Journal of the American Medical Informatics Association, 27(12), 2020, 2020–2023
doi: 10.1093/jamia/ocaa094
Advance Access Publication Date: 18 June 2020
Perspective

Perspective

Latent bias and the implementation of artificial intelligence in medicine

Matthew DeCamp[1] and Charlotta Lindvall[2,3,4]

[1]Department of Medicine, University of Colorado, Aurora, Colorado, USA, [2]Department of Psychosocial Oncology and Palliative Care, Dana-Farber Cancer Institute, Boston, Massachusetts, USA, [3]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA, and [4]Harvard Medical School, Harvard University Boston, Massachusetts, USA

*Corresponding Author: Matthew DeCamp, MD, PhD, University of Colorado-Center for Bioethics & Humanities, Fulginiti Pavilion for Bioethics and Humanities - Mailstop B137, 13080 E. 19th Avenue, Aurora, CO 80045, USA (matthew.decamp@-cuanschutz.edu)

Received 16 April 2020; Revised 30 April 2020; Editorial Decision 1 May 2020; Accepted 8 May 2020

ABSTRACT

Increasing recognition of biases in artificial intelligence (AI) algorithms has motivated the quest to build fair models, free of biases. However, building fair models may be only half the challenge. A seemingly fair model could involve, directly or indirectly, what we call "latent biases." Just as latent errors are generally described as errors "waiting to happen" in complex systems, latent biases are biases waiting to happen. Here we describe 3 major challenges related to bias in AI algorithms and propose several ways of managing them. There is an urgent need to address latent biases before the widespread implementation of AI algorithms in clinical practice.

Key words: artificial intelligence, machine learning, bias, clinical decision support, health informatics

INTRODUCTION

Artificial intelligence (AI) in general, and machine learning in particular, by all accounts, appear poised to revolutionize medicine.[1–3] With a wide spectrum of potential uses across translational research (from bench to bedside to health policy), clinical medicine (including diagnosis, treatment, prediction, and healthcare resource allocation), and public health, every area of medicine will be affected. Estimates suggest upwards of $6 billion of investment in AI and healthcare by 2021.[4]

Health care leaders are optimistic about the future dissemination and implementation of AI in health systems; in some surveys, more than half expect AI to be in widespread use within the next few years.[5] Physicians are somewhat less optimistic; some worry it could replace them.[6] Patients may be supportive of AI, particularly if it lowers healthcare costs and improves their care,[7] though some fear it could interfere with their relationships with clinicians.[8]

Among the many concerns about AI that have garnered widespread public attention, the most controversial and pressing may be the challenge of identifying biases in AI algorithms.[9,10] These biases include those related to missing data and patients not identified by algorithms, misclassification, observational error, and misapplication. One university considered using AI to direct case management resources to patients for early discharge, until leaders recognized that doing so would preferentially benefit wealthy white patients and disadvantage poorer African-Americans.[11] A commercial algorithm to guide resource allocation in healthcare was found to be profoundly biased against black patients.[12]

Increasing recognition of biases in AI algorithms has motivated the quest to build fair models, free of biases. This quest, though laudable, is no easy feat. Moreover, building fair models may be only half the challenge. In this perspective, we imagine the development of a fair AI predictive model that is free of bias, implemented within the electronic health record (EHR), and adaptive (ie, it is not "locked" but instead continues to learn and improve in performance over time). This hypothetical model operates in a decision support

manner; it is not autonomous, meaning patients and clinicians retain the final decision authority.

However, even this seemingly fair model could involve, directly or indirectly, what we call "latent biases." Just as latent errors are generally described as errors "waiting to happen" in complex systems, latent biases are biases waiting to happen. Here we describe 3 major challenges related to bias in AI algorithms and propose several ways of managing them (Figure 1). There is an urgent need to address latent biases before the widespread implementation of AI algorithms in clinical practice.

THREE CHALLENGES

The first major bias-related challenge for this hypothetically fair algorithm is that, as an adaptive model, it can become biased over time. This can occur in a number of ways. An AI algorithm trained to operate fairly in 1 context could learn from disparities in care in a different context and start to produce biased results; or the algorithm might simply learn from pervasive, ongoing, and uncorrected biases in the broader healthcare system that lead to disparate care and outcomes.

