

Taylor & Francis
Taylor & Francis Group

The American Journal of Bioethics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uajb20

Reasons in the Loop: The Role of Large Language Models in Medical Co-Reasoning

Sebastian Porsdam Mann, Brian D. Earp, Peng Liu & Julian Savulescu

To cite this article: Sebastian Porsdam Mann, Brian D. Earp, Peng Liu & Julian Savulescu (2024) Reasons in the Loop: The Role of Large Language Models in Medical Co-Reasoning, The American Journal of Bioethics, 24:9, 105-107, DOI: 10.1080/15265161.2024.2383121

To link to this article: https://doi.org/10.1080/15265161.2024.2383121

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.

Published online: 03 Sep 2024.

Submit your article to this journal

Article views: 121

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uajb20

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Eleanor Gilmore-Szott http://orcid.org/0009-0001-9426-3383
Ryan Dougherty http://orcid.org/0000-0002-7949-1214

REFERENCES

Akgun, A. E., G. S. Lynn, and J. C. Byrne. 2003. Organizational learning: A socio-cognitive framework. *Human Relations* 56 (7):839–868. doi:10.1177/00187267030567004.

Blumenthal-Barby, J., B. Lang, N. Dorfman, H. Kaplan, W. B. Hooper, and K. Kostick-Quenet. 2022. Research on the clinical translation of health care machine learning: Ethicists experiences on lessons learned. *The American Journal of Bioethics* 22 (5):1–3. doi:10.1080/15265161.2022.2059199.

Burrell, J. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3 (1):205395171562251. doi:10.1177/2053951715622512.

Dougherty, R. J., and J. J. Fins. 2024. Toward a social bioethics through interpretivism: A framework for healthcare ethics. *Cambridge Quarterly of Healthcare Ethics: CQ: The International Journal of Healthcare Ethics Committees* 33 (1):6–16. doi:10.1017/S0963180123000452.

Gilmore-Szott, E. 2023. Epistemic value of digital simulacra. *The The American Journal of Bioethics* 23 (9):63–6. doi:10.1080/15265161.2023.2237448.

Frith, C. D., and U. Frith. 2007. Social cognition in humans. *Current Biology: CB* 17 (16):R724–R732. doi:10.1016/j.cub.2007.05.068.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (5):206–15. doi:10.1038/s42256-019-0048-x.

Salloch, S., and A. Ericksen. 2024. What are humans in the loop doing? Co-Reasoning and practical judgment when using machine learning-driven decision aids. *The American Journal of Bioethics* 24 (9):67–78. doi:10.1080/15265161.2024.2353800.

Sullivan, E. 2022. Understanding from machine learning models. *The British Journal for the Philosophy of Science* 73 (1):109–33. doi:10.1093/bjps/axz035.

Zednik, C. 2021. Solving the black box problem: A normative framework for explainable artificial intelligence. *Philosophy & Technology* 34 (2):265–88. doi:10.1007/s13347-019-00382-7.

THE AMERICAN JOURNAL OF BIOETHICS
2024, VOL. 24, NO. 9, 105–107
https://doi.org/10.1080/15265161.2024.2383121

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

OPEN ACCESS Check for updates

Reasons in the Loop: The Role of Large Language Models in Medical Co-Reasoning

Sebastian Porsdam Mann[a,b], Brian D. Earp[b,c], Peng Liu[d] and Julian Savulescu[b,c]

[a]University of Copenhagen; [b]University of Oxford; [c]National University of Singapore; [d]Zhejiang University

Salloch and Eriksen (2024) present a compelling case for including patients as co-reasoners in medical decision-making involving artificial intelligence (AI). Drawing on O'Neill's neo-Kantian framework (O'Neill 1989), they argue that the "human in the loop" concept should extend beyond physicians to encompass patients as active participants in the reasoning process. While we commend this perspective, we suggest it can be extended further: certain types of medical AI, particularly large language models (LLMs), should also be considered as potential co-reasoners.

Salloch and Eriksen state that doctors should encourage "patients to take active part in a process of reasoning," and that "doctors need to articulate reasons to support their reliance on automated decisions." This active participation, they argue, "is not primarily a matter of registering desires or preferences, but of being *responsive to valid considerations* concerning what the situation demands" (emphasis in original). This framing elevates the patient's role from a mere

CONTACT Brian D. Earp brian.earp@philosophy.ox.ac.uk Uehiro Centre for Practical Ethics, Faculty of Philosophy, University of Oxford, 16-17 St. Ebbes Street, Oxford, OX1 1PT, UK; Centre for Biomedical Ethics, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, 117597, Singapore.
© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

source of preferences to an active participant in the reasoning process, helping to address issues such as automation bias (Khera, Simon, and Ross 2023).

