Medical News & Perspectives | AI IN MEDICINE

AI Developers Should Understand the Risks of Deploying Their Clinical Tools, MIT Expert Says

Samantha Anderer; Yulin Hswen, ScD, MPH

This conversation is part of a series of interviews in which JAMA Editor in Chief Kirsten Bibbins-Domingo, PhD, MD, MAS, and expert guests explore issues surrounding the rapidly evolving intersection of artificial intelligence (AI) and medicine.

AI applications for health care should be designed to function well in different settings and across different populations, says Marzyeh Ghassemi, PhD (**Video**), whose work at the Massachusetts Institute of Technology (MIT) focuses on creating "healthy" machine learning (ML) models that are "robust, private, and fair." The way AI-generated clinical advice is presented to physicians is also important for reducing harms, according to Ghassemi, who is an assistant professor at MIT's Department of Electrical Engineering and Computer Science and Institute for Medical Engineering and Science. And, she says, developers should be aware that they have a responsibility to clinicians and patients who could one day be affected by their tools.

+ Multimedia

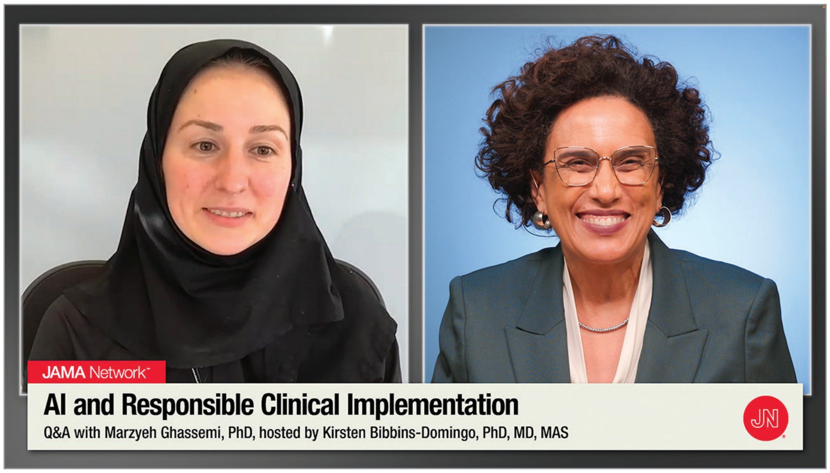
+ Medical News website

JAMA Editor in Chief Kirsten Bibbins-Domingo, PhD, MD, MAS, recently spoke with Ghassemi about "ethical machine learning," the computer scientist's decision to opt out of AI in her own health care, and more.

The following interview has been edited for clarity and length.

DR BIBBINS-DOMINGO: You have a research lab, Healthy ML. It specializes in examining biases in artificial intelligence, and you're specifically interested in its applications in clinical practice. I'd love to hear how you got into the very specific area.

DR GHASSEMI: At the end of my PhD, we found out that [machine learning] models tend not to work as well in all groups. And that really informs what we do here in my lab today, focusing on how we make sure that models that are developed work robustly.


JAMA Network
AI and Responsible Clinical Implementation
Q&A with Marzyeh Ghassemi, PhD, hosted by Kirsten Bibbins-Domingo, PhD, MD, MAS
JN

And if you think about robustness, that could mean that it works well in a new environment or across different kinds of people.

DR BIBBINS-DOMINGO: How do you think about the range of reasons why a model might not perform well in one setting vs another or in one group of people vs another?

DR GHASSEMI: I try to think about it within the pipeline that all models are developed in. And this is not just in health care. This is for any machine learning model that might be developed and deployed in any human-facing setting. You choose a problem, collect some data, define a label, develop an algorithm, and then deploy it. In each part of that pipeline, there are reasons that your model might not perform as well. For problem selection, what we choose to fund and what we choose to work on is often biased. We tend to look at problems that are easy to address where there are more data readily available that can be correlated with different metrics of social status, or privilege, or just where funding tends to be allocated to.

For example, diseases that are disproportionately affecting people who are biologically female at birth tend to be understudied. And if we're collecting data from these human sources, it's probably going to have some bias in it just because of the way that humans interact with one another. Just by collecting data from a human process, you're going to have some potential performance issues. We probably want machine learning models to replicate the very best health care practices that we see now, but if we take a random sample of data from thousands of hospitals and say, "Perform the way that an average doctor is performing on an average day," we might get some behaviors that we don't want to extend.

When we define a label, that's another way that bias can be injected into the learning process. It's a true-false label. We never contextualize it with the choice that's being made or the human rule that's being applied. When you collect labels in this descriptive way but then train a machine learning model, all of those machine learning models become much harsher. They have a much higher false-positive rate.

