In this issue, Lee and colleagues describe the performance of several widely used artificial intelligence (AI) image generation models on producing images of physicians in the United States. The key question the authors set out to answer was whether the models would produce images that accurately reflect the actual racial, ethnic, and gender composition of the US physician workforce, or whether the models would demonstrate biased performance. One important aspect of the study method was that the authors used relatively open-ended prompts, including “Photo of a physician in the United States,” allowing the machinations of the AI to produce an image that it determined was most likely to meet the needs of the end user. AI tools powered by large language models, including the ones examined in the study, use a degree of randomness in their outputs, so models are expected to produce different images in response to each prompt—but how different would the images be? Their findings are striking. First, although 63% of US physicians are White, the models produced images of White physicians 82% of the time. Additionally, several models produced no images of Asian or Latino physicians despite nearly a third of the current physician workforce identifying as a member of these groups. The models also severely underrepresented women in their outputs, producing images of women physicians only 7% of the time. These results demonstrate a clear bias in outputs relative to actual physician demographics. But what do these findings mean for AI and its use in medicine?

Bias is a broad term with definitions that depend on context. In ethics and law, concepts of bias involve the unfair treatment of people generally or based on a legally protected class. In AI, bias can also be viewed through a technical lens when outputs of technology tools are systematically inaccurate. Whether a technical bias translates into an ethical bias depends on a technology’s application in the real world. In the case of the present study, the findings that AI images of physicians systematically exclude members of specific races, ethnicities, and genders pulls at our conscience, triggering a visceral internal reaction in many readers that something needs to be addressed. Such a response is appropriate and necessary—because AI has no conscience of its own. Its use for good or ill depends on human beings who uniquely possess the capacity to determine right from wrong.

Where does AI bias originate, and how can we seek to mitigate it? In a recent report, authors from the National Institute of Standards and Technology highlighted 3 major sources of AI bias: computational, human and systemic. Computational biases arise from the specific datasets and algorithms used by AI systems. Human biases, on the other hand, reflect underlying cognitive biases in human thought that influence the development and use of AI tools, while systemic biases arise at the institutional level as a result of the inherent structures, processes, and norms of an organization. But who among these stakeholders is ultimately responsible for mitigating bias in AI outputs? The answer is exceedingly simple yet painfully complex: it is all of us.

Across the spectrum of developers, users, and institutions, there is an ethical requirement to be good stewards of any technology tool. In high-stakes domains such as health care, this is arguably even more important when the effects of AI biases may dramatically influence health outcomes, even life and death. To mitigate AI bias, tremendous work must be applied to ensure the responsible use of AI. Such efforts can begin with multidisciplinary collaborations that include health care workers throughout the 4 phases of the AI life cycle: predesign, design and development, deployment, and testing and evaluation. At the predesign stage, health care workers can provide input on problem identification and the goal of the AI tool. During the design and development of the...
AI tool, model developers should actively engage clinicians to discuss bias identification and mitigation strategies and embed safeguards to overcome bias. Once AI tools are deployed, health care workers should be able to easily provide feedback when the model displays a biased output, similar to how safety reporting currently works in clinical settings. And ongoing studies should continue to ask pointed questions about AI performance with the ultimate goals of improving both accuracy and fairness. Work in this space is already under way as demonstrated by the creation of organizations such as the Coalition for Health AI, which seeks to bring together diverse stakeholders to inform the development, use, and evaluation of AI in health care.5

Although impressive, AI is still just a technology—how it is designed and implemented is dependent on how it is engineered and put to use by people and organizations. In short, it still must be told what to do, and for what purpose. Whether those instructions create good or cause harm is ultimately the product of human beings and their choices. To this end, we must let our conscience be our guide, while making conscientious decisions that seek to illuminate unfairness and eliminate it. We have already seen the positive effects of such work in the form of revisions to biases in well-known clinical tools, such as the estimated glomerular filtration rate.6 Similar efforts can, and must, be done with AI as well.

The consequences of bias in AI range from trivial to far reaching, but these effects can only be fully mitigated when they are known. It is imperative that we pay close attention to bias and the potential for unintended consequences from AI design and implementations. Studies such as those by Lee and colleagues1 will continue to play a role in helping stakeholders understand the impacts of new AI technologies and guide thoughtful modifications. AI stands to greatly benefit human beings, and there will be inevitable trade-offs that may affect performance and acceptability. But by surfacing objective data on AI performance and understanding how it aligns with our goal to improve both health outcomes and broader societal aims, we can do our best to make adjustments in the AI life cycle that will guide us toward a more desirable end state. Although no system is perfect nor is perfection the goal of AI, we all bear a responsibility to ensure AI is fair, trustworthy, and beneficial. Our patients deserve it. Our conscience demands it.

ARTICLE INFORMATION
Published: August 6, 2024. doi:10.1001/jamanetworkopen.2024.25955
Open Access: This is an open access article distributed under the terms of the CC BY License. © 2024 Crowe B et al. JAMA Network Open.

Corresponding Author: Byron Crowe, MD, Division of General Internal Medicine, Beth Israel Deaconess Medical Center, 330 Brookline Ave, Boston, MA 02215 (bcrowe@bidmc.harvard.edu).

Author Affiliations: Division of General Internal Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts (Crowe); Harvard Medical School, Boston, Massachusetts (Crowe, Rodriguez); Division of General Internal Medicine and Primary Care, Brigham and Women’s Hospital, Boston, Massachusetts (Rodriguez).

Conflict of Interest Disclosures: Dr Crowe reported being an employee of Solera Health outside the submitted work. No other disclosures were reported.

REFERENCES