Cambridge Quarterly of Healthcare Ethics (2024), 33: 3, 307–309
doi:10.1017/S0963180123000671

CAMBRIDGE UNIVERSITY PRESS

EDITORIAL

Editorial: The Ethical Implications of Using AI in Medicine

Orsolya Friedrich and Sebastian Schleidgen

Institute of Philosophy, FernUniversitaet in Hagen, Hagen, Germany
Corresponding author: Sebastian Schleidgen; Email: sebastian.schleidgen@fernuni-hagen.de

The use of artificial intelligence (AI) and machine learning (ML) plays an increasingly significant role in medical research and clinical practice. In clinical practice, AI promises to support healthcare professionals in topics such as prevention, diagnosis, prognosis, treatment planning, and overall decision-making. On the one hand, the (future) use of AI systems in clinical practice is associated with the hopes of increasing the efficiency and effectivity of diagnostic and therapeutic processes as well as improving their quality and thus enhancing medical staff and patient satisfaction.

On the other hand, a number of ethical concerns are debated, for instance regarding data protection and possible negative effects on doctor–patient relationships. A key issue in ethical debates is the opaqueness of many AI, especially in ML-based systems, and the resulting lack of transparency of their internal mechanisms and hence the lack of traceability of their outputs: should ML-based systems be explainable, i.e., be designed to process information transparently and, as a consequence, allow doctors and/or patients to understand the realization of outputs? Is the requirement of explainability necessary for giving informed consent to medical treatment or for trusting doctors and the treatment options offered by ML systems? What kind of transparency or explainability and to what degree should be required to establish appropriate informed consent or trust in the field of medicine? To what extent do ML systems increase or decrease personal autonomy in medical contexts? How should we evaluate the requirements for transparency and explainability in comparison with the requirements we place on doctors as medical experts? Furthermore, there are questions of justice and discrimination related to the use of ML systems, as they have been shown to incorporate various bias effects that can systematically discriminate certain patient groups. Finally, from a legal perspective, it has not yet been clarified whether and how errors in ML systems can lead to civil liability and consequences under criminal law.

This special section takes up these questions, digs deeper into them, and discusses further relevant ethical implications regarding the use of ML-based systems in medical contexts. The collection is not intended to call the progress of AI and its positive effects into question, but rather to help clarify important conceptual ambiguities as well as to discuss some of the most relevant normative and social implications of the medical use of ML-based systems. With this, we hope to contribute to the conceptual and ethical reflection of a rapidly developing field.

In *Hammer or Measuring Tape? Artificial Intelligence Tools and Justice in Healthcare*, Jan-Hendrik Heinrichs contends that the principle of justice in healthcare must be revised due to the influence of AI on decisionmaking processes. The article claims that current procedural approaches to justice lack sensitivity in considering how seemingly fair principles can have varying effects on different groups of individuals. Consequently, according to Heinrichs, enhancing the principle of justice in healthcare necessitates a combination of procedures for selecting equitable criteria and principles, alongside leveraging algorithmic tools for accurately measuring their actual impact.

The second article of the section, *The Virtues of Interpretable Medical AI,* by Joshua Hatherley, Robert Sparrow, and Mark Howard discusses explainable AI as a potential response to "black box" AI systems. They argue that it would be unjustified to prioritize accurate AI systems over interpretable ones, since such prioritizing could prove detrimental to patient well-being. Clinicians might prefer interpretable systems over more accurate black boxes, which would be a reason to favor

interpretability over accuracy. Furthermore, the authors posit that the utilization of advantages offered by AI would depend on how doctors and patients decipher its results. Therefore, the focus on accuracy in medical AI could have negative consequences in terms of patient health if its interpretability is disregarded.

