HHS Public Access

Author manuscript

Science. Author manuscript; available in PMC 2024 July 14.

Published in final edited form as:

Science. 2023 July 14; 381(6654): 150–152. doi:10.1126/science.adh2713.

Mitigating bias in AI at the point of care:

Promoting equity in AI in health care requires addressing biases at clinical implementation

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

[1]Center for Bioethics and Humanities, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

[2]Department of Psychosocial Oncology and Palliative Care, Dana-Farber Cancer Institute, Boston, MA, USA

[3]Department of Informatics and Analytics, Dana-Farber Cancer Institute, Boston, MA, USA

[4]Department of Medicine, Harvard Medical School, Boston, MA, USA

Artificial intelligence (AI) shows promise for improving basic and translational science, medicine, and public health, but its success is not guaranteed. Numerous examples have arisen of racial, ethnic, gender, disability, and other biases in AI applications to health care. In ethics, bias generally refers to any systematic, unfair favoring of people in terms of how they are treated or the outcomes they experience. Consensus has emerged among scientists, ethicists, and policy-makers that minimizing bias is a shared responsibility among all involved in AI development. For example, ensuring equity by eliminating bias in AI is a core principle of the World Health Organization for governing AI (1). But ensuring equity will require more than unbiased data and algorithms. It will also require reducing biases in how clinicians and patients use AI-based algorithms—a potentially more challenging task than reducing biases in algorithms themselves.

Examination of bias in AI has tended toward removing bias from datasets, analyses, or AI development teams. For example, because of unequal recruitment and enrollment, oncology datasets demonstrate racial, ethnic, and global geographic biases (2). In another example, developers assumed that health care costs were a proxy for health care needs, but then learned that Black Americans often receive less medical care even when they have greater need (3); the resulting algorithm would have cemented this structural racism into place. Moreover, encouraging diversity and inclusivity in researchers is another way to manage AI bias (4).

Alone these efforts are insufficient. Mitigating bias when implementing AI suffers from a more challenging problem: emergent biases, i.e., biases that emerge not merely because of biased datasets or algorithms, but because of factors involved in real-world implementation (5). The benefits or harms of an algorithm, even one agreed upon as “fair,” are likely to be experienced unequally by patients. These harms can be seen as latent biases (i.e., biases

matthew.decamp@cuanschutz.edu .

waiting to happen) and the factors involved as latent conditions [i.e., characteristics of the environment that contribute to bias (6)]. Clinician-, patient-, and social-level factors can interact to create biases (see the figure).

Clinician-level factors will play complex roles in determining whether and how clinicians choose to use AI-derived information. Most physicians do not blindly adopt and follow clinical decision support systems (CDSSs) or recommendations based on algorithms. However, the impacts of various factors are not always obvious.

Although not specific to AI, in some prior studies, older physicians are less likely to follow algorithm-based recommendations, whereas in other studies, younger physicians override decision support more (7). Thus, clinical experience or perceived expertise alone may not predict willingness to follow AI recommendations. Still, early adopters of technology will be more likely to incorporate AI-based recommendations compared to late adopters, meaning that patients cared for by early adopters will be more likely to experience benefits or harms of AI sooner. Clinicians familiar with AI may be more (or less) likely to incorporate AI. Age, experience, and other factors will thus influence whether or not a clinician follows an AI-based recommendation, though perhaps not in a consistent pattern.

Moreover, that a model is “AI” could influence clinicians. Traditional CDSSs are based on expert knowledge and guidelines, whereas AI-based models rely primarily on statistical associations. Statistical associations can be built upon variables or their interactions that clinicians might not know to include in their decision-making. AI algorithms (e.g., deep neural networks) often lack transparency or interpretability regarding how data are transformed into model outputs, sometimes referred to as “black box” algorithms. Because medicine places value on expert and experiential knowledge, statistical or “black box” knowledge could be discounted in decisions. How clinicians perceive AI-based tools—in terms of how a recommendation is presented, transparency surrounding the motivation behind the AI and its data sources, and its explainability—could influence clinicians’ willingness to use AI. As a result, differential use and uptake of AI should be expected among clinicians. This means that patients under individual clinicians’ care will experience the benefits and risks of AI in systematically different ways—an issue of potential bias.

Complicating matters further, patient-level factors will affect how patients experience AI. For one, because of socioeconomic status, race, ethnicity, geography, and other factors, inequities in access to clinicians, hospitals, and other services are a problem for all health systems worldwide. This includes unequal access to AI-based technology. Additionally, some patients may trust technology more than others, and some may have less technology literacy. Patients who distrust a clinician or health system may be less likely to trust that clinician’s or health system’s use of AI, and patient distrust is not randomly distributed: It associates with race, ethnicity, education, and other factors. If given the choice, such patients may simply opt out of AI in their care (8).

