Taylor & Francis
Taylor & Francis Group

The American Journal of Bioethics

ISSN: 1526-5161 (Print) 1536-0075 (Online) Journal homepage: www.tandfonline.com/journals/uajb20

Broadening the Ethical Scope

Margaret Levi, Michael Bernstein & Quinn Waeiss

To cite this article: Margaret Levi, Michael Bernstein & Quinn Waeiss (2022) Broadening the Ethical Scope, The American Journal of Bioethics, 22:5, 26-28, DOI: 10.1080/15265161.2022.2055219

To link to this article: https://doi.org/10.1080/15265161.2022.2055219

THE AMERICAN JOURNAL OF BIOETHICS
2022, VOL. 22, NO. 5, 26–28
https://doi.org/10.1080/15265161.2022.2055219

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Broadening the Ethical Scope

Margaret Levi, Michael Bernstein, and Quinn Waeiss

Stanford University

McCradden and colleagues' (2022) argues that machine learning in health care poses new challenges to appropriate evaluation for safe use in clinical care. It also claims that "the longstanding system of institutional ethics oversight for clinical research can be adapted to health care ML" (11) and offers an approach for doing so. There is a lot to applaud in the proposed process, particularly in stages 2 and 3. Our main concerns focus on stage 1. We would also be interested to learn more about the potential of implementation of the whole, of the reasons the authors believe it will be accepted by universities and firms and what the obstacles might be.

The authors begin by describing problems that AI/ML create when applied in clinical health practice. Some of these have long been recognized in AI, i.e., the introduction of algorithmic biases, particularly in relation to race, gender, and class. Other problems, such as violations of privacy, have also been noted, but the authors are correct to emphasize that personal health data creates a raft of different complexities than consumer or other administrative data. Additionally, there are issues of proprietary techniques and interventions that the medical world has long dealt with but which receive less attention in this piece.

The authors do point out some important lacuna in the system of institutional ethics, particularly as constituted in institutional review boards (IRBs). These make less convincing their claim that the system could be adapted for ML, or at least easily adapted, as they seem to imply. The first lacuna is that only publicly funded research is required to be reviewed, but, as the authors themselves point out, the development of most of the ML applications is privately funded. And this raises serious questions about how and at what point an ethical review is initially required, as we argue below.

The second is the fact that "… many applications of technology in medicine are validated without research ethics oversight" (9). While the authors argue that this is not necessarily a big problem and that it is dealt with through the data governance framework offered in Phase 1, this framework is not yet standard practice. Thus, it is worrisome that a crucial feature of ML development is not subject to scrutiny.

Of equal concern to us is that emphasis on the current methods means an emphasis only on the harms to human subjects. IRBs' rules and regulations in the United States (Department of Health, Education, and Welfare 1979) focus on risks to human subjects, not risks to human society. In contrast, many of the risks embedded in AI research are best understood as risks to society rather than risks to human subjects: for example, when training data or stakeholders do not include communities likely to be impacted by the algorithm's deployment, when the algorithm might be deployed by malicious actors or in ways that the researchers never intended, or when subgroups in society are harmed as through job displacement. From our work, we have identified a number of potential ethical concerns inherent in medical research employing AI. For example, research that uses electronic medical records to develop a model that informs patient treatment options could embed the societal biases inherent in the EMRs within the model itself; the development of a model intended to predict likely

CONTACT Margaret Levi mlevi@stanford.edu Center for Advanced Study in the Behavioral Sciences, Stanford University, 75 Alta Rd, Stanford, CA 94305-8090, USA.

adverse health events in patients, without adequate safeguards to data access, could be used by insurance companies to discriminate in coverage decisions; and some organizations may exert financial or institutional pressure to replace human expert judgment with cheaper algorithmic solutions that were originally intended to augment rather than replace human experts.

Unfortunately, the U.S. Department of Health and Human Services' definition of human subjects research specifically excludes review of risks to society, stating, "The IRB should not consider possible long-range effects of applying knowledge gained in the research […] as among those research risks that fall within the purview of its responsibility" (United States Department of Health and Human Services 2018). As a result, when AI research does not directly involve human subjects, many IRBs decline to review research that embeds these harms. Some efforts, such as the Microsoft Research Ethics Review Program established in 2013, take a more expansive view (Gray, Watts, and Horvitz n.d.)—a view that is unfortunately rare in many research contexts. The second approach of recommended processes such as checklists (e.g., Madaio et al. 2020), volunteer drop-ins, and product reviews (e.g., Gebru et al. 2018; Madaio et al. 2020) all rely on voluntary usage, so while they are valuable, they are limited to those who self-select. A third approach, requirements by academic forums such as NeurIPS (e.g., Abuhamad and Rheault 2020) and the Future of Computing Academy (Gibney 2018) to add ethics content to research paper submissions, are worthwhile but are only enforced at the end of the research process. In contrast, ethics and societal reflections are best considered at the beginning of the research process, before researchers ossify any decisions in stakeholders, models, data, or evaluation strategy. So, there remains a void for an institutional process in research, which can begin the conversation early on in the project lifecycle and engage with all relevant projects rather than just self-selected opt-ins.