Importantly, this approach identifies responsiveness to valid considerations—that is, identifying and endorsing reasons—as the key factor in ethical use of medical AI systems. It matters less whether reasoning is initially offered by the doctor or patient than whether both respond to and endorse valid reasons once raised and discussed.

We agree with this and propose to take it further. If what matters is not the source of reasons, but their relevance and the ability of patients and doctors to rationally endorse them, then we propose that "the loop" can be extended to include certain types of medical AI, particularly LLMs, as providers of reasons. These AI systems, when properly designed and implemented, can generate relevant reasons and considerations that could significantly enrich the co-reasoning process (see Demaree-Cotton, Earp, and Savulescu 2022; Haupt and Marks 2023; Truhn, Reis-Filho, and Kather 2023).

Support for this idea comes from a recent study (Chu et al., under review) where we asked LLM models to consider moral dilemmas, make judgments, and provide reasons for those judgments. The reasons generated were predominantly—though not always—relevant and sensible, allowing for evaluation and further discussion by human experimenters. This demonstrates that LLMs can provide "inputs" for deliberation—not final answers—which humans can then consider in reaching their own conclusions.

In the medical context, we are concerned with what a person (doctor, patient, administrator) should do. This is a call for reasons. A human expert, or an AI, might be 95% likely to get the right answer to an empirical (e.g., diagnostic) problem, based on past evidence. This is important to know as one reason to believe it will get the right answer to similar types of problems in the future. It is one justification. Explanation is another reason to believe it will get the right answer—if there is a plausible explanation for how these predictions arose, then that increases confidence. But while a human expert or AI system might have high predictive accuracy based on past evidence, and explanations can increase confidence in predictions, these do not exhaust the set of reasons.

Empirical reasons can have a variety of sources, including evidence, causal explanation, but also other reasons based on argument as well as evidence. Moreover, normative reasons require values, and there can be a multitude of values relevant to any decision. LLMs can provide candidates of both empirical and normative reasons which can then be evaluated by humans.

How might this apply to a medical context? Our experiments involved the use of "standard," off-the-shelf LLMs, which were trained on large volumes of general data (i.e., without being specialized for any particular domain), and still they were able to provide reasons for—that is, coherent, logically relevant justifications of—the responses given. But LLMs in medicine tend to be fine-tuned on, or given access to, specific datasets that are likely to be even more pertinent to a given decision-making context (Allen et al. 2024; Zohny et al. 2024).

Suppose that, upon appropriate reflection, and with all necessary safeguards in place to ensure that the LLMs' "input" is not unduly privileged over other perspectives or sources of information, the reasons provided by an LLM in the context of some decision-making process are ultimately *endorsed* by both doctors and patients because they are rationally compelling.

If so, we argue that it should not matter (in and of itself) that the reasons were generated by an LLM, in the first instance, rather than by a doctor or patient. What matters is that the humans in the loop can reflectively endorse, or reject or override, the reasons provided by the LLM in reaching an all-things-considered conclusion.

One might object that LLMs merely generate statistically likely responses without genuine understanding (Bender et al. 2021), or that they lack access to relevant feelings and contextual nuances because they are unlikely to be conscious (Chalmers 2023). We acknowledge that, on most plausible current theories, LLMs do not have *subjective* or *experiential* access to such phenomena (although they may still encode relevant functional cues to such phenomena) (Chemero 2023). However, we argue that this is ultimately irrelevant to the potential value of LLM-generated reasons in medical co-reasoning. What matters is the quality and relevance of those reasons as judged by the humans involved (cf. Maslej et al. 2023).

For example, we have previously proposed the use of personalized LLMs (as in Porsdam Mann et al. 2023) to infer what a formerly competent but currently incapacitated patient would want (a "Personalized Patient Preference Predictor" or P4; Earp et al. 2024). Such predictors could turn out to be more reliable (i.e., statistically accurate) than the patient's family. However, what also matters is their capacity to offer reasons for their prediction. For example, a P4 might offer the following reason for why a patient would not wish to be kept alive when unable to communicate with family: "On June 6, 1990, John wrote an email to Jack expressing his sadness that Jack's father could not communicate after a stroke, and said he would never want to be kept alive in a state like that." That is a relevant reason.

It is not decisive, but it ought to be evaluated in the context of all that is known about John.

We propose that the same standard should apply to reasons generated by medical AI as to reasons provided by humans. If an LLM generates a consideration that both the doctor and patient find valid and relevant after discussion, it should be incorporated into the decision-making process, regardless of its origin. This approach maintains the centrality of human judgment—the doctor and patient remain the ultimate arbiters of which reasons are valid and relevant—while leveraging the unique capabilities of AI to enrich the co-reasoning process.

