Response to our reviewers

Durán, Juan Manuel; Jongsma, Karin Rolanda

DOI
10.1136/medethics-2021-107531

Publication date
2021

Document Version
Accepted author manuscript

Published in
Journal of Medical Ethics

Citation (APA)
Durán, J. M., & Jongsma, K. R. (2021). Response to our reviewers. *Journal of Medical Ethics*, *47*(7), 514-514. [107531]. https://doi.org/10.1136/medethics-2021-107531

Response to our reviewers

Juan Manuel Durán, Delft University of Technology

Karin Rolanda Jongsma, University Medical Center Utrecht

We would like to thank the authors of the commentaries for their critical appraisal of our feature article Who is afraid of black box algorithms? Their comments, suggestions and concerns are various, and we are glad that our article contributes to the academic debate about the ethical and epistemic conditions for medical XAI.

We would like to bring to attention a few issues that are common worries across reviewers. Most prominently are the merits of computational reliabilism (CR) -- in particular, when promoted as an alternative to transparency -- and CR as necessary but not sufficient for delivering trust. We finalize our response by addressing concerns about the place and role of AI in medical decision-making and the physician's responsibilities.

We understand the concerns and reservations that some of the reviewers express regarding the epistemic merits of CR. We believe that, in part, this is due to a practice too deeply rooted in transparency. But upon closer inspection, transparency presents a series of concerns (epistemological and moral) that should not go unnoticed. CR is advanced as a framework that is capable of delivering trust in the results of medical AI and, as such, an alternative to transparency. To this end, CR picks out indicators relevant for sanctioning the reliability of the algorithm and the trustworthiness of its output. CR, then, delivers different degrees of trust based on the different merits of the reliability indicators that are at play. Thus understood, CR conveys our best epistemic efforts aimed at trusting the algorithm and its outputs, and offers a genuine alternative to transparency. In particular, CR neither requires "opening" the algorithm -- a quite debatable methodological and epistemological move -- nor inherits the limitations of transparency (e.g., the algorithm regress mentioned in our paper). This is of course not to say that CR is free of concerns. In particular, as admitted in our article, we had no pretense to exhaust all possible reliability indicators for medical AI, nor to entrench a hierarchy among them. This is an important point, since some of the reviewers seem to believe that CR is a fixed, one-solution-fits-all approach to trust. Contrary to this belief, CR admits additions and order in the reliability indicators. For instance, taking stock of Grote's comment, one could consider uncertainty qualification (UQ) as a reliability indicator external to a medical AI and capable of contributing to the overall reliability of said system. As briefly described by Grote, UQ by itself cannot grant the desired trust on the results of medical AI. Rather, it is in combination with other indicators that it conveys the kind of information that entrenches a medical AI as a reliable autonomous decision-making system.

As Grote and Veliz et al. point out, the place of black box algorithms in the clinical workflow and the weight it should have in determining treatment decisions is still an open question. While empirical evidence in this respect is scarce, as integration of medical AI in clinical care is still in its infancy, we agree that the ways in which physicians use and rely on such systems is a question that deserves further ethical and epistemological scrutiny. In this respect, we believe that medical AI should be understood as socio-technical systems. Indeed, for ethical medical AI not only the design of the technology and its opaqueness matter, but the competence of physicians inusing these systems responsibly and their skills in recognizing harm and risk are similarly important. As we outline elsewhere in more detail (Sand, Durán, Jongsma, 2021), physicians' responsibility should be understood not only as being accountable for any mistakes (backward looking responsibility, after harm has occurred), but should also include the prevention of harm (forward looking responsibility), as Lang also seems to suggest in his commentary. This notion of responsibility also has implications for the training of physicians. Physicians should, as a minimum, be able to understand and critically assess whether the AI system's outputs are reasonable given a certain diagnostic procedure and be able to understand the input data and its quality. We then agree with Lang that it would be "over-demanding to expect physicians to double as programmers or computer engineers." Aside from knowledge about the system, awareness of one's own experience and skill decline is also important for physicians to use these systems in a way that can improve the diagnostic process or clinical decision-making and prevent overreliance on such systems.

References:

Sand, M., Durán, JM, and Jongsma, KR (2021) "Responsibility beyond design: physician's requirements for ethical medical AI. Bioethics (in press)