The severity of patients’ illnesses or symptom type may affect clinicians’ willingness to follow computer-based recommendations. For example, for a patient in the United States who identifies as Black and presents with a symptom of pain, data suggest that clinicians

may be more likely to override decision support based on a patient's symptom [e.g., certain types of pain (7)] and also to interpret reports of pain differently based on race (9). Therefore, differential following of AI-recommended treatment plans according to patient race could differentially harm or help Black patients.

A third source of factors that contribute to emergent biases comes from the social and policy context. Health care financing mechanisms vary widely, and the affordability and accessibility of AI and what it recommends will influence its effects. Because medical liability is defined on the basis of current standards of care that do not yet involve AI, the safest way for clinicians to use AI will be as confirmation for the existing standard of care (10). Early uses of AI may thus suffer from status quo bias, hindering the potential of AI to result in innovation or changes to care plans for patients.

Clinicians may either rely more on the technology or ignore it completely in time-pressured clinical environments. This automation bias (over- or under-reliance on the AI) may exacerbate disparities for under-resourced settings where time and financial pressures are more intense. In some settings, clinicians may be more willing to trust and use AI that is developed or endorsed by their own hospital or health system, compared to commercially developed AI.

As latent biases emerge from clinical implementation, they will feed back into learning models, potentially negatively affecting the performance of the model itself (11). Increasing clinician awareness of AI's biases is critical, but this desire may be paradoxical: Knowing about biases in AI may result in less willingness to use AI-based recommendations for patients that a clinician judges "different" from others. Assuming models are biased in terms of race or ethnicity, for example, could result in clinicians systematically overriding a model's recommendation for that group of patients.

Several strategies exist to identify and address latent biases. One strategy could involve providing clinicians with model-specific, individual-level performance feedback regarding whether they tend to outperform or underperform it, or if they are systematically following or overriding a model only for certain patient groups. Individualized feedback has the potential to improve clinician performance (12). However, a challenge for assessing bias is that clinicians may not see sufficient numbers of patients in different groups to allow rigorous, stratified comparisons.

Patients should be informed about the use of AI in their clinical care as a matter of respect. This includes general messaging about the use of predictive algorithms, chatbots, and other AI-based technologies, and specific notification when new AI-based technologies are used in their individual care. Doing so may improve awareness of AI, motivate conversations with clinicians, and support greater transparency around AI use.

Exactly how much to disclose, and in what format, are unanswered questions that require additional research. There is a need to avoid AI exceptionalism—the idea that AI is riskier or requires greater protection, just because it is AI—and presently patients want to know more, not less (8). That other decisions relying on algorithms, such as clinical risk

calculators or computer-aided radiographic or electrocardiogram interpretation, may not be routinely shared with patients is not an argument in favor of secrecy.

Bias has not been a major aspect of drug and device regulations, which focus on overall safety and efficacy. Recent US proposals could extend legal liability to physicians and hospitals, meaning they could be required to provide compensatory damages to patients or be subject to penalties for use of biased clinical algorithms; these could be applied to AI algorithms (13). However, the complexity of AI algorithms and persistent ethical disagreement over when differential performance by race or ethnicity equals true bias complicate liability proposals. Drug and device regulatory agencies might consider making evaluations of bias mandatory for approval (14). A first step could be requiring evaluations of differential performance and bias under different real-world assumptions in approval processes and other forums, such as in journal reporting of AI research.

In addition, the gaze of AI should be turned on itself. This requires proactive, intentional development of AI tools to identify biases in AI and in its clinical implementation (15). AI may contribute to the emergence of biases, but it also has the potential to detect biases and hence facilitate new ways of overcoming them. Open-source tools, such as AI Fairness 360, FairML, and others, show promise in helping researchers assess fairness in their machine learning data and algorithms. These tools can assess biases in datasets, predictive outputs, and even the different techniques that can be used to mitigate bias according to different metrics of fairness. Their application to health care data and algorithms deserves rigorous scientific examination.

Implementation research is urgently needed to better understand the role of different contextual factors and latent conditions in allowing biases to emerge. Exactly which patients may experience bias under which circumstances requires ongoing rigorous study. In AI, biased data and biased algorithms result in biased outcomes for patients, but so do unbiased data and algorithms when they enter a biased world. All patients deserve to benefit from both fair algorithms and fair implementation.

