

The ethics of machine learning in healthcare: Is exceptionalism required after all?

Char and colleagues (2020) describe an interesting and useful approach in their paper, “Identifying ethical considerations for machine learning healthcare applications.” Their proposed framework, which seeks to identify and address relevant ethical concerns along the pipeline of conception, development, implementation, evaluation, and oversight of machine learning healthcare applications (ML-HCA), can support relevant stakeholders in “put[ting] the problem well” when appraising the ethical issues accompanying this new technology. However, the framework (as any suggested, ethical framework) calls for critical scrutiny.

The main question we raise in this commentary is whether the structural and epistemic underpinning of this framework is robust enough to capture the ethically relevant aspects of ML-HCAs. We question the assumption that ML-HCAs should not be expected to require an exceptional ethical approach compared to other health technologies. Instead, we argue that the distinctions between “locked” and “continuously learning” and between “assistive” and “autonomous” ML-HCAs, as emphasized by the authors, suggest that the continuously learning and autonomous ML-HCAs call for exceptionalism regarding the *identification of ethical concerns* in a way locked and assistive MC-HCAs do not. We also question whether the authors’ strategy “to put the problem well” based only on taking one step back and assessing a broad range of ethical concerns, as described in the pipeline framework, constitutes an adequate methodological approach toward the trustworthy ethical assessment of ML-HCAs. To properly address the problem, we suggest taking at least three steps back to safeguard the appropriateness and usefulness of an ethical framework, such as that proposed by Char et al. (2020).

Exceptional technologies-exceptional ethics

Contrary to Char et al.'s (2020) claim that a framework “does not need to be focused on exceptions, even as it should leave space for exceptional considerations to be identified” (p7) we believe integrating an intentional search for exceptionalism is required for an ethical framework tasked with assessing this new technology. Identifying ethical concerns shared between different technologies is important, but investigating potential differences can logically settle the question regarding ethical exceptionalism. Thus, researchers must proactively seek ethically relevant properties between old and new technologies before drawing a valid conclusion on whether or not a new technology should be assessed using an already established normative framework. If this is not done carefully, unique ethical issues may slip through the net. We support this view by highlighting the foreseeable disruptive and transformative effects of continuous learning and autonomous ML-HCAs on society and on human interactions and conditions.

The signifying characteristics of continuously learning and autonomous ML-HCAs is that they “automatically update using inputs during use, as opposed to locked ML-HCAs, which are deterministic” and are able to “provide direct diagnosis and management statements without any clinical or any other human interpretation or supervision” (Char et al. 2020, pp11–12). Conversational agents (CAs) used in clinical settings is an example of the type of ML-HCAs that can have transformative effects, which the framework of Char et al. must be able to adequately address. According to McGreevey, Hanson, and Koppel (2020, p552), CAs “...are artificial intelligence (AI) programs that engage in a dialogue with users by interpreting their questions or concerns and replying to them in a text message, image, or voice format...They typically imitate human conversation by applying natural language processing and machine learning and stand in contrast to text-based engagement platforms that accept discretely formatted human inputs and reply with preset messages.” CAs

autonomously use continuous learning based on inputs from the conversation and work without the involvement of a human operator.

A specific characteristic of these CAs is that their aim is not only to assist clinicians with, for example, identifying patients at risk of self-harm, but to replace them altogether. At this point, CAs *become* the healthcare providers, and this is precisely where the disruptive and transformative nature of this technology lies. Working alongside their human “colleagues,” broadly recognized concerns are that CAs will not be able to explain their reasoning regarding their care management decisions the way that humans can, nor will it be possible to hold them accountable on the same premises of designated responsibility as is the case with their human counterparts. Such a lack of transparency and accountability pathways might challenge overall trust in the healthcare system (although not necessarily in the CA). The opaqueness in the way decisions are made might lead people to doubt the values underpinning these decisions without giving them the opportunity to openly question and challenge them (Kerasidou 2020).

Alternatively, based on their potential ability to perform better than humans according to specific endpoints, CAs may lead to more trust in, or more accurately, more reliance on healthcare systems (Kerasidou 2021, forthcoming), but at the cost of progressively devaluing investments in healthcare personnel to develop the relevant competencies (Kerasidou 2020). Prioritizing the technological performance of CAs that supersedes human abilities may also impact the fundamental conditions that foster the practice of ethics by healthcare personnel. If autonomous ML-HACs increasingly replace humans, professionals’ experiences with patients’ social, emotional, and existential challenges may decrease. Human contact is a basis for compassion, empathy, solidarity, and recognition of injustice. These are all important

elements in the “social fabric” among humans that influences motivations, actions, practical ethics, and political ideology and prompt people to raise their voice on behalf of those who cannot do so (Bærøe and Gundersen 2019). The trustworthy application of machine learning in healthcare assumes human-based, ethical control over the technology system, but transformative ML-HCAs threaten to undermine this crucial control (Bærøe, Miyata-Sturm and Henden 2020). This type of concern can ultimately lead to the “exceptional” situation that renders reliance on ethical assessments based on human abilities useless.

