The American Journal of
BIOETHICS
November 2020, Volume 20, Number 11
Identifying Ethical Considerations for Machine Learning Healthcare Applications

The American Journal of Bioethics

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

AI Ethics Is Not a Panacea

Stuart McLennan , Meredith M. Lee , Amelia Fiske & Leo Anthony Celi

To cite this article: Stuart McLennan , Meredith M. Lee , Amelia Fiske & Leo Anthony Celi (2020) AI Ethics Is Not a Panacea, The American Journal of Bioethics, 20:11, 20-22, DOI: 10.1080/15265161.2020.1819470

To link to this article: https://doi.org/10.1080/15265161.2020.1819470

Published online: 26 Oct 2020.

ACKNOWLEDGMENTS

The authors thank John Fahrenbach and Alvin Rajkomar for their review of portions of an earlier version of this manuscript.

DISCLOSURE STATEMENT

Michaela Hardt is employed by Amazon, but her views expressed in this commentary do not represent the views of her employer.

FUNDING

Dr. Chin was supported in part by the Robert Wood Johnson Foundation Advancing Health Equity: Leading Care, Payment, and Systems Transformation Program Office, the Merck Foundation Bridging the Gap: Reducing Disparities in Diabetes Care National Program Office, and the Chicago Center for Diabetes Translation Research [NIDDK P30 DK092949].

REFERENCES

Agarwal, A., A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. 2018. A reductions approach to fair classification. *Proceedings of Machine Learning Research* 80:60–69.

Char, D. S., M. D. Abràmoff, and C. Feudtner. 2020. Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics* 20 (11):7–17. doi: 10.1080/15265161.2020.1819469.

Fawcett, T. 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8):861–874.

Hardt, M., E. Price, and N. Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29:3315–3323.

Kleinberg, J., and S. Mullainathan. 2019. Simplicity creates inequity: Implications for fairness, stereotypes, and interpretability. Proceedings of the 2019 ACM Conference on Economics and Computation, 807–808.

Platt, J. C. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*, ed. A. J. Smola, Peter Bartlett, B. Schölkopf and D. Schuurmans, 61–74. Cambridge, MA: MIT Press.

Rajkomar, A., M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine* 169 (12):866–872.

Woodworth, B., S. Gunasekar, M. I. Ohannessian, and N. Srebro. 2017. Learning non-discriminatory predictors. *Proceedings of Machine Learning Research* 65:1920–1953.

Zhang, B. H., B. Lemoine, and M. Mitchell. 2018. Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335–340. New Orleans, LA, USA.

THE AMERICAN JOURNAL OF BIOETHICS
2020, VOL. 20, NO. 11, 20–22
https://doi.org/10.1080/15265161.2020.1819470

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

AI Ethics Is Not a Panacea

Stuart McLennan[a,b], Meredith M. Lee[c], Amelia Fiske[a], and Leo Anthony Celi[d,e,f]

[a]Technical University of Munich; [b]University of Basel; [c]UC Berkeley Division of Computing, Data Science, and Society; [d]Beth Israel Deaconess Medical Center; [e]Harvard–Massachusetts Institute of Technology; [f]Harvard T.H. Chan School of Public Health

From machine learning (ML) and computer vision to robotics and natural language processing, the application of data science and artificial intelligence (AI) is expected to transform health care (Celi et al. 2019). While the rapid development of technological capabilities offers paths toward new discoveries and large-scale analysis, numerous critical ethical issues have been identified, spanning privacy, data protection, transparency and explainability, responsibility, and bias.

Last year, for instance, a commercial prediction algorithm affecting millions of patients was shown to exhibit significant racial bias, dramatically underestimating the health needs of Black patients (Obermeyer et al. 2019). Trained using health care cost as the proxy for the need for more comprehensive care, the algorithm had been designed specifically to exclude race as a feature, in an attempt to avoid bias—but cost was clearly not a race-neutral measure of health care need. Studies have repeatedly illuminated racial disparities in

CONTACT Leo Anthony Celi lceli@mit.edu Division of Health Sciences and Technology, Harvard–Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA 02139, USA.

the provision of primary care services: Black patients incur approximately US$1800 less in medical costs per year compared to white patients with the same number of chronic conditions, and are less likely to be identified as high-risk for complex care in the future. But even if another proxy, such as probability of death, had been used to train the algorithm, would it have led to a "better" algorithm and fair patient outcomes?