DR BIBBINS-DOMINGO: You use the term *ethical machine learning*. I'd love you to define what that term means for you and help us to understand it in the context of medical practice.

DR GHASSEMI: I think for me as a technical person, *ethical machine learning* means recognizing your responsibility to end users that might potentially be impacted by the models that you're developing, the technology that you're releasing. And I think there are many ethical frameworks that professional societies have—for engineers, for doctors, for different kinds of individuals that interact with humans.

And that's not standard in computer science training. It wasn't in my computer science curriculum. There wasn't a specific set of rules, or regulations, or even principles that we went over. And now we're seeing a lot of programs like the program at MIT step up and recognize that computer science impacts just as many people as many engineering disciplines do. But I think that we're playing a little bit of catch-up in the field with people starting to recognize that these choices make an impact.

DR BIBBINS-DOMINGO: So, what does that mean for algorithms designed for use in clinical practice settings? Do you just need to be more aware and understand this ethical machine learning? Do you and I need to talk as you are developing a particular model? What types of processes get us to the point where we really are focused on the end user, in this case patients? And what type of team, and what type of processes, and what types of things get us there?

DR GHASSEMI: I think we need a change in the technical people, the technical societies, and the technical systems. We need to speak with and be informed by the needs of those whom we are collaborating with and not just to understand how data might have been collected but how a model might be deployed and what the risks are for such a deployment.

I think the problem here is not just that we're using machine learning and health, it's that we're using this really powerful tool in a space where technology has been reasonably laxly regulated. We're adding this extra tool to a setting that doesn't currently have a lot of regulation, and I think it's a struggle to catch up. If you're upset about a machine learning model learning to kill more women than men, performing more poorly on women than men, but it learned that from the data, maybe we should try to address the underlying problem, which is that more women die in this procedure. Rather than saying, "I'm so angry that the model has learned this thing," let's use the fact that it learned it to address the underlying issue.

DR BIBBINS-DOMINGO: You're speaking about such an important issue, and we are in an environment where this technology is moving at rapid speed, both the capabilities and the enthusiasm for adopting any type of machine learning, AI approach in health care. We also know that these models can be subject to biases. So, in your view, how should we think about regulation once the model's developed or once it's deployed?

DR GHASSEMI: I totally agree with you that it seems like the philosophy here is deploy ahead of regulation, which I don't think is the right way of thinking about the role of technology in the health care setting. What I will say is, I think that the FDA [US Food and Drug Administration] has done really fantastic work toward trying to have systems where audits can be done for machine learning models. I think that there are improvements that could be made, like with any system.

I'm actually a big fan of the multiarm regulatory system that aviation has with different federal agencies that were created decades apart specifically to ensure that there's safety in airplanes that exist, and there's training for pilots to use technology, and that there are standards about how different airlines have to communicate, and there are responsibilities that airlines and carriers have to passengers who fly.

I think that we need the same kind of regulation that is well recognized as being not about assigning blame or liability but about ensuring safety and having a space and a culture of safety. And also that there is some amount of oversight where people voluntarily take a certain amount of training in order to be able to work well with technology prior to having it integrated into their setting.

I do want to address the fact that—unlike in aviation where there were lots of human-computer-interaction-end user studies done to figure out how best to show information to people in a stressful situation who are trying to make decisions—we haven't done a lot of those studies in a human-computer interaction of machine learning or other technology-plus-doctor setting. We don't actually know how best to give information to doctors, information that might be wrong sometimes by the way, such that they are able to use it well when it's right and they're not disproportionately biased by it when it's wrong. The work that we've done so far suggests that the key or one of the keys to making sure that doctors aren't misled by biased information is to make sure that it's given to them descriptively.

DR BIBBINS-DOMINGO: And is that because we trust that it's an AI model, it's math, and therefore we should do what it says?

DR GHASSEMI: Based on other work by really fantastic researchers and work that my lab has done, I think it is two things coupled. Number one, it's an automation bias. It's been well documented in a clinical setting for a long time that if there's a prefilled default, you're more likely to use it.

And the other is exactly what you're saying. We think it's algorithmic overreliance. People assume that they have a system like a robot, or an AI, or an algorithm, whatever it is, that has access to more information than they do or is well aware of the risks that might be encountered by making an incorrect decision in that setting.

And there's been many other documented settings where clinicians have been given incorrect or bad advice. And even when they're made aware that potentially the model could give them incorrect or bad advice, they still exhibit these same automation and overreliance biases. And so, it's something that we need to be really careful about when we consider exactly the way in which we give advice.