In other words, we should move from "human in the loop" to "reasons in the loop."

AI USE ACKNOWLEDGEMENT

The LLM Claude 3 was used to shorten and edit text supplied by the authors. All ideas and arguments were provided by the authors.

DISCLOSURE STATEMENT

JS is a Bioethics Committee consultant for Bayer. JS is an Advisory Panel member for the Hevolution Foundation.

FUNDING

SPM's research for this paper was supported by a Novo Nordisk Foundation Grant for a scientifically independent International Collaborative Bioscience Innovation & Law Programme (Inter-CeBIL programme – grant no. NNF23SA0087056). JS is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-GV-2023-012). This work was supported by the Wellcome Trust [Grant number: 226801] for Discovery Research Platform for Transformative Inclusivity in Ethics and Humanities Research (ANTITHESES).

ORCID

Sebastian Porsdam Mann http://orcid.org/0000-0002-1867-2097
Brian D. Earp http://orcid.org/0000-0001-9691-2888
Peng Liu http://orcid.org/0000-0003-4929-0531
Julian Savulescu http://orcid.org/0000-0003-1691-6403

REFERENCES

Allen, J. W., B. D. Earp, J. Koplin, and D. Wilkinson. 2024. Consent-GPT: Is it ethical to delegate procedural consent to conversational AI? *Journal of Medical Ethics* 50 (2):77–83. doi:10.1136/jme-2023-109347.

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of Stochastic Parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* 610–623. Virtual Event, Canada.

Chalmers, D. J. 2023. Could a large language model be conscious? *arXiv*. https://doi.org/10.48550/arXiv.2303.07103.

Chemero, A. 2023. LLMs differ from human cognition because they are not embodied. *Nature Human Behaviour* 7 (11):1828–9. doi:10.1038/s41562-023-01723-5.

Chu, Y., P. Liu, J. Savulescu, and B. D. Earp. (under review). Road rage against the machine: Humans and GPT-4 share a blame bias against driverless cars. Manuscript under review.

Demaree-Cotton, J., B. D. Earp, and J. Savulescu. 2022. How to use AI ethically for ethical decision-making. *The American Journal of Bioethics* 22 (7):1–3. doi:10.1080/15265161.2022.2075968.

Earp, B. D., S. Porsdam Mann, J. Allen, S. Salloch, V. Suren, K. Jongsma, M. Braun, D. Wilkinson, W. Sinnott-Armstrong, A. Rid, et al. 2024. A personalized patient preference predictor for substituted judgments in healthcare: Technically feasible and ethically desirable. *The American Journal of Bioethics* 24 (7):13–26. 1–14. doi:10.1080/15265161.2023.2296402.

Haupt, C. E., and M. Marks. 2023. AI-generated medical advice—GPT and beyond. *JAMA* 329 (16):1349–50. doi:10.1001/jama.2023.5321.

Khera, R., M. A. Simon, and J. S. Ross. 2023. Automation bias and assistive AI: Risk of harm from AI-driven clinical decision support. *JAMA* 330 (23):2255–7. doi:10.1001/jama.2023.22557.

Maslej, M. M., S. Kloiber, M. Ghassemi, J. Yu, and S. L. Hill. 2023. Out with AI, in with the psychiatrist: A preference for human-derived clinical decision support in depression care. *Translational Psychiatry* 13 (1):210. doi:10.1038/s41398-023-02509-z.

O'Neill, O. 1989. *Constructions of reason: Explorations of Kant's practical philosophy*. Cambridge: Cambridge University Press.

Porsdam Mann, S., B. D. Earp, N. Møller, S. Vynn, and J. Savulescu. 2023. AUTOGEN: A personalized large language model for academic enhancement—Ethics and proof of principle. *The American Journal of Bioethics* 23 (10):28–41. doi:10.1080/15265161.2023.2233356.

Salloch, S., and A. Eriksen. 2024. What are humans doing in the loop? Co-reasoning and practical judgment when using machine learning-driven decision aids. *The American Journal of Bioethics* 24 (9):67–78. doi:10.1080/15265161.2024.2353800.

Truhn, D., J. S. Reis-Filho, and J. N. Kather. 2023. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine* 29 (12):2983–4. doi:10.1038/s41591-023-02594-z.

Zohny, H., S. Porsdam Mann, B. D. Earp, and J. McMillan. 2024. Generative AI and medical ethics: The state of play. *Journal of Medical Ethics* 50 (2):75–6. doi:10.1136/jme-2023-109834.