Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220709

Ethical Issues in the Utilization of Black Boxes for Artificial Intelligence in Medicine

Diva BELTRAMIN[a] , Eugenia LAMAS[b] and Cédric BOUSQUET[a,b,1]
[a] *Service de santé publique et information médicale, CHU Saint Etienne, France*
[b] *Sorbonne Université, Université Sorbonne Paris Nord, INSERM, Laboratoire d'informatique médicale et d'ingénierie des connaissances en e-santé, LIMICS, F-75006 Paris, France*

Abstract. Artificial Intelligence (AI) has made major progress in recent years in many fields. With regard of medicine however, the utilization of AI raises numerous ethical questions, especially since newer and much more accurate algorithms function as black boxes. A trade-off must then be made between having algorithms being very accurate and effective, and algorithms that are explainable but less proficient. In this paper we address the ethical implications of utilizing black box algorithms in medicine.

Keywords. Black boxes, Artificial intelligence, Ethics; Decision making, Trust

1. Introduction

Artificial Intelligence has proven to be, in some cases, just as efficient as some experts at very specific tasks [1,2]. However, several doctors are still skeptical about its application in the medical field, questioning the reliability of modern artificial intelligence (AI). In fact, a possible explanation of this mistrust is that some of the latest AI methods utilize algorithms that work as black boxes. Rudin defines black boxes as a model that could be "either (1) a function that is too complicated for any human to comprehend or (2) a function that is proprietary [3]." In this paper, we mainly focus on the first kind, especially deep learning models. These modern architectures implement several layers of hidden, artificial neural networks that, by means of complex associations between statistical weights and adjustments according to the resulting prediction, are capable of learning. Therefore, algorithms featuring black boxes may be exceptionally efficient, but it is impossible to know exactly the relation between the parameters learned by training the algorithm and the predictions. For example, the Deep Gestalt system, which is an accurate model to predict the presence of facial phenotypes associated with genetic disorders, cannot explain to the end user what are the facial characteristics leading to the prediction [4].

Therefore, it becomes evident that ethical problems arise when an algorithm that is very accurate but not explainable guides a diagnosis. In this regard, in 2018 a high-level expert group on AI elaborated a document entitled "Ethics guidelines for trustworthy AI" [5]. This group evaluated the implications of the use of black boxes in medicine and

[1] Corresponding Author, Dr Cedric Bousquet, SSPIM, Bâtiment CIM42, chemin de la Marandière, Hôpital Nord, 42055 Saint Etienne, France; E-mail: cedric.bousquet@chu-st-etienne.fr.

concluded that the degree of explicability required is highly dependent on the consequences for the patient if the output is inaccurate or even erroneous, and that explicability is to be made an ethical imperative for AI algorithms.

Nevertheless, different points of view exist with regard of the utilization of black boxes in medicine. For Kundu, AI in medicine must be explainable and not rely on a black box at all [6]. Rudin [3] is very reluctant on black boxes, advocating for the utilization of explainable models instead. On the other hand, Babic [7] is convinced that enforcing black boxes to be explainable would not be beneficial, and instead give users an unwarranted sense of security when interpreting the predictions. Indeed, explanations are based on a substitute model that is trained to mimic the output of the black box. However, the substitute model imperfectly reproduces the predictions of the black box and is not sufficiently robust. The WHO Guidance document [8] has listed ensuring transparency, explicability and intelligibility as a core principle, and considers that regulators, clinicians and patients should be able to understand decisions made by AI to the maximum possible extent. At the same time, it acknowledges that the requirement of explicability may not always be possible or desirable in medicine. The ability to explain how AI systems arrive at judgment should not take precedence on the evaluation of the system's performances when it could improve the delivery of health care in prevention, diagnosis and treatment of disease. In this paper, we briefly explore the ethical issues raised by the use of black box algorithms in medical practice.

2. Arguments opposing the utilization of black boxes in medicine

Considering ethical implications of AI in medicine, the possibility of biases may concern any artificial intelligence system. Biases generally fall into three main categories [9]. The first category is when imbalanced or misrepresentative data is fed as training data to algorithms, that could completely ignore misrepresented classes, for example using datasets not featuring enough data about vulnerable groups. The second category is bias generated by a faulty algorithm. Finally, the third category is that related to human error, indeed the subtlest, as it is a result of long-held societal prejudices. In the case of black boxes however, biases appear to be much more difficult to identify and prevent.

This has been highlighted very well by Obermeyer [10], evidencing that the U.S. health care system was using a commercial algorithm which had an evident racial bias against Black patients. Obviously, the last thing we wish is for an algorithm to be utilized despite having inherent biases, causing erroneous medical decisions. If black boxes are to be utilized in the clinical practice, extra care should be put in order to prevent the insurgence of the said biases. During the COVID-19 pandemic, researchers all over the world tried to build effective models capable of diagnosing the disease from chest radiographs and computed tomography scans. However, as extensively documented by Roberts et al. [11], none of the 320 papers featuring artificial intelligence yielded satisfying results. In fact, among the papers considered in the final analysis by Roberts et al., 55 out of 62 had a high risk of bias.

Another interesting and fundamental aspect of ethics related to black boxes is that the relationship between the patient and the physician is undoubtedly altered. Furthermore, Kundu [6] raised the issue that if a physician does not know why an algorithm suggests a said diagnostic, he would not be able to effectively communicate and justify his decision to the patient. In turn, this would lead to the patient potentially bearing incertitude on whether or not to trust the physician.