The aforementioned concerns might be perceived as only speculative given the current state of ML-HCAs. Undoubtedly, technology's development and future applications remain uncertain. Nevertheless, these are the kind of foreseeable and significant societal consequences any systematic ethical framework trying to assess the social impact of new technologies should capture and address. Researchers’ and practitioners’ understanding of ethical challenges related to “old” technologies have evolved over the years through people’s experiences, reflections, and negotiations over how to address them. Thus, new technologies with disruptive and transformative societal consequences should not be interpreted in light of historical “ethical schemes” by default. Rather, they call for pro-active approaches to screen for exceptional concerns. Unless an intentional search for exceptionalism is integrated, there is less reason to trust that those new problems are accurately identified and described. Moreover, these consequences must be deliberated and addressed within ethical and political frames and structures that fit the purpose and receive wide societal acceptance.

Taking not one, but three steps back to put an ethical problem well

This leads us to the next point we want to convey in this commentary. We question whether the authors’ strategy “to put the problem well,” based on only taking *one step* back

and providing a broad range of ethical concerns described in the pipeline framework, constitutes a sufficient methodological approach toward the ethical assessment of ML-HCAs. People can be expected to disagree on how decisions about socially important issues should be made (Daniels and Sabin 2002). This is likely to be true about what consequences diverse stakeholders would consider when assessing ML-HCAs, including the value attached to their potential transformative impact described above (Bærøe and Gundersen 2019). Less disagreement can be expected regarding the conditions for a fair decision-making process (Daniels and Sabin 2002). The collaboration of stakeholders can support the identification, interpretation, and contextualization of generalized ethical concerns. Nevertheless, to safeguard the ethics of an ethical framework, such as that proposed by Char et al. (2020) “to put a problem well,” requires a *second step* back to also provide justification of *who* should contribute to identifying the ethical concerns in the first place, *how* they do so, *when* in the different phases of the pipeline should they do it, and *where* in the decision-making process their impact can be traced (Bærøe 2014). An ethical framework for assessing health technologies is not based on fair conditions, unless it also involves easily understandable justification of how inputs from diverse stakeholders should be allowed to broadly impact the people's understanding of the problem. In turn, the justifications of the “who,” “what,” “when,” and “where” can be scrutinized theoretically as coherently promoting fairness by taking a *third step* back. In principle, this opens for an infinite regression of steps of justification. However, for practical purposes, it should involve at least three steps to protect against the arbitrary justification of the use of power.

Further development and implementation of the proposed framework should keep distinct versions of ML systems apart and carefully address the exceptional conditions related to continuously learning and autonomous ML-HCAs. Finally, further work on the framework

and the proposed *three steps* is necessary in ensuring the ethical and trustworthy development and applications of this technology. To make such advances, drawing on work explicitly focusing on stakeholders involvement in the field of priority setting in health can be helpful (e.g., Abelson et al. 2016; Jansen, Baltussen, and Bærøe 2018).

References

- Abelson, J., F. Wagner, D. DeJean, S. Boesveld, F.-P. Gauvin, S. Bean, R. Axler, S. Petersen, S. Baidooobonso, G. Pron, M. Giacomini, and J. Lavis. 2016. Public and patient involvement in health technology assessment: A framework for action. *International Journal of Technology Assessment in Health Care* 32 (4): 256–264. doi: 10.1017/S0266462316000362.
- Bærøe, K. 2014. Translational ethics: an analytical framework of translational movements between theory and practice and a sketch of a comprehensive approach. *BMC Medical Ethics* 15 (1): 71.
- Bærøe, K., and T. Gundersen. 2019. Social Impact Under Severe Uncertainty: The role of neuroethicists at the intersection of neuroscience, AI, ethics, and policymaking. *AJOB Neuroscience* 10 (3): 117–119. doi: 10.1080/21507740.2019.1632965.
- Bærøe, K., A. Miyata-Sturm, and E. Henden. 2020. How to achieve trustworthy artificial intelligence for health. *World Health Organization. Bulletin of the World Health Organization* 98 (4): 257–262.
- Daniels, N., and J. Sabin. 2002. *Setting limits fairly: can we learn to share medical resources?* Oxford: Oxford University Press.
- Jansen, M.P.M., R. Baltussen, and K. Bærøe. 2018. Stakeholder participation for legitimate priority setting: a checklist. *International Journal of Health Policy and Management* 7 (11): 973.
- Kerasidou, A. 2020. Artificial intelligence and the ongoing need for empathy, compassion and trust in healthcare. *Bulletin of the World Health Organization* 98: 245–250. doi: <http://dx.doi.org/10.2471/BLT.19.237198>.

Kerasidou, A. 2021 (forthcoming). Trusting institutions in the context of global health research collaboraitons. In *Cambridge Handbook of Health Research Regulation*, edited by Graem Laurie and Agomoni Mitra. Cambridge: Cambridge University Press.

McGreevey, J.D., III, C. William Hanson, III, and R. Koppel. 2020. Clinical, legal, and ethical aspects of artificial intelligence-assisted conversational agents in health care. *JAMA* 324 (6): 552–553. doi: 10.1001/jama.2020.2724.