Trust Through Explanation? On the claim for explainable medical decision support systems

Sebastian Schleidgen, FernUniversität in Hagen (Germany)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: medical decision support systems, explainability, doctor-patient relationship

Extended Abstract

Artificial intelligence (AI)-based systems will play an increasingly significant role in clinical practice. For example, decision support systems (for tumor identification and classification in diagnostic imaging) or AI-based systems for automatic data analysis with regard to diagnosis, prognosis and prediction are already being developed and increasingly used (McKinney et al. 2020, Holzinger et al. 2019).

On the one hand, this is associated with the hope that AI will be able to answer certain questions faster, more efficiently, and more effectively than human physicians, and that AI will thus be able, for instance, to compensate for the shortage of medical personnel (Buch 2018). On the other hand, a number of ethically relevant problems is raised by the eventual use of such technologies, in particular issues of quality and safety (Challen 2019, Pisano 2020) as well as of unfairness (Mehrabi et al. 2022), which are primarily reflected in systematically inaccurate or erroneous results of AI-based systems. To complicate matters, the black box-nature, especially of deep learning-based systems, may result in such inaccuracies and errors being noticed only at a late stage.

In addition, the risk for so-called "workflow disruptions" (Yu et al. 2018, Challen et al. 2019) has been documented for contexts of using AI-based decision support systems in clinical practice: there is, first, a tendency of physicians to decision-making passivity, i.e., to adopt AI diagnoses without further review. Second, physicians tend to automation complacency, meaning that in cases where physician and AI diagnoses match, the latter is not verified. And third, physicians often seem to show some alert fatigue, i.e., they do not take AI-generated indications of possible positive diagnoses seriously. Consequently, such workflow disruptions may increase the risk of systematically erroneous diagnoses, and thus may further exacerbate the immanent problems of AI use in clinical practice.

It has been argued that both the immanent problems of AI and possible workflow disruptions may have a negative impact on the doctor-patient relationship or even patient trust in medical care (Ferretti et al. 2018, Astromske et al. 2021). A frequent proposal to address these issues consists in the demand for explainable deep learning-based decision support systems, i.e. systems whose operations, roughly speaking, "can be understood by human[s]" (Adadi et al. 2018: 52141).

Regarding medical decision support systems, this claim is often supported by two types of arguments: first, it is stated that, if medical decision support systems were explainable, causes of quality, safety and unfairness issues could be understood, and therefore be remedied or avoided effectively and efficiently. Knowledge of this would strengthen the trust of patients in such systems and, thus, also the

doctor-patient relationship (A1). Second, it is argued that informed consent – as part of a good doctor-patient relationship – requires a certain level of patient information, for which to gain explainable medical decision support systems would be necessary (A2).

In the proposed talk, I will address the question as to whether (or in what sense) these arguments succeed, or as to whether (or in what sense) explainable medical decision support systems may contribute to a good doctor-patient relationship. Apart from the principal question of whether the demand for explainable deep learning-based systems is reasonable or whether such systems can be explainable at all, regarding A2 I will argue that medical decision support systems need not be explainable to patients with a view to a good doctor-patient relationship. This, I will claim, follows from the basic reasons we have for the use of such systems as well as the conditions for supporting these very reasons. However, in contexts of using medical decision support systems, doctors play an important role in supporting these reasons, i.e., they must be able to provide reasons to patients regarding the use of such systems. In view of this, regarding A1, I will argue that medical decision support systems must be explainable to doctors in a specific sense to enable them as reason-givers when it comes to using medical decision support systems. With this, I will not only examine the two arguments for explainable medical decision support systems, but also try to shift the debate from claims of explainability aiming at patient understanding to explainability necessary for doctor understanding.

References

Adadi A, Berrada M (2018): Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6: 52138-52160

Astromské K, Peičius E, Astromskis P (2021): Ethical and legal challenges of informed consent applying artificial intelligence in medical diagnostic consultations. AI & SOCIETY 36: 509-520

Buch VH, Ahmed I, Maruthappu M (2018): Artificial intelligence in medicine: current trends and future possibilities. British Journal of General Practice 68 (668): 143-144

Challen R, Denny J, Pitt M et al (2019): Artificial intelligence, bias and clinical safety. BMJ Quality & Safety 28 (3):231-237

Ferretti A, Schneider M, Blasimme A (2018) Machine learning in medicine. European Data Protection Law Review 4 (3): 320-332

Holzinger A, Langs G, Denk H et al (2019): Causability and explainability of artificial intelligence in medicine. WIREs Data Mining and Knowledge Discovery 9 (4): e1312

McKinney SM, Sieniek M, Godbole V et al (2020): International evaluation of an AI system for breast cancer screening. Nature 577: 89-94

Mehrabi N, Morstatter F, Saxena N et al (2022): A Survey on bias and fairness in machine learning. arXiv: 1908.09635

Pisano ED (2020): AI shows promise for breast cancer screening. Nature 577: 35-36

Yu K-H, Kohane IS (2018): Framing the challenges of artificial intelligence in medicine. BMJ Quality & Safety 28 (3): 238-241