For example, an algorithm to predict patient mortality or an individual patient's response to particular treatments could learn from existing racial, ethnic, and socioeconomic disparities in care and predict worse outcomes for those patients. In effect, a negative feedback loop could be created whereby biases are reinforced over time, further worsening biases in prediction. This matters clinically because prediction can help direct healthcare resources and make subsequent treatment recommendations (eg, palliative care), among other uses. Even more importantly, it is now known that this is possible even when the algorithm is not permitted to produce output based on the variable in question (such as race or zip code) or when the dataset does not even include that variable. This can happen when other variables are correlated or act as proxies for the variable that was removed. This makes the otherwise intuitive strategy for managing biases (ie, excluding variables of concern, such as race, zip code, and so on) infeasible.

A second set of bias-related challenges arises from the interaction of AI with clinical environments that include their own implicit and explicit biases.[11,13] Two phenomena within the setting of patient-clinician interaction are worth noting. One is the phenomenon of automation bias (ie, treating AI-based predictions as infallible or following them unquestioningly).[14] Even an AI-based algorithm that operates in principle as merely a decision support tool can become *de facto* autonomous when its predictions are almost always followed. Busy clinicians who are pressed for time or who fear increased legal liability for overriding (rather than following) an algorithm's output may therefore unintentionally fail to notice biased outputs.

The other is the phenomenon of privilege bias (ie, disproportionately benefiting individuals who already experience privilege of 1 sort or another).[11] Even a perfectly fair algorithm can perform unfairly if it is only implemented in certain settings, such as clinics serving mainly wealthy or white patients. Historical distrust of the health system in general can cause certain patients to distrust the algorithm and hence not follow the recommendations it makes.

A third and final manner in which bias can arise even for fair algorithms is in the choice of what the model is intended to promote (ie, the goal or outcome of interest). This may not at first glance appear to be a "bias" akin to traditional racial, ethnic, and socioeconomic biases. However, when the outcomes of interest or the problems chosen to be solved by AI do not reflect the interests of individual patients or the community, this is in effect a bias: preferentially selecting or encouraging 1 outcome over others. To illustrate with an analogy, 1 reason many clinical trials have failed to improve clinical care is because the outcomes chosen for studies may be surrogates, composites, or other endpoints not relevant to the patients themselves. For example, heart failure outcomes are often measured using a change in physiological parameters (eg, left ventricular ejection fraction) instead of a change of symptoms (eg, fatigue). This has led to an increasing movement toward patient-reported outcome measures (PROMs) in both research and clinical care in order to overcome bias in the choice of outcome.

The outcomes of interest to various healthcare stakeholders—from patients to clinicians to health system leaders to payers and beyond—vary widely between stakeholders. Patients, for example, are likely to care most about improving their own health outcomes and/or lowering their out-of-pocket costs and to care relatively less about efficient resource allocation at the system level. As a result, it is important to acknowledge and address potential biases related to how and why decisions are made to use AI for some purposes and not others.

MANAGING THE CHALLENGES OF EMERGENT BIASES

Biases that emerge from adaptive AI-based algorithms after they are deployed can be best described as "latent biases," (ie, biases waiting to happen). Here we purposely draw from the concept of latent errors in complex systems, understood as errors waiting to happen. Latent errors are failures of organizational design or process that, under the right circumstances, can lead to real errors and harm to patients. Like latent errors, latent biases are not intentional, nor are they unavoidable; instead, they are predictable outcomes that will occur with some level of probability or risk and with some magnitude of harm. This means there is an affirmative ethical obligation to begin addressing them. We propose doing so in 3 ways.

First, addressing latent bias in AI algorithms should be seen as a patient safety issue to be identified and addressed proactively and preferably *ex ante*—not after the fact. Efforts to demonstrate biases in AI algorithms after they have been deployed are laudable and important. Approaches are needed, however, to detect biases in advance and in real time. AI algorithms need to be monitored for biases in predictive performance and also for biases in the way their predictions are used in clinical care. Decades of experience in the struggle to ameliorate health disparities have shown that unequal processes and outcomes cannot be addressed only after the fact (ie, by waiting for biases to arise and then taking steps to remediate their effects). The idea of allowing AI applications to be a proverbial rising tide that initially lifts all boats, followed by separate efforts to remediate inequalities, may be intuitively appealing but is disproven by history. Characterizing latent biases as a patient safety issue helps ensure they will be addressed in advance.

