency mapping. However, a major criticism of current XAI approaches is that their explanations are too unreliable to be

Received Sep. 23, 2022; revision accepted Mar. 17, 2023.
For correspondence or reprints, contact Tyler Bradshaw (tbradshaw@wisc.edu).
Published online Apr. 28, 2023.

DOI: 10.2967/jnumed.122.264949

clinically beneficial. Studies have found inconsistencies in the explanations provided by different XAI techniques and have demonstrated their sensitivity to clinically inconsequential changes in input images (*7*). With different XAI methods providing different explanations, how can users know which one is correct? And if both the predictions and the explanations can be wrong, this adds another avenue for an AI system to err. A more fundamental challenge is that explanations from current XAI methods superficially represent the computational complexity that underlies a prediction (*3*). Additionally, the best approach for evaluating the quality of an explanation is uncertain. Different approaches have been used to evaluate XAI, including user feedback studies, simulation studies, and studies measuring XAI impact on diagnostic accuracy, but each has potential shortcomings, such as subjectivity, lack of reality, and high costs. Although the current challenges facing XAI should not dissuade us from pursuing explainability as a goal, there is no guarantee that XAI will become sufficiently reliable, especially as AI complexity continues to increase.

TRANSPARENCY AND TRUSTWORTHINESS

Another potential benefit of XAI is that it could enhance the transparency and trustworthiness of AI systems. XAI provides users with a better understanding of an AI system's reasoning, which can lead to trust between user and algorithm. Furthermore, clinicians generally prefer AI systems with XAI over systems without it (*8*), and providing them with XAI could lead to broader adoption of beneficial AI tools. A rebuttal to this argument is that a mechanistic understanding of how an intervention works is not necessary for either trust or transparency. Many drugs have unknown mechanisms of action, yet we learn the conditions under which they should and should not be used. Furthermore, given the questionable reliability of XAI, the transparency offered by XAI may not be the kind of transparency that is valuable and could be worse than no information at all. As long as developers provide detailed information on the development and validation of their AI system, including its expected benefits and risks, the additional transparency provided by XAI may not be necessary.

PATIENT-CENTERED CARE

Lastly, it can be argued that XAI empowers clinicians to provide more patient-centered care. Without explainability, clinicians lack power to adequately critique an AI decision on behalf of their patient. XAI gives an opportunity for providers and patients to understand AI decisions, giving both of them greater control over clinical decisions. Additionally, clinicians are accountable for obtaining informed consent from patients, which may not be well served by black-box algorithms. The counterargument is that XAI may lead to less emphasis on patient input and testimony in decision-making, particularly given the possibility of overreliance on AI systems (*9*). Currently, neither AI nor XAI considers patient preferences or values (*10*), and XAI may shift clinicians further into a decision-making mode that assumes that the AI system holds all the knowledge to guide clinical decisions. Additionally, informed consent has never required a mechanistic understanding of an intervention, only its risks and benefits.

CONCLUSION

XAI may play an essential role in the era of collaborative AI–human intelligence systems within medicine. But the potential benefits of XAI need to be carefully weighed against potential risks. A cautious approach to the clinical adoption of XAI is warranted. Future directions for research should include improved robustness of XAI with more standardized methods of objectively measuring explanation quality. Furthermore, an understanding of the multifaceted impact that XAI will have on clinical decision-making is needed, which will require a concerted multidisciplinary effort.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Jacobs M, Pradier MF, McCoy TH, Perlis RH, Doshi-Velez F, Gajos KZ. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Transl Psychiatry.* 2021;11:108.
2. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: a review of machine learning interpretability methods. *Entropy (Basel).* 2021;23:18.
3. Evans T, Retzlaff CO, Geißler C, et al. The explainability paradox: challenges for xAI in digital pathology. *Future Gener Comput Syst.* 2022;133:281–296.
4. Miller RJH, Kuronuma K, Singh A, et al. Explainable deep learning improves physician interpretation of myocardial perfusion imaging. *J Nucl Med.* 2022;63:1768–1774.
5. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat Mach Intell.* 2021;3:610–619.
6. Gaube S, Suresh H, Raue M, et al. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit Med.* 2021;4:31.
7. Arun N, Gaw N, Singh P, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol Artif Intell.* 2021;3:e200267.
8. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J Am Med Inform Assoc.* 2020;27:592–600.
9. McCradden M, Hui K, Buchman DZ. Evidence, ethics and the promise of artificial intelligence in psychiatry. *J Med Ethics.* December 29, 2022 [Epub ahead of print].
10. Birch J, Creel KA, Jha AK, Plutynski A. Clinical decisions using AI must consider patient values. *Nat Med.* 2022;28:229–232.