AI explicability in medicine and healthcare: fighting against the return to the paternalism

Lorella Meola, Department of Cultural Heritage, University of Salerno (Italy)

International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: explanability, paternalism, autonomy

Extended Abstract

The use of Artificial Intelligence (AI) in medicine and healthcare increases the efficiency and accuracy of clinicians and their medical decision-making process. In particular, deep learning technologies can gather and analyse a great amount of patient data, whose connection can improve the precision of diagnosis, prognosis, therapy, prediction, and prevention of disease [1; 2]. AI can change and improve medical practice and biomedical research [3]; however, it raises several ethical challenges and problems [4; 5; 6].

Notably, clinicians can be assisted or even replaced by AI sytems in the decision making process, improving care quality, that means care accuracy, and optimizing the time spent with the patients [7]. However, the lack of transparency and the opacity of AI [8; 9; 10], in particular of machine learning algorithms, could mine the therapeutic relationship; thus, it could weaken trust and discourage patients and destroy the basis of their self-determination.

Self-learning AI systems are often defined as black-boxes [11] because of the lack of algorithmic transparency in data interpretation. Thus, AI systems can rise the ethical and the epistemological issue of explicability [12]. The widespread AI decision-making urges an explanation of how it works. Furthermore, a new set of methods and techniques in the application of AI technology, AI Xplainable (XAI), is spreading. XAI is based on the principle that the results of the solution can be understood by humans and it contrasts with the concept of the "black box" in machine learning where even their designers cannot explain why the AI arrived at a specific decision. It is paramount that explicability is not just a technical issue, but it requires a multidisciplinary approach [13]. Despite not having a common definition of explicability, however, it is widely considered as a necessary principle to regulate the use of AI, in particular in medicine and healthcare. Luciano Floridi argues the need of explicability, considered as the synthesis of the epistemological sense of intelligibility (as an answer to the question “how does it work?”) and ethical concept of accountability (“who is responsible for the way it works?”)[14]. Adopting this principle, it could be possible to identify systematic errors and improve AI performance on one hand, and find out responsibilities on the other hand, drawing the boundaries between autonomy and accountability of humans and machines, reducing human risks and safeguarding their safety. This means restoring the balance of human-machine relations [15].

Explicability is becoming an urgent issue in medicine too [16; 17]; it is as important as the four principles of biomedical ethics, set out by James Childress and Tom Beauchamp, in the biomedical literary classic *Principles of biomedical ethics*.

Some scholars consider explicability a way to foster medicine trust and restore the relation between the doctor and the patients in their specific singularity [18]; others consider it a way to preserve the empathy between clinician and patient [19].

On the other hand, instead, some scholars define explicability as a useless criterion for medical AI. For example, Alex London argues that, having medicine often applied treatments without understanding clearly their causal relations, there is no reason a different standard for AI, above all considering that AI systems are more accurate than traditional medical methodologies. According to London, the criterion of accuracy is more important than that of explicability [20].

A dichotomy rises from the increasing literature about AI: generally speaking, we can summarize that some scholars consider result accuracy more important than the comprehension of causal mechanism results, because it is the result that matters, not the means employed; instead, other scholars consider that we cannot give up the comprehension of the medical choice's reasons, by shifting the power from patient-clinician dialogue to an opaque machine, for the sake of accuracy.

We are going to analyse the ethical consequences of the epistemological dilemma between explicability and accuracy by supporting the importance of AI explicability in medicine [21]. We are going to reread this dilemma in the broader and classical framework of the links and oppositions between medicine based on empircal evidence and centred-patient medicine. We are going to outline that a lack of algorithmic transparency could compromise patient autonomy, a pitting point of biomedical ethics since the second post-war period. The failure of the clinician to mediate between the rigor of science and the complex singularity of each patient and the unavailability of information for the patient, who is not able to decide, and for the doctor, who is not able to explain, seem to introduce a new form of paternalism. The delegation of the power to a machine could tighten up the paternalistic effects on patients and society: if we assume a "simmetric relation" with the machines, so that we are not entitled to ask for explanations about AI decisions, we are going to interact with a system that claims to know what is good for us because it knows health standards. This knowledge is based on a strict classification of the patients, according to standardized parametres, that develop a quantification of the concepts of health and disease. Moreover, it does not take into account the patients in their singularity, with their specific biographies, in which health and disease become meaningful. Giving up to the traditional doctor-patient relationship might important ethical and political effects not only on patients but on the whole society.