Latent biases in AI performance are important no matter what, but they are more significant in some use cases than others. High stakes medical decisions (eg, about chemotherapeutic recommendations or mechanical ventilation), decisions that affect the resources patients do or do not receive (eg, care management or post-hospital discharge support), and decisions that are automated or made only by the machine deserve special scrutiny. Decision support systems that make claims such as "Patients like you have chosen [X] in similar

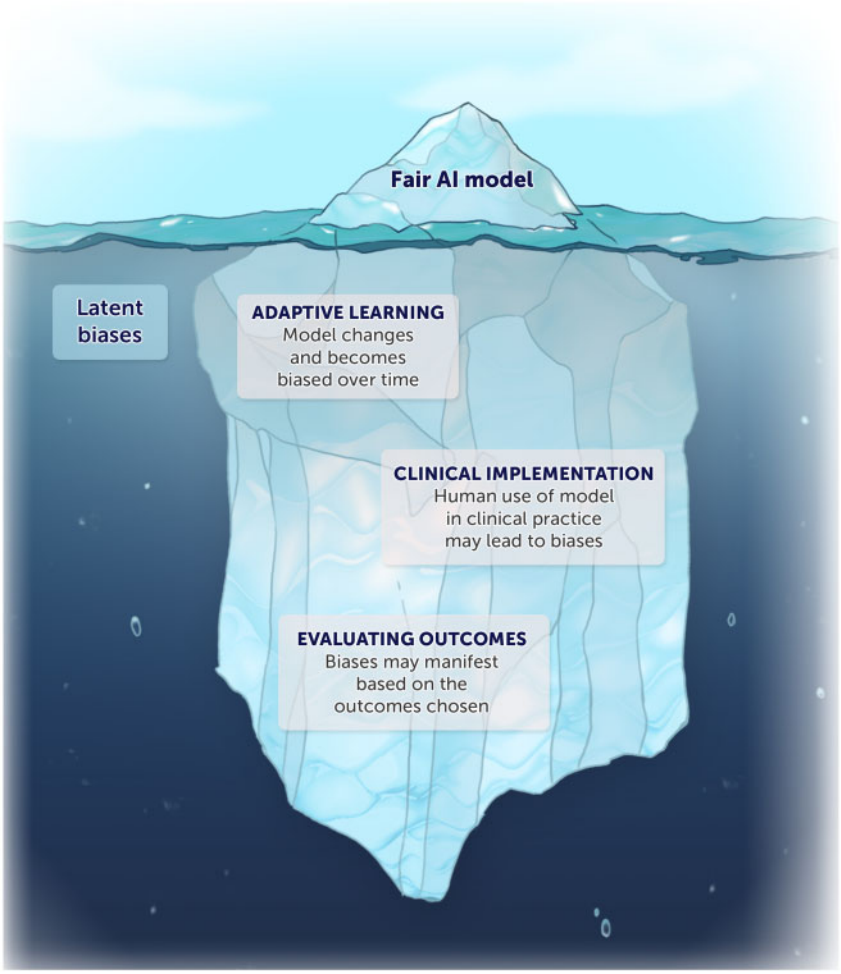

Figure 1. A seemingly fair AI model could involve latent biases after clinical implementation.

circumstances" do as well. Such an algorithm could be biased because of adaptive preferences, whereby choice patterns are affected by past restricted or inappropriately influenced choice options. An algorithm that suggests African-American patients are more likely to choose aggressive care at the end of life, for example, could be based on historical care patterns or health system distrust that may or may not apply to an individual patient's or family's treatment preferences.

Second, regulatory frameworks governing AI and machine learning algorithms must explicitly include reference to monitoring for biases in performance, including those that emerge. Proposed guidance from the US Food and Drug administration on adaptive algorithms, for example, is appropriately tailored to the riskiness of a particular medical application (eg, whether the medical condition is critical, serious, or nonserious, and whether the information provided by the AI aims merely to inform clinical decisions or to treat the disease).[15] However, concerns about bias are not included. Practically, biases that emerge over time should be treated as adverse events; in practice, biases mean that some patients can benefit from an AI application while others are harmed. This would imply that the disparate impacts of AI that result from biases and that cause harm to patients should be part of mandatory device reporting requirements. If a drug product were found to benefit only certain patients and harm others, we would expect this to be reported and managed; we should expect the same of AI algorithms.

