Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI?

Aimilia Gastouniotti, PhD • Despina Kontos, PhD

Aimilia Gastouniotti, PhD, is a research associate in the department of radiology at the University of Pennsylvania. Her research focuses on artificial intelligence with an application focus on cancer imaging phenotypes related to risk prediction. She has co-authored 22 journal articles, three book chapters, 25 conference proceedings papers, as well as 22 abstracts in premier scientific meetings. Dr Gastouniotti is an associate member of the American Association for Cancer Research (AACR).

Despina Kontos, PhD, is an associate professor of radiology at the University of Pennsylvania. Dr Kontos holds a PhD in computer and information science. Her research focuses on investigating imaging as a biomarker for precision-medicine decisions in cancer screening, prognosis, and treatment. She has authored more than 60 peer-reviewed publications and is leading multiple NIH studies using machine learning to integrate imaging and genomic markers for augmenting precision cancer care.

Within a short time, artificial intelligence (AI) has taken center stage in radiology, with a large number of publications that is growing at a swift pace. Radiologists routinely assess medical images and report findings to detect, characterize, and monitor diseases. Such assessment is based both on education and experience and can be, at times, subjective. To augment such qualitative interpretation, AI excels at recognizing complex patterns in imaging data and can offer novel tools for quantitative medical image interpretation in an automated fashion. Deep learning, in particular, is a subset of AI that utilizes neural network architectures loosely inspired by the human brain. Although most of the earlier AI methods led to applications with subhuman performance, deep learning has recently demonstrated performance that matches and often surpasses humans (1–3). Recognizing its power, medical researchers and scientists have been exploring the role of deep learning methods in a variety of radiologic applications (4), including early detection, risk prediction, prognosis assessment, treatment planning, as well as lesion and organ segmentation, for various diseases.

However, AI models are often considered “black boxes”; although we may be able to get accurate decisions and

predictions, we cannot easily or clearly explain or interpret the logic behind an AI model’s outputs. But how can we gain useful insights? Which aspects of an AI model should we strive to elucidate, and what are the computational tools we need to achieve that? These questions are important to address when the issue of AI model interpretability or explainability is raised.

Although interpretability can be a broad, poorly defined concept, in AI it can be thought of as the ability to explain or to present in understandable terms to a human the cause of a model’s decision. The more interpretable or explainable an AI model is, the easier it is for someone to comprehend why certain decisions or predictions have been made. The concept of interpretability in AI is not new. Inherently interpretable AI models, such as decision trees and decision rules, have received a great deal of attention from the research community. Moreover, several methods for feature importance analysis have been developed to offer insights into what the model is learning and what factors might be relevant. However, in the case of deep learning, the situation is more complicated. A key difference is that traditional AI systems primarily rely on predefined, engineered (ie, handcrafted) features based on expert knowledge. In contrast, deep learning allows the algorithm to learn discriminative features from data automatically, offering the ability to approximate very complex nonlinear relationships through several layers and neurons. Therefore, deep learning interpretability becomes even more challenging, as we need not only to explain the relationship between features and decisions or predictions, which usually involves a vast number of parameters, but also to understand the learned features themselves.

In this issue of *Radiology: Artificial Intelligence*, Reyes et al address this important topic by reviewing computational approaches to interpret deep learning models, and discuss gaps and challenges that need to be addressed, focusing primarily on radiologic applications and their translation into clinical practice (6). The authors are to be commended for the comprehensive summary of existing interpretability technologies, as well as for the intuitive, yet detailed, descriptions of such sophisticated computational tools. Reyes and colleagues have categorized various interpretability methods, provided explanatory figures, demonstrated related applications, and offered a thorough discussion on the advantages and limitations.

If a deep learning model performs well, why shouldn’t we simply trust the model and ignore why it made a specific decision? A lot of this has to do with the impact that

From the Department of Radiology, University of Pennsylvania, 3700 Hamilton Walk, Richards Bldg, Room D702, Philadelphia, PA 19104. Received May 5, 2020; revision requested May 6; revision received May 13; accepted May 14. Address correspondence to D.K. (e-mail: Despina.Kontos@upennmedicine.upenn.edu).

See also the article by Reyes et al in this issue. Conflicts of interest are listed at the end of this article.