At present, a key evaluation metric for machine learning in health care applications is accuracy. To inspect an algorithm for bias, an additional step is often undertaken to measure performance across different subpopulations, aiming for consistent accuracy across race, gender, country, and other categories where disparities exist. But just because an algorithm is deemed accurate does not mean it will support fairness in health care applications. In an ideal world, only individual patient health and disease factors would determine—and guide prediction of—clinical outcomes. However, studies have repeatedly demonstrated that this is far from the case. For example, mortality from critical illness has been shown to be higher in disproportionately minority-serving hospitals after adjustment for illness severity and other biological factors that pertain to the patient and to the disease (Rush et al. 2020; Danziger et al. 2020).

Data routinely collected in the process of care are heavily influenced by long-standing social, cultural, and institutional biases. Unless the underlying inequities in our communities are addressed, algorithms will perpetuate, if not magnify, existing health disparities.

The bioethics community has an essential role in providing thoughtful, ethical consideration of machine learning health care applications (ML-HCAs), and the model pipeline framework proposed by Char and colleagues (2020) is an important step forward in promoting the systematic identification of relevant ethical concerns. However, it would be misguided to think that an insufficient consideration of ethics is the sole factor leading to poorly designed systems that harm users (Mittelstadt 2019). Indeed, despite a wide range of AI ethics frameworks and principles being published in recent years, the challenges of building ethical AI are as acute and pervasive as ever. It remains unclear whether the influx of guidelines has actually made any impact on improving the ethical development and implementation of AI.

As the bioethics community increasingly turns its attention to ML-HCAs, it is therefore crucial for individuals and organizations to think beyond narrow definitions of machine learning or specialized disciplines, and to engage in addressing disparities with a holistic perspective. ML-HCAs exist in complex societal contexts, rife with biases, disparities, and ethical issues—requiring deeper commitment than a general adoption of ethical frameworks and principles. Efforts that do not actively address the disparities they are informed by, and within which they operate, will always fall short (500 Women Scientists Leadership 2020). As Timnit Gebru has noted, it is not just a matter of biased inputs leading to biased outputs. Ethical considerations of fairness or bias need to be considered beyond any given ML or AI application (Gebru and Denton 2020). In a health care setting for the benefit of patients, this means stakeholders including bioethicists must intentionally extend the typical range of concerns considered when thinking about what constitutes ethical or fair AI. As Chen and colleagues note:

> Researchers often frame theoretical problems of disparity around achieving algorithmic fairness. Looking forward, these conversations should be expanded to acknowledge the systematic dimension of health disparity, taking into account that data is collected in the context of a flawed and unjust system. The research community itself should continue to promote and drive diversity within the field of AI, as more diverse perspectives will ensure that the right questions are asked. (Chen et al. 2020, 16)

A model pipeline framework such as the one proposed by Char and colleagues can certainly help to design better ML-HCAs, but only if the scope of ethical deliberations is attuned to matters of structural inequity and how technologies can be used and co-opted to marginalize subpopulations.

With the scale of patient impacts at stake, the AI community must recognize the call to action, for instance, in hiring, building, and supporting more diverse teams of developers, engineers, and research scientists. For the AI community and the field of bioethics, conversations around how to address persistent inequities in relation to both AI development and clinical consequences are deeply needed—and so too are scholarship and discussion focused on topics including social determinants of health, racial and social justice, LGBTQ+ ethics, disability ethics, and factors that contribute to vulnerable populations. As Keisha Ray recently noted in an opinion piece on #BlackBioethics, the field's bias in this respect represents a failure to address critical needs and scholarship (Ray 2020). For the worlds of both AI and bioethics to enable the change needed in modern health care, it will be essential to co-design approaches with those in marginalized communities who have experienced harmful effects of technologies (Technology can't fix this 2020).