Third, recognizing the challenges posed by the different perspectives on appropriate uses and applications of AI in medicine, there is an ever-present need to engage all healthcare stakeholders in the design and implementation of AI. Dissemination and implementation (D&I) science increasingly recognizes engagement as critical from start to finish in order to ensure the effective and ethical implementation of interventions. Engagement provides a way to help avoid biases related to challenges in defining which applications are appropriate for AI and which are not. To illustrate, there may be disagreement about whether an AI machine should be used to calculate and predict an individual's mortality or time of death. For some, the accuracy afforded by a machine may aid their personal decision-making, but for others, this may not be the sort of decision or judgment an algorithm should make. Studies are beginning to explore how patients, clinicians, and health system leaders perceive AI, but more research is needed in this area.

In addition, engaging patient stakeholders requires, as a matter of respect, understanding the circumstances under which they should be informed when AI is being used in their care. This may not be necessary in all cases. Clinicians are not ordinarily required to disclose that an algorithm read an electrocardiogram, for example, but we might think there is an obligation to report AI was used to inform chemotherapeutic decisions.

CONCLUSION

Biases in AI-based algorithms can result not only from biased training data but also from how the algorithms learn over time and are used in practice. Given the pervasiveness of biases, no excuse exists for not taking them seriously. A failure to proactively and comprehensively mitigate all biases—including latent ones that only emerge over time—risks exacerbating health disparities, eroding public trust in healthcare and health systems, and somewhat ironically, hindering the adoption of AI-based systems that could otherwise help patients live better lives.

FUNDING

A portion of Dr. Lindvall's salary was supported by the Cambia Health Foundation Sojourns Scholar Leadership Program.

AUTHOR CONTRIBUTIONS

Dr. DeCamp and Dr. Lindvall developed the conceptual idea and wrote the manuscript together. Both authors approved the final version.

ACKNOWLEDGMENTS

The authors would like to thank Kai-ou Tang for making the illustration.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Topol E. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books; 2019.
2. Miller DD. The medical AI insurgency: what physicians must know about data to practice with intelligent machines. *Npj Digit Med* 2019; 2 (1): 62.
3. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care: a report from the National Academy of Medicine. *JAMA* 2020; 323 (6): 509–10.
4. Frost & Sullivan. Cognitive Computing and Artificial Intelligence Systems in Healthcare: Ramping Up a $6 Billion Dollar Market Opportunity. 2015. https://store.frost.com/cognitive-computing-and-artificial-intelligence-systems-in-healthcare.htmlAccessed March 28, 2020.
5. Obermeyer Z, Weinstein JN. Adoption of artificial intelligence and machine learning is increasing, but irrational exuberance remains. *NEJM Catal Innovations Care Deliv* 2020; 1 (1): https://doi.org/10.1056/CAT.19.1090.
6. Blease C, Bernstein MH, Gaab J, *et al*. Computerization and the future of primary care: a survey of general practitioners in the UK. *PLoS One* 2018; 13 (12): e0207418.
7. Pearson D. Americans ready to embrace healthcare AI on one condition. 2020. https://www.aiin.healthcare/americans-ready-embrace-healthcare-ai-one-conditionAccessed March 28, 2020.
8. Syneos Health Communications. Artificial intelligence for authentic engagement: patient perspectives on health care's evolving AI conversation. 2018. https://syneoshealthcommunications.com/perspectives/artificial-intelligence-for-authentic-engagementAccessed March 28, 2020.
9. Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. *PLoS Med* 2018; 15 (11): e1002689.
10. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med* 2018; 378 (11): 981–3.
11. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018; 169 (12): 866–72.
12. Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
13. Parikh RB, Teeple S, Navathe AS. Addressing bias in artificial intelligence in health care. *JAMA* 2019; 322 (24): 2377–8.
14. Challen R, Denny J, Pitt M, *et al*. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019; 28 (3): 231–7.
15. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD). US Food and Drug Administration; 2019:20. https://www.fda.gov/media/122535/downloadAccessed March 28, 2020.