To begin earlier in the process requires not only new frameworks and requirements but also, as McCradden et al recognize, close collaboration among ML researchers and clinicians while simultaneously overcoming their cultural differences in research and application. However, unlike other professions such as law and medicine, computing lacks widely applied professional ethical and societal review processes. Progress is being made in many dimensions, including algorithmic advances (e.g., Agarwal et al. 2018), norm changes, improved product design processes (e.g., Gebru et al. 2018; Madaio et al. 2020), and both academic and public activism on these issues.[1] However, ongoing progress requires that everyone involved, not just those who care enough to participate, consider societal impact and ethics in their work. Organizing the behavior of an entire group is the role that institutional structures—rules, incentives, and processes that apply to all—are designed to play. Examples include journal peer review processes, environmental review of new construction, and the rules of evidence and argumentation followed by courts. In today's university environment, there are few institutional structures to facilitate computing and AI researchers in addressing issues of societal and ethical harm. And there are even fewer that bring together those in the private sector with those in the academy.

As one possible approach, we have introduced the Ethics and Society Review Board (ESR), an institutional process that facilitates researchers in mitigating the negative societal and ethical aspects of AI research by serving as a requirement to access funding: grant funding from a large AI institute at our university is not released until the researchers complete the ESR process on the project (Bernstein et al. 2021b). Researchers submit a brief ESR statement alongside their grant proposal that describes their project's most salient risks to society, to subgroups in society, and globally. This statement articulates the principles the researchers will use to mitigate those risks and describes how those principles are instantiated in the research design. For example, researchers building a reinforcement learning AI to support long-term student retention might identify that the AI might learn to focus on the learners who it is most likely to be able to retain rather than those most at risk, then describe how they have brought in a collaborator who studies inclusive educational experiences for marginalized communities and will evaluate the system in part for robust transfer across groups.

The funding program conducts its typical grant merit review, then the ESR begins its process on the grants that are recommended for funding. The ESR convenes an interdisciplinary panel of faculty representing expertise in technology, society, and ethics to review the proposals and ESR statements: our most recent panel included faculty from Anthropology, Communication, Computer Science, History, Management Science & Engineering, Medicine,

[1]For a more detailed review of the progress researchers have made in developing ethical and societal review processes in AI, please see Bernstein et al. (2021a).

Philosophy, Political Science, and Sociology. The ESR considers risks and mitigations in the context of possible benefits to society. Its goal is not to eradicate all possible or actual negative impacts—which is often impossible—but to work with the researchers to identify negative impacts and devise reasonable mitigation strategies. It engages in iterative feedback to the researchers, which can include raising new possible risks, helping identify collaborators or stakeholders, conversations, and brainstorming. When the process is complete, the ESR submits its recommendation to the funding program, and funds are released to the researchers. We present this process as not the only way to structure such a review—we will certainly iterate its design—but as one such approach that has had some success.

What McCradden et al offer is a way to prevent harms to patients by establishing a process for vetting and testing ML applications and, importantly, the standard of equipoise as the basis of a moral justification for human trials. The methods and standards these authors use will influence our own efforts and, hopefully, that of many others. However, we do feel strongly that more attention must also be given to: (1) the very early stages of product development by private actors and industries, as well as university researchers; (2) the incentives for implementation of the approach; and (3) the societal consequences as well as ethical implications of ML innovation in clinical health.

FUNDING

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

ORCID

Margaret Levi http://orcid.org/0000-0001-6672-4552
Michael Bernstein http://orcid.org/0000-0001-8020-9434
Quinn Waeiss http://orcid.org/0000-0002-4880-3043

REFERENCES

Abuhamad, G., and C. Rheault. 2020. Like a researcher stating broader impact for the very first time. *arXiv preprint arXiv:2011.13032*.

Agarwal, A., A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. 2018. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Bernstein, M. S., M. Levi, D. Magnus, B. A. Rajala, D. Satz, and Q. Waeiss. 2021a. ESR: Ethics and society review of artificial intelligence research. *arXiv preprint arXiv: 2106.11521*.

Bernstein, M. S., M. Levi, D. Magnus, B. A. Rajala, D. Satz, and Q. Waeiss. 2021b. Ethics and society review: Ethics reflection as a precondition to research funding. *Proceedings of the National Academy of Sciences* 118 (52): 1–8. doi:10.1073/pnas.2117261118.

Department of Health, Education, and Welfare. 1979. The Belmont report: Ethical principles and guidelines for the protection of human subjects of research. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Gebru, T., J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. Daumé, III, and K. Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*.

Gibney, E. 2018. The ethics of computer science: This researcher has a controversial proposal. *Nature News Q&A*, July 31, 2018 doi:10.1038/d41586-018-05791-w.

Gray, M., D. J. Watts, and E. Horvitz. n.d. Microsoft Research Ethics Review Program & IRB. Accessed September 1, 2021. https://www.microsoft.com/en-us/research/microsoft-research-ethics-review-program-irb/.

Madaio, M. A., L. Stark, J. W. Vaughan, and H. Wallach. 2020. Codesigning checklists to understand organizational challenges and opportunities around fairness in AI. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 1–14.

McCradden, M. D., J. A. Anderson, E. A. Stephenson, E. Drysdale, L. Erdman, A. Goldenberg, and R. Zlotnik Shaul. 2022. A research ethics framework for the clinical translation of healthcare machine learning. *The American Journal of Bioethics* 22(5): 8–12. doi:10.1080/15265161.2021.2013977.

United States Department of Health and Human Services. 2018. Common rule. *Code of Federal Regulations Title 45, §46.111*.