Cureus
Part of SPRINGER NATURE

Open Access
Abstract

Cureus

Black Boxing in Healthcare Artificial Intelligence

Open Access
Abstract
Published 02/05/2026

Shikhar Gupta [1]

1. Penn Healthcare Review, University of Pennsylvania, Philadelphia, USA

Corresponding author: Shikhar Gupta, shikharg@sas.upenn.edu

Categories: Healthcare Technology
Keywords: artificial intelligence in medicine, blackboxing, transparency, transparency in ai, transparency in ai and llms black box

How to cite this abstract
Gupta S (February 05, 2026) Black Boxing in Healthcare Artificial Intelligence. Cureus 18(2): a1665

Abstract

Introduction

With applications ranging from Face ID to Google searches, artificial intelligence is gaining popularity due to its potential to drastically increase the efficiency and accuracy of systems. A subset of AI, called “machine learning,” is being utilized, especially in the field of medicine. These machine, or deep learning models, are systems that are trained with relevant data sets in order to automate the analyzation process. However, as the technology continues to be adopted, concerns regarding understanding are being raised. This is referred to as a “black box.” The black box metaphor refers to a box where the inputs and outputs are known without any context of the steps in between. Many physicians find themselves lacking an understanding of how machine learning systems work, which leads to the first issue: a lack of transparency. Physicians also wish to add biological logic to algorithms instead of using a purely data-driven approach, which leads to the second problem: program interpretability. As a result, a push for greater transparency in the application of such technology is gaining traction. This article examines the recommendation for physician transparency and interpretability in medical artificial intelligence applications.

Methods

While this technology has proved to be of substantial benefit to the industries that utilize them, the metaphorical black box that they sit in has raised questions, notably among physicians. The complaints suggest that healthcare providers simply do not have enough information or expertise to effectively understand the technology. That is, when working with AI models, clinicians are unable to answer a series of critical questions:

- Can I verify that this model is actually suitable for the particular patients I am seeing?
- Can I have faith that it will provide a reliable result to inform my clinical decision making?
- Does this model fit what I already understand about human disease?
- What exactly has the model discovered?

The list of approved algorithms continues to grow while healthcare providers are often left clueless while choosing one. Once the physician knows what they want their AI system to do, they then go to databases such as Data Science Institute AI Central or the FDA list of approved products to pick any program that they believe will work for them. If it doesn’t work, the physician refers back to the published list. This inefficiency all stems from a lack of understanding from the physician since they aren’t able to choose a program they know will fit their needs.

In an effort to address this problem, the European Commission issued the world’s first legislation aiming to regulate AI in April 2021. The Artificial Intelligence act deemed an AI system adequately transparent if it allowed its users to understand an AI output and apply it accordingly. However, this outline is extremely vague and has led to many different interpretations on the problem.

Results

There are multiple viewpoints on how to address this problem of a lack of transparency. In the case of the

lists of approved machine learning algorithms, Dr. Keith Dreyer, the Chief Science Officer of the American College of Radiology's Data Science Institute, recommends that the FDA should standardize the information that is available in their catalogs. According to him, metrics such as evaluation parameters, test demographics, and findings would help physicians better understand a program before selecting it for their specific need. This also helps eliminate the inefficiencies and false negative risks associated with choosing the incorrect model.

Along with altering the information available when selecting a system, some experts suggest that the systems themselves should be modified to include transparency measures. Poon and Sung elaborate on this idea with their model of black box interpretability. They argue that transparency can be increased in a stepwise method. In their model, large volumes of structural data are inputted into an algorithm. From there, the algorithm combines biological knowledge with its own computational knowledge in a learning process that can be supervised. The system accomplishes this through well known tools such as Bayesian networks, fuzzy logic, and random forests. After that, the system enters the black box stage but Poon and Sung propose the use of popular interpreting techniques such as Garson's Algorithm, Lek's profile, and local interpretable model-agnostic explanation (LIME) method. This model aims to lighten up the metaphorical black box surrounding steps in the system. As the interpretability is increased, physicians and patients can place greater amounts of trust in artificial intelligence.