As the world moves towards more and more automation, the ethical discussion of the role of AI in healthcare is becoming more and more relevant. It poses new risks and our concerns do not deal with the role of the doctor, but with the autonomy of the patient in medical artificial intelligence. While medicine seems to turn into a mere application of empirical facts it urges to define which are the goals of medicine too.

[1] A. Becker*, Artificial Intelligence in medicine: What it is doing for us today?*, “Health Policy and Technology”, 8(2), 2019, 198-205.

[2] E. Topol, *High-performance medicine: the convergence of human and artificial intelligence*, in “Nature Medicine”, 25, 2019, 44-56.

[3] A. Blasimme, E. Vayena*, The Ethics of AI inBiomedical Research, Patient Care, and Public Health*, in M. S. Dubber, F. Pasquale, S. Das, *The Oxford Handbook of Ethics of AI*, Oxford University Press, 2020, pp. 703- 718.

[4] D. S. Char, M. D. Abràmoff, C. Feudtner, *Identifying ethical considerations for machine learning healthcare applications*, “The American Journal of Bioethics”, 20(11), 2020, 7–17. [5] T. Grote, P. Berens, *On the ethics of algorithmic decision-making in healthcare*, 2Journal of Medical Ethics2, 46(3), 2019, 205–211.

[6] E. Vayena, A. Blasimme, I. G. Cohen, *Machine learning in medicine: Addressing ethical challenges*, “PLoS Medicine”*,* 15(11), 2018, 1- 4.

[7] E. J. Topol, *Deep Medicine - How Artificial Intelligence Can Make Healthcare Human Again*. 2019, New York, Basic Books.

[8] B. Heinrichs, S. B. Eickhoff, *Your Evidence? Machine Learning algorithms for medical diagnosis and prediction*, in 2Human Brain Mapping”, 41, 2020,1435- 1444.

[9]J. Burrel, *How the machine ‘thinks’: Understanding Opacity in machine learning algorithms*, in “Big Data Society”, 2016, <https://doi.org/10.1177/2053951715622512>.

[10] T. Zarsky, *The Trouble with Algorithmic Decisions: An Analytic Road Map to Examine Efficiency and Fairness in Automated and Opaque Decision Making*, in “Science, Technology and Human Values”, 41 (1), 2016, 118–132.

[11] M. Carabantes, *Black-box artificial intelligence: an epistemological and critical analysis*, in “AI & Society”, 35, 2020, 309-317.

[12] D. Doran, S. Schulz, T. R. Besold, *What does explainable AI really mean? A new conceptualization of perspectives*, in T. R. Besold, O. Kutz, (Eds.), *Proc. First Int. Workshop on Comprehensibility and Explanation in AI and ML*, Volume 2071 of *CEUR Workshop Proceedings*, 2017, 1-8.

[13] J. Borrego-Díaz, J. Galán-Paèz, *Explainable Artificial Intelligence in Data Science*, “Minds and Machines”, 32, 2022, 485- 531. [14] L. Floridi et alii, *AI4People –an ethical framework for a good AI society: opportunities, risks, principles, and recommendations*, “Minds and Machines”,28, 2018, 689- 707.

[15] G. Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci, Roma, 2020

[16] J. M. Durán, M. Sand, K. Jongsma, *The ethics and epistemology of explanatory AI in medicine and healthcare*, “Ethics and Information Technology”,42, 2022.

[17] J. Amann et alii, *Explainability for Artificial Intelligence in healthcare: a multidisciplinary perspective*, “BMC Medical Informatics and Decision Making”, 20, 310, 2020. [18] C. Poppe, G. Starke, *Karl Japsers*

and Artificial neural nets: on the relation of explaining and understanding artificial intelligence in medicine, "Ethics and Information Technology", 24, 3, 2022. [19] F. Cabitza, R. Rasoini, G. F. Gensini, *Unintended Consequences of Machine Learning in Medicine*, in "*JAMA",* 318 (6): 517–518, 2017.

[20] A. J. London, *Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability*. "Hastings Center Report", 49(1), 2019, 15–21.

[21] C. Herzog, *On the Ethical and Epistemmological Utility of Explicable AI in medicine*, "Philosophy & Technology", 50, 2022.