Radiology: Artificial Intelligence 2020; 2(3):e200088 • <https://doi.org/10.1148/ryai.2020200088> • Content code: **IN** • ©RSNA, 2020

This copy is for personal use only. To order printed copies, contact reprints@rsna.org

a model might have in the real world. Models used in a low-risk environment, where a mistake will not have serious consequences, will have a far less impact than the ones used in radiology, for example, to assess the risk of malignancy in screening mammographic images. Reyes et al illustrate the critical role of model interpretability on many levels in radiology, including auditability, model verification, enhancing trust, and clinical translation. Deep learning models can be debugged, audited, and verified only when they can be interpreted, and their interpretability is valuable in the development phase as well as after deployment. On the one hand, interpretation of an erroneous decision or prediction helps one understand the cause of the error and delivers a direction for how to fix it. On the other hand, an interpretation of a correct decision or prediction helps verify the logic for a specific conclusion, making sure that causal relationships are picked up and alleviating potential suspicion about confounding or bias. Moreover, it is easier for radiologists and patients to trust a model that explains its decisions, including its failures, compared with a “black box”. Together, by verifying the performance of the model and enhancing user trust, interpretability can ultimately accelerate the translation of useful deep learning tools into routine clinical practice.

Additional intriguing aspects brought up by the authors are related to the potential of model interpretability methods to identify mislabeled data, as well as challenging cases. The identification of mislabeled data, or even inconsistent data from different clinical sites, would allow for better quality control of training data. Moreover, it could help identify and alleviate potential effects of different imaging acquisition hardware, scanning protocols, and patient populations on the acquired imaging data, that may substantially affect the ability of an AI system to generalize to previously unseen images. Such improvements can substantially enhance the reliability of the model, as well as its robustness to continuous technological advancements in imaging systems and protocols. Moreover, access to challenging cases, as well as matched or counterexamples, seen by the model, can prove to be particularly useful in the training process of new radiology experts. Interpretability methods could even lead to hypothesis generation by providing useful insights beyond the diagnostic or prediction task of the model. Identifying previously unappreciated patterns in data and interactions among multiomic features (5) may indicate novel hypotheses that can then be explored at the mechanistic level.

Interpretability in AI is evolving into a whole new research field. Moving forward, this research area is expected to receive

more and more attention, particularly for radiologic applications. Despite the current availability of various interpretability methods, applications in radiology are still mostly unexplored. We likely will soon start seeing more studies, conferences, and special journal issues focusing on interpreting AI models in radiology, as well as new task-specific interpretability methods aiming to elucidate how AI models make decisions, for example, in medical image segmentation, survival analysis, and prognosis assessment.

The advent of deep learning is poised to change the delivery of health care dramatically in the near future. In radiology, the potential is immense due to the desire to provide fully automated interpretation, enhance workflow efficiency, and lower health care costs. To support these goals, interpretability methods will be critical, as will be the involvement of radiology experts in the development of such approaches. What is the scope and level of interpretability needed for different diagnostic and prediction tasks? What is the preferable way to access interpretable insights in the daily clinical routine? How much time could be allowed for interpretability methods in clinical practice? To answer these questions, researchers, AI engineers, and radiologists will need to work hand-in-hand. With the help of interpretability methods, the next generation of radiology will likely experience full use and enhanced trust in AI and deep learning.

Disclosures of Conflicts of Interest: A.G. Activities related to the present article: institution receives grant from Susan G. Komen for the Cure Breast Cancer Foundation (PDF17479714). Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. D.K. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: institution receives grant from Hologic. Other relationships: disclosed no relevant relationships.

References

1. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst* 2019;111(9):916–922.
2. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature* 2015;518(7540):529–533.
3. Xiong W, Droppo J, Huang X, et al. Toward human parity in conversational speech recognition. *IEEE Trans Audio Speech Lang Process* 2017;25(12):2410–2423.
4. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018;18(8):500–510.
5. Haas R, Zelezniak A, Iacovacci J, Kamrad S, Townsend S, Ralser M. Designing and interpreting ‘multi-omic’ experiments that may change our understanding of biology. *Curr Opin Syst Biol* 2017;6:37–45.
6. Reyes M, Meier R, Pereira S, et al. On the interpretability of AI in radiology: challenges and opportunities. *Radiol Artif Intell* 2020;2(3):e190043.