Ordish et al, describe a system detecting breast cancer risk as one instance where an existing model could be modified with transparency in mind. In this new system, for instance, a heat map method could be used to allow the radiologist to focus on specific regions of the mammogram that the AI is considering before making its decision. The system could also utilize standard text samples to describe its justification for its choices. This potential modification would allow for the operating physician with greater insight into how the system itself is working. This idea of modification with the goal of transparency is being implemented at Moorfields Eye Hospital in London by DeepMind. Their machine learning model first triages images of patient eyes before analyzing 3D eye scans. The system is able to then identify cases that require a referral. In order to provide the overseeing physician with maximum transparency, the program is able to provide and rate multiple possible explanations for its decisions and visually identify parts of the eye.

On the other hand, many experts refute the idea of pausing the usage of machine learning and argue that transparency isn't a necessity. Eric Topol, the director of the Scripps Research Translational Institute says, "If prospective trials validate these models… there's every reason to bring them forward to the clinic with the hope of achieving a happy symbiosis between doctors and machines - even if doctors have no idea why the models work." His sentiment is echoed by many experts. They believe that, while safety is a priority, physicians should not allow black boxing to prevent the usage of reliable algorithms. As highlighted previously, there are many operational and potential machine learning programs capable, for example, of capturing subtle visual nuances that humans may not discern. Regina Barzilay, a breast cancer deep learning model creator from MIT, asks, "does it really mean that we shouldn't be benefiting from [artificial intelligence] because our visual capacity is limited?"

Conclusion

We find ourselves in a reality where artificial intelligence systems are being widely adopted in healthcare settings. While they have a great and proven potential to improve care, it is not always explicitly clear to healthcare providers and patients how such systems work. Some experts argue that the black boxing of medicine has led to a need for modifications to improve interpretability. Others argue that black boxing isn't an inherent issue and can be overlooked as long as a system is reliable and safe. This current debate calls upon artificial intelligence developers to reconsider how they are creating healthcare programs while also encouraging healthcare providers and patients to reevaluate their priorities when considering their treatment plan.

REFERENCES

1. Bender, Eric. "Unpacking the Black Box in Artificial Intelligence for Medicine." Undark Magazine, December 4, 2019. https://undark.org/2019/12/04/black-box-artificial-intelligence/.
2. Dreyer, Keith. "Why Does AI Medical Device Transparency Matter." www.acr.org, October 21, 2021. https://www.acr.org/Advocacy-and-Economics/Voice-of-Radiology-Blog/2021/10/21/Why-Does-AI-Medical-Device-Transparency-Matter.
3. Kiseleva, Anastasiya, Dimitris Kotzinos, and Paul De Hert. "Transparency of AI in Healthcare as a Multilayered System of Accountabilities: Between Legal Requirements and Technical Limitations." Frontiers in Artificial Intelligence 5 (May 30, 2022).

https://doi.org/10.3389/frai.2022.879603.

4. Ordish, Johan, Colin Mitchell, Hannah Murfet, Tanya Brigden, and Alison Hall. "Black Box Medicine and Transparency." PHG Foundation. University of Cambridge, n.d. https://www.phgfoundation.org/report/black-box-medicine-and-transparency.
5. Poon, Aaron I F, and Joseph J Y Sung. Development of Algorithm by Artificial Intelligence (AI)/Machine Learning. March 2021. Online image. Journal of Gastroenterology and Hepatology. https://onlinelibrary.wiley.com/doi/10.1111/jgh.15384#.
6. Poon, Aaron I F, and Joseph J Y Sung. "Opening the Black Box of AI Medicine." Journal of Gastroenterology and Hepatology 36, no. 3 (March 2021): 581–84. https://doi.org/10.1111/jgh.15384.
7. Zhang, Zhongheng, Marcus W. Beck, David A. Winkler, Bin Huang, Wilbert Sibanda, and Hemant Goyal. "Opening the Black Box of Neural Networks: Methods for Interpreting Neural Network Models in Clinical Applications." Annals of Translational Medicine 6, no. 11 (June 2018): 216–16. https://doi.org/10.21037/atm.2018.05.32.