The model pipeline framework is a solid step forward for AI ethics in health care, but ultimately frameworks are never enough. If we want to make sure that AI applications in health care contribute to a society we want to live in, we each have to ensure we are building that society—both in AI ethics and beyond.

ORCID

Stuart McLennan http://orcid.org/0000-0002-2019-6253
Meredith M. Lee http://orcid.org/0000-0002-1405-577X
Amelia Fiske http://orcid.org/0000-0001-7207-6897
Leo Anthony Celi http://orcid.org/0000-0001-6712-6626

REFERENCES

Celi, L. A., B. Fine, and D. J. Stone. 2019. An awakening in medicine: The partnership of humanity and intelligent machines. *The Lancet Digital Health* 1 (6):e255–e257. doi: 10.1016/S2589-7500(19)30127-X.

Char, D. S., M. D. Abràmoff, and C. Feudtner. 2020. Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics* 20 (11):7–17. doi: 10.1080/15265161.2020.1819469.

Chen, I. Y., S. Joshi, and M. Ghassemi. 2020. Treating health disparities with artificial intelligence. *Nature Medicine* 26 (1):16–17. doi: 10.1038/s41591-019-0649-2.

Danziger, J., M. Ángel Armengol de la Hoz, W. Li, et al. 2020. Temporal trends in critical care outcomes in U.S. minority-serving hospitals. *American Journal of Respiratory and Critical Care Medicine* 201 (6):681–687. doi: 10.1164/rccm.201903-0623OC.

Gebru, T., and E. Denton. 2020. Tutorial on fairness accountability transparency and ethics in computer vision at CVPR. https://sites.google.com/view/fatecv-tutorial/schedule

Mittelstadt, B. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1 (11):501–507. doi: 10.1038/s42256-019-0114-4.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366 (6464): 447–453. doi: 10.1126/science.aax2342.

Ray, K. 2020. Black bioethics and how the failures of the profession paved the way for its existence. *Hastings Bioethics Forum*. https://www.thehastingscenter.org/black-bioethics-and-how-the-failures-of-the-profession-paved-the-way-for-its-existence/

Rush, B., J. Danziger, K. R. Walley, A. Kumar, and L. A. Celi. 2020. Treatment in disproportionately minority hospitals is associated with increased risk of mortality in sepsis: A national analysis. *Critical Care Medicine* (7): 962–967. doi: 10.1097/CCM.0000000000004375.

Technology can't fix this. 2020. *Nature Machine Intelligence* 2:363. Available at: https://www.nature.com/articles/s42256-020-0210-5

500 Women Scientists Leadership. 2020. Silence Is Never Neutral; Neither Is Science: Ignoring science's legacy of racism or a wider culture shaped by white supremacy doesn't make scientists "objective". *Scientific American*. https://blogs.scientificamerican.com/voices/silence-is-never-neutral-neither-is-science/

THE AMERICAN JOURNAL OF BIOETHICS
2020, VOL. 20, NO. 11, 22–24
https://doi.org/10.1080/15265161.2020.1819471

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

Where Bioethics Meets Machine Ethics

Anna C. F. Lewis

Harvard University

Char et al. (2020) question the extent and degree to which machine learning applications should be treated as exceptional by ethicists. It is clear that of the suite of ethical issues raised by machine learning applications, many are familiar from other settings. The framework for identifying these issues offered by Char et al., alongside numerous others (Jobin et al. 2019), can be useful in mapping them out. There is at least one clear way in which machine learning is exceptional, and that is the degree of formalism—of codification—demanded of the moral frameworks employed in the development of applications. This topic has spawned the sub-field of machine ethics, part of the broader AI Ethics (Anderson and Anderson 2011). Through getting into

CONTACT Anna C. F. Lewis annalewis@fas.harvard.edu Harvard University, Cambridge, MA 02138, USA.

This article was originally published with errors, which have now been corrected in the online version. Please see Correction (http://dx.doi.org/10.1080/15265161.2020.1858635)