The American Journal of Bioethics

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

Ethical Algorithmic Advice: Some Reasons to Pause and Think Twice

Torbjørn Gundersen & Kristine Bærøe

To cite this article: Torbjørn Gundersen & Kristine Bærøe (2022) Ethical Algorithmic Advice: Some Reasons to Pause and Think Twice, The American Journal of Bioethics, 22:7, 26-28, DOI: 10.1080/15265161.2022.2075053

To link to this article: https://doi.org/10.1080/15265161.2022.2075053

Published online: 23 Jun 2022.

THE AMERICAN JOURNAL OF BIOETHICS
2022, VOL. 22, NO. 7, 26–28
https://doi.org/10.1080/15265161.2022.2075053

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Ethical Algorithmic Advice: Some Reasons to Pause and Think Twice

Torbjørn Gundersen[a] and Kristine Bærøe[b]

[a]Oslo Metropolitan University; [b]University of Bergen

Machine learning and other forms of artificial intelligence (AI) can improve parts of clinical decision making regarding the gathering and analysis of data, the detection of disease, and the provision of treatment recommendations. The target article "Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept" (Meier et al. 2022) explores the less-examined possibility of using this technology to provide ethical advice. The article examines the feasibility of an algorithmic advisory system for clinical ethics called METHAD, which is designed to provide recommendations to clinicians facing difficult ethical questions. METHAD utilizes a form of machine learning model called *fuzzy cognitive maps* and is based on Beauchamp and Childress' four principles of biomedical ethics, namely, beneficence, non-maleficence, autonomy, and justice, and is trained on data from clinical ethics committees. The article provides an illuminating and highly interesting exploration of how ethical principles can be operationalized into an algorithmic model, which clinicians could use as an advisory tool and even defer to for moral judgments, similar to how they might defer to people concerning ethical issues. The authors also display a sensible degree of expert humility on behalf of METHAD and are explicit about the technical and ethical challenges regarding the reliability and acceptability of the recommendations that the algorithm provides. In this commentary, we wish to draw attention to some key challenges an ethical algorithmic advisory system, such as METHAD, encounters pertaining to ethical algorithmic design and the operationalization of ethics.

OPERATIONALIZATION OF ETHICAL PRINCIPLES

When designing METHAD, developers face fundamental problems with ethical algorithmic design and the operationalization of ethics. The authors have chosen the four biomedical principles articulated by Beauchamp and Childress (2013) as the moral framework for the algorithmic model since "it provides a set of decision factors common across case types which lends itself to being translated into machine-readable values" (Meier et al. 2022, 9). As the authors rightly emphasize, there is no agreement on what constitutes a superior ethical theory for clinical ethics. A further problem which they also mention is whether clinical decision-making is, in fact, based on any such ethical theories (Meier et al. 2022). When clinicians make decisions about the treatment of patients, it is far from clear that they base these choices on any explicit ethical theory, such as these principles. Their decision-making might be much more tacit and pragmatic in nature and, one could add, wisely so. Moreover, to ensure that medical decision making is ethically acceptable and trustworthy, clinicians must be able to draw on a set of institutional and legal standards, patient and stakeholder perspectives, and group deliberation to reach individual and collective judgments in weighing concerns and trading off interests. Thus, while one could agree that the four principles of biomedical ethics should be a central part of a moral framework for ethical algorithmic design, it is, like any ethical theory, *immensely insufficient* for "ethical decision making," as the title of the target article formulates it. Consequently, the operationalization of ethical principles by encoding them directly into an algorithm like METHAD has a rather narrow scope of relevant sources for making an ethically acceptable decision compared to health

CONTACT Torbjørn Gundersen torgun@oslomet.no Centre for the Studies of Professions, Oslo Metropolitan University, Oslo 0130, Norway.

professionals. Finally, when encoding ethical principles directly into an algorithm, there is the risk that essential ethical deliberation and collaboration between AI developers, professionals, patients, and other stakeholders will be short-circuited (Gundersen and Bærøe 2022).

These challenges raise several questions. What are the consequences for the development of clinical ethics decisions in changing contexts (as for example medical technology develop or the pool of resources change), if the advice of an AI system is based on a too narrow ethical basis? Where and how will we place responsibility for the ethical decisions? When the AI system aims to recommend decisions that otherwise are carried out in committees (as clinicians have requested assistance), would the added normative source of perceived legitimacy not be lost regarding the collective efforts of reaching a conclusion through a fair process?

WHAT IS THE RATIONALE OF ETHICAL ADVICE MADE BY AI?

In our view, the authors do not elaborate much on the main purpose of developing an algorithmic advisory system, such as METHAD, for clinical ethics. To approach this issue, we find it useful to distinguish between two main modes of output from AI, which generate various ethical and epistemological challenges, namely, informing and recommending output (Gundersen 2018). In the informing mode, AI *provides evidence* for decision making, that is, by providing analyses, probability estimates, or predictions from vast datasets that human decision-makers can use when deciding. When human decision-making is assisted by AI-generated information, several well-known epistemological and ethical pitfalls pertain to the lack of transparency, explainability, and accountability of AI (Grote and Berens 2020). In the recommending mode, on the other hand, AI is used to directly influence what people should do, such as whether doctors should give a patient a certain treatment. METHAD clearly belongs to the recommending mode of output and surely provides a particular kind of recommendation, namely, ethical recommendations about what clinicians should do according to the central principles of bioethics.