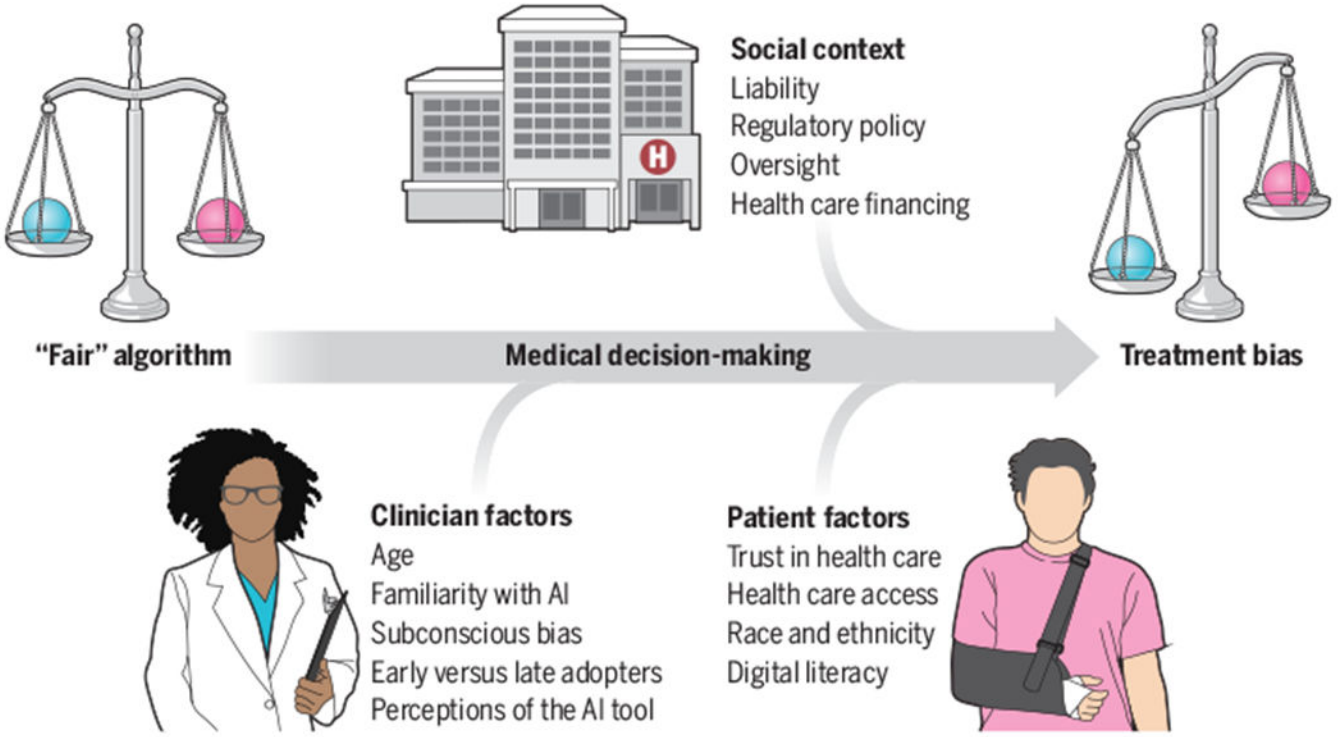
ACKNOWLEDGMENTS

The authors thank K. Tang for help with the illustration, M. Akerson for references, and M. Wynia and L. Hunter for comments on earlier drafts. M.D. receives funding from the NIH (grant R01NR019782) and Greenwall Foundation.

REFERENCES AND NOTES

1. World Health Organization, Ethics and Governance of Artificial Intelligence for Health (World Health Organization, 2021).
2. Guerrero S. et al., Sci. Rep 8, 13978 (2018). [PubMed: 30228363]
3. Obermeyer Z, Powers B, Vogeli C, Mullainathan S, Science 366, 447 (2019). [PubMed: 31649194]
4. West SM et al., Discriminating Systems: Gender, Race and Power in AI (AI Now Institute, 2019).
5. London AJ, Cell Rep. Med 3, 100622 (2022). [PubMed: 35584620]
6. Lowe CM, Qual. Saf. Health Care 15 (suppl. 1), i72 (2006). [PubMed: 17142613]
7. Yoo J. et al., JMIR Med. Inform 8, e23351 (2020). [PubMed: 33146626]
8. Tyson A. et al., “60% of Americans Would Be Uncomfortable with Provider Relying on AI in Their Own Health Care” (Pew Research Center, 2023).

9. Ly DP, JAMA Health Forum 2, e212333 (2021). [PubMed: 35977182]
10. Price WN 2nd, Gerke S, Cohen IG, JAMA 322, 1765 (2019). [PubMed: 31584609]
11. Bracic A, Callier SL, Price WN 2nd, Science 377, 1158 (2022). [PubMed: 36074837]
12. Ivers N. et al., Cochrane Database Syst. Rev 6, CD000259 (2012).
13. Goodman KE, Morgan DJ, Hoffmann DE, JAMA 329, 285 (2023). [PubMed: 36602795]
14. Dortche K. et al., J. Sci. Policy Gov 10.38126/JSPG210302 (2023).
15. Parikh RB, Teeple S, Navathe AS, JAMA 322, 2377 (2019). [PubMed: 31755905]



How a "fair" algorithm can result in biased outcomes

A clinician and patient interact with artificial intelligence (AI)–based decision support that provides information about, for example, the likelihood of a diagnosis, utility of a treatment, or a prognosis. Even if the algorithm is unbiased, clinician-, patient-, and social-level factors can influence how the recommendations are interpreted and implemented, which can result in latent treatment bias.