3. Arguments in favor of the utilization of black boxes in medicine

There are certain cases in which there is not an imperative need for algorithms to be explainable, in order to be utilized. Algorithms can work side-by-side with physicians, giving them an extra tool to work with, rather than substituting their decisions. For example, by utilizing black boxes in medicine to automatically chart symptoms during medical consultations [12]. In this specific case the algorithm does not need to be explainable, because it is limited at collecting what it hears and does not interfere with the diagnostic process.

4. Discussion

Ethical challenges posed by the utilization of black boxes are way more complex than those issued by simpler artificial intelligence methods [13]. Indeed, some voices in the scientific community prefer to be very cautious and are not keen on utilizing black boxes at all [6]. Others prefer to prioritize the trade-off between efficacy and explicability. On the matter of responsibility, when a clinician does not agree with the prediction made by the algorithm, the WHO states that the responsibility is the clinician's only [8]. In fact, different levels of liability exist, depending on whether it is the clinician or the AI making a mistake and whether or not there are consequences for the patient [14].

The question of using black boxes in the medical practice has proven to be controversial. We limited our paper to present some arguments in favor or against the usage of black boxes in Medicine. Perspectives are to perform a scoping review that would provide more in-depth analyses of the different points of view.

As the utilization of black boxes seems to be more and more preponderant in medicine, the European Commission began to ponder upon ethical issues that are being introduced [5]. The high-level expert group on artificial intelligence set up by the European commission introduced explicability as the fourth ethical principle, as it was deemed imperative for Artificial Intelligence to be explainable. Moreover, in order to induce physicians and patients alike to trust modern AI, seven essential requirements have been proposed to be introduced in future algorithms, and among them is transparency. This requirement is the product of taking into account three dimensions: traceability, explicability and communication. Traceability allows to explore how data has been gathered and models were trained. Explicability lets end users understand how a prediction has been reached. Finally, communication is necessary between AI designers and physicians, in order to build algorithms that make sense clinically.

Complex data such as medical images or hospital discharges imply nonlinear relations between inputs and predictions, and imply deep learning models with millions of parameters. These black box models have large potential to improve disease prevention, access to medical care and could become valuable tools for physicians. However, relying on opaque methods is not an optimal choice, and that is why the European group of experts introduced explicability as an imperative ethical principle when recommending use of AI in medicine [5]. A current approach in order to overcome this issue is to implement features that make sense for physicians based on findings of deep learning.

An inspiring example of this approach is an algorithm called DeepLive [15], designed to help dermatologists in the precocious diagnosis of skin lesions. The system exploits the power of deep learning to perform keratinocyte nuclei segmentation, and to

infer a series of quantitative, reproducible and biologically relevant metrics to describe keratinocytes. Then, atypia is defined using different algorithms: simple rules based on expert knowledge, and machine learning definitions. Models interpretability is achieved using weights for Logistic Regression, feature importance for Isolation Forrest, Shapley values and feature importance for XGBoost. Clinicians taking advantage of this tool can more easily evaluate a lesion, assess its severity and evaluate the effects of treatment.

5. Conclusions

In order to exploit the potential of artificial intelligence, it is desirable to add functionalities that allow to better understand how decisions are made. For the time being some algorithms, such as Deep Gestalt, are still not explainable and thus should be evaluated mainly on their performances. Nevertheless, algorithms such as DeepLive pave the way for AI systems that consist of both a black box and an interpretable algorithm. Such approach may improve trust by patients and physicians alike, and also would be much easier to regulate from a legislative point of view.

References

[1] McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, Back T, Chesus M, Corrado GS, Darzi A, Etemadi M. International evaluation of an AI system for breast cancer screening. Nature. 2020 Jan;577(7788):89-94.
[2] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. nature. 2017 Feb;542(7639):115-8.
[3] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206-215.
[4] Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, Basel-Salmon L, Krawitz PM, Kamphausen SB, Zenker M, Bird LM. Identifying facial phenotypes of genetic disorders using deep learning. Nature medicine. 2019 Jan;25(1):60-4.
[5] High-Level Expert Group on AI. Ethics Guidelines for trustworthy AI. 2019. Available at: https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1.
[6] Kundu S. AI in medicine must be explainable. Nat Med. 2021;27(8):1328.
[7] Babic B, Gerke S, Evgeniou T, Cohen IG. Beware explanations from AI in health care. Science. 2021;373(6552):284-286.
[8] World Health Organization, Ethics and governance of artificial intelligence for health: WHO Guidance, Geneva. 2021. Available at:https://www.who.int/publications/i/item/9789240029200.
[9] Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. Patterns. 2021 Oct 8;2(10):100347.
[10] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-53.
[11] Roberts M, Driggs D, Thorpe M. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. Nat Mach Intell. 2021;3:199-217.
[12] Rajkomar A, Kannan A, Chen K, Vardoulakis L, Chou K, Cui C, Dean J. Automatically charting symptoms from patient-physician conversations using machine learning. JAMA internal medicine. 2019 Jun 1;179(6):836-8.
[13] Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In Artificial intelligence in healthcare. Academic Press. 2020;295-336.
[14] Price WN, Gerke S, Cohen IG. Potential liability for physicians using artificial intelligence. JAMA. 2019;322(18):1765-1766.
[15] Fischman S, Pérez-Anker J, Tognetti L, Di Naro A, Suppa M, Cinotti E, Viel T, Monnier J, Rubegni P, Del Marmol V, Malvehy J. Non-invasive scoring of cellular atypia in keratinocyte cancers in 3D LC-OCT images using Deep Learning. Scientific reports. 2022 Jan 10;12(1):1-1.