Next, in *Learning to Live with Strange Error: Beyond Trustworthiness in AI,* Charles Rathkopf and Bert Heinrichs take a critical look at various concepts of trustworthy AI. The authors' primary statement is that they unjustifiably rely on anthropocentric assumptions when describing trustworthiness. On this basis, they aim to show that replacing "trustworthy" with "confers epistemic justification" is not appropriate, as the justification of AI judgments differs substantially from the justification of human expert judgments. Indeed, reliability in AI models would show a differing epistemic weight and be associated with "strange errors" related to ML classification problems. If such strange errors exist, the use of ML could lead to unforeseeable risks and associated harms for which no one can be blamed or held responsible. For such harms, the authors call for systems of compensation that do not depend on identifying blameworthy parties.

Catrin Misselhorn's contribution, *Machine Ethics in Care: Could a Moral Avatar Enhance the Autonomy of Care-Dependent Persons?,* discusses the increasing development of AI systems as a possible solution to the shortage of caregivers. This development would be supported by arguments stating that such systems could contribute to increasing the autonomy of users by increasing the scope of action for people in need of care. Misselhorn investigates how an artificial moral agent could be designed as a moral avatar of its user that adapts to the user's value profiles with the aim of increasing user autonomy.

In *Misplaced Trust and Ill-placed Distrust: How not to Engage with Medical AI*, Georg Starke and Marcello Ienca discuss demands for so-called trustworthy AI. On the one hand, the authors point out the difficulty in talking about trust in contexts of AI usage, as this would represent a mistake in categories regarding nonhuman agents. Furthermore, there would be fears that the concept of trust could be misused for ethics washing. Against this background, the authors consider cases in which the concepts trust and mistrust are misplaced. They systematize these cases and offer a taxonomy of misplaced trust and distrust.

Andreas Wolkenstein's article, *Healthy mistrust. Medical Black Box Algorithms, Epistemic Authority and Preemptionism,* addresses the question of whether and in what sense medical "black box algorithms" (BBAs) can be understood as epistemic authorities (EAs) and whether their results completely supersede patients' existing convictions. Wolkenstein denies the latter in a critique of so-called preemptionism and discusses some requirements for dealing with BBAs as EAs.

Finally, Pim Haselager et al., in *Reflection Machines: Supporting Effective Human Oversight over Medical Decision Support Systems*, address the problem that an increasing use of medical decision support systems (MDSSs) may lead to people relying more and more in an uncritical manner on machines to make decisions. They propose the development of "reflection machines" (RMs), which could be used to improve human control in decisionmaking situations. Such RMs would be less of a supplier of decisions and more of a support on the path to decisions by encouraging reflection through questions. A RM could, for example, point out possible counterarguments in relation to a possible solution and thus uncover unreflected aspects of automated decisionmaking processes. The article refers to a RM prototype to show what an application could look like.

The contributions collected here show that ML systems used in medicine have to take into account not only their expected benefits, but also a wide range of ethical challenges. Many of them concern the epistemic level and not only raise the question of to what extent and how practitioners and patients need to understand ML processes and outcomes, but also open up a more fundamental question about the difference in the epistemic status of human versus machine judgments in doctor–patient relationships. New ML-based technological possibilities also force us to think again about questions of autonomy, trust, and justice in human–technology relationships in medicine and in general. Since such technological developments are very fast, it is important to address such questions from an ethical perspective in a rapid but well-founded manner. The manuscripts collected here are intended to contribute to this endeavor.

Acknowledgments. The work on this special section was funded by the German Research Foundation (DFG), project number 418201802. We would like to thank all the authors for their contributions and Tomi Kushner for her endless patience and continued support during the production of this special section.

Competing interest. The authors declare having no competing interest.

Cite this article: Friedrich O and Schleidgen S (2024). Editorial: The Ethical Implications of Using AI in Medicine. *Cambridge Quarterly of Healthcare Ethics* 33: 307–309, doi:10.1017/S0963180123000671

Alexandra Exter, Robot, 1926, Cardboard, fabric, wood, glass, and string. Location: The Art Institute of Chicago/Chicago/USA, Photo Credit: The Art Institute of Chicago/Art Resource, NY Reproduced by Permission.