Now, in cases where human experts provide recommendations to laypeople on technical policy issues, there are several ways in which they could potentially be valuable. By coupling advanced knowledge with political and ethical premises, scientific experts and medical doctors may describe and rank the available options from which policymakers and patients might choose. Such recommendations could be used in an *advisory manner* by providing policy alternatives from which laypeople could learn from, build on, and modify in ways that empower their decision-making capabilities and enable public deliberation. In some cases, where people lack a required understanding of the available knowledge about the issue at hand or are uncertain about their own ethical judgment, they could in some cases even reasonably *defer* to experts and follow the recommendation if they have reason to think that this will improve their decisions (for instance by generating more welfare or reduce harm to others, see Enoch (2014) for a defense of moral deference). Indeed, the target article does indicate that METHAD could play an advisory role when mentioning that it could be used as a part of education (Meier et al. 2022) and as a potential source of moral deference in emergency situations in which it might replace clinicians (Meier et al. 2022).

However, while the algorithmic model might be used for educational purposes, it is unclear what advisory role an algorithmic model could play and how clinicians might learn from it. One possible route, which the authors do not mention, could be that the algorithmic model provides a set of plural and conditional recommendations based on different interpretations of existing values and expands the available lines of action for clinicians. Concerning moral deference, in its current form, in which METHAD provides problematic recommendations due to the downplaying of patient autonomy, it seems that the prospect of rationally deferring to METHAD is unrealistic.

A THOUGHT EXPERIMENT: THE PARADOX OF SUCCESSFUL REPLACEMENT

METHAD is still in its pilot phase, and the authors display a commendable degree of humility on behalf of METHAD. However, let us assume that some of the main technical and ethical difficulties mentioned earlier in designing METHAD are defeated. Let us assume, for the sake of argument, that METHAD provides ethical recommendations faster, cheaper, in line with central principles of medical ethics and medical ethicists' justification, and more consistently so than human clinicians. Let us also imagine that this AI system is showing a positive effect on clinical practice and health outcomes in clinical trials by comparing the use of the algorithm with standard practices. Due to this immense success, this ethical advisory algorithm is being applied in healthcare on a large scale, and clinicians have begun to defer to the algorithms as a part of their practice. In other words, let us assume that METHAD, or a similar kind of ethical advisory algorithm, is excellent in all relevant regards and that METHAD replaces the ethical judgment of medical doctors and clinical ethics committees. A likely scenario is that medical doctors who defer to the algorithm will tend to use their own ethical judgment less and spend less time deciding on what one should do. There simply will be less need for this, and they could spend their time on tasks other than ethical reasoning. An unfortunate consequence is that it will likely limit the development of the ethical sensitivity otherwise obtained through engaging with challenging ethical cases (Rest and Narvaez 1994). Correspondingly, their ability to exercise good ethical judgment in an autonomous manner will deteriorate. The success of the algorithm might thus undermine its key source of ethical standards, which it is supposed to reflect. If the ethical sensitivity and judgment of clinicians and members of clinical ethics committees are reduced, the very standards by which the algorithm is assessed diminish. What this means for the future of clinical ethics and the role of clinical ethicists is, indeed, worth further consideration.

FUNDING

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

REFERENCES

Beauchamp, T. L., and J. F. Childress. 2013. *Principles of biomedical ethics*. 7th ed. New York: Oxford University Press.

Enoch, D. 2014. A defense of moral deference. *The Journal of Philosophy* 111 (5):229–58.

Grote, T., and P. Berens. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics* 46 (3):205–11. doi:10.1136/medethics-2019-105586.

Gundersen, T. 2018. Scientists as experts: A distinct role? *Studies in History and Philosophy of Science Part A* 69: 52–9. doi:10.1016/j.shpsa.2018.02.006.

Gundersen, T., and K. Bærøe. 2022. The future ethics of artificial intelligence in medicine: Making sense of collaborative models. *Science and Engineering Ethics* 28 (2): 1–16. doi:10.1007/s11948-022-00369-2.

Meier, L. J., A. Hein, K. Diepold, and A. Buyx. 2022. Algorithms for ethical decision-making in the clinic: A proof of concept. *The American Journal of Bioethics* 22 (7):4–20. doi:10.1080/15265161.2022.2040647.

Rest, J. R., and D. Narvaez. 1994. *Moral development in the professions: Psychology and applied ethics*. Hillsdale, NJ: Lawrence Erlbaum.

THE AMERICAN JOURNAL OF BIOETHICS
2022, VOL. 22, NO. 7, 28–30
https://doi.org/10.1080/15265161.2022.2075050

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Check for updates

Algorithmic Ethics: A Technically Sweet Solution to a Non-Problem

Aurelia Sauerbrei, Nina Hallowell, and Angeliki Kerasidou

University of Oxford, Nuffield Department of Population Health, Ethox Centre

In their proof-of-concept study, Meier et al. (2022) built an algorithm ("METHAD") to aid ethical decision making. In the limitations section of their paper, the authors state a frequently cited axiom; namely, that the fact "that one can do something does not imply that they should." They continue by asking whether developing and ultimately implementing an AI supported ethical decision-making tool in the clinic would be a good idea. For the remainder of the paper, they fail to properly engage with this question. We suggest that, first, the authors fail to convincingly describe the problem that they are trying to solve. Second, they do not explain why an AI driven system is the appropriate tool to aid ethical decision-making. We conclude that the project takes a methodologically flawed approach to problem-solving and is effectively driven by technical sweetness.

According to the EU *Ethics Guidelines for Trustworthy AI* (AI HLEG 2019), one of the things that must be considered when building AI tools is their social value and their impact on "social relationships" and "effect on institutions." Recent guidance by the WHO (2021) also warns against using AI as a first recourse solution to any potential problem. Consideration of the social value and impact of a technology should take place prior to the development

CONTACT Aurelia Sauerbrei aurelia.sauerbrei@ethox.ox.ac.uk Department of Population Health, University of Oxford Nuffield, Oxford OX3 7LF, United Kingdom.