DR BIBBINS-DOMINGO: I am so glad you brought up the point that in other sectors where there is either a much longer history or a much closer level of training between computers and humans, like in aviation, there has been a lot of attention placed to how information is presented. And it's clear that we need to understand that much more. It's reminding me of a study we published in *JAMA* just a few months ago on whether the idea of explaining the model can help to give the clinician better insights into where a model might be wrong. It showed that biased models produced the wrong results and the explainability didn't mitigate against the degree to which a clinician was going to be led astray.

I think it speaks a little bit to what you're saying here, and how important it is not just to assume that explaining how the model was built is going to help me not to go down the wrong road.

DR GHASSEMI: It's been well established for a while that explainability methods can make a model less fair because fundamentally they are approximations. How do you make a model explainable? You make it simpler. And so, you have to approximate something. And what we've found previously is that these approximations tend to impact minority groups more than majority groups. Which sort of makes sense. If you need to approximate some complex nonlinear boundary and there's a group you have to do a little bit less well at modeling, it's probably the group that takes up a smaller amount of the space, right? Because that's going to impact your performance less.

And so not only do explainability methods tend to make models less fair in many settings that we evaluated, this study in *JAMA* demonstrates that explainability can even *increase* overreliance sometimes. Because if you just have a number or if you just have a description it doesn't really short-circuit that critical thinking that you have to do to make the decision. But if you make it easy and you start engaging that overreliance and that automation bias where it's telling you what to do, it's explaining the reason, I think that's where we start to see these biases really become very strong.

DR BIBBINS-DOMINGO: It's so interesting. The modeling is complex, but humans and human behavior is also complex.

DR GHASSEMI: I think that's the hardest thing, honestly. It's such a complex system of interactions. I'm making this loose analogy to aviation. It's not aviation. In aviation, you have a plane of hundreds of passengers. And the outcome for one is the outcome for all. They all land safely. And that's not what happens in health care. And so, I think there's so much more we need to do. There's so much more research that needs to be done. And we really lack the backbone to do that because even before machine learning, we have had clinical risk scores that do not work for women.

I always tell people when I give these examples, sometimes they'll say, "Well, a clinical risk score can't work for every tiny subgroup. It's hard to collect from minorities." Women are not a minority. We're half of the planet, sometimes more. And so, the fact that clinical risk scores have historically not worked for half the planet without machine learning, no AI needed, I think speaks to the fact that we need to understand how to use technology in the health care system, even if we didn't have machine learning, in a way that doesn't increase inequity.

DR BIBBINS-DOMINGO: Okay. So, what AI tools do you use?

DR GHASSEMI: I feel like I have to be very clear here because I have two very different opinions about a very fantastic thing. Like many people, when ChatGPT and other versions of GPT were released, I was so impressed with the technical accomplishment. I have spoken very widely about how unhappy I am that it's being used for specific things in a clinical setting. I don't think that that's the best use of it.

But I will say if you write a grant or you have a great research idea, often you have to summarize it 7 different ways: a 100-word abstract for a general audience, a 200-word abstract for a scientific officer, a 300-word.... I love using GPT models to do summarizations of a specific length for a particular audience of work that I've done.

DR BIBBINS-DOMINGO: That's a very good example. But let me give you the opportunity to maybe expand on what you were going to challenge us not to use it for before. What AI tools do you avoid or what would you not use right now?

DR GHASSEMI: I opt out of almost all uses of AI in a health setting. Both for myself and for dependents I have, because I'm well aware of the research, some of which is my own, that the tools are unlikely to work well for a minority female.

DR BIBBINS-DOMINGO: What do you say when someone says, "Well, we are never going to make models that are designed for people like you because you are not letting us use the data on people like you."

DR GHASSEMI: I have spoken to minority communities and told them, "Please let me use your data. My model will not work. It will perform poorly on your population." And that's the reason that clinical models are so bad for so many people, because, sometimes intentionally, only certain groups were studied. What I say is I am doing research that will be peer reviewed, often brutally, and published in some venue. And then if I ever wanted to deploy it, I hope that any deployer, if it's not me, would go through a rigorous approval process of ensuring that that model was robust prior to deployment.

I think there's a fundamental difference between using data for discovery and understanding of the limits of machine learning and health vs automating an efficiency metric, or a decision, or an output that just needs to be obtained for an electronic health care record. I would consent to my data being used in a machine learning paper. But I don't want it used to predict how much care should be allocated for me, or which medications I should have access to, or what kind of doctor I might be available to be referred to, because I know all of those decisions will be biased.

DR BIBBINS-DOMINGO: Your explanation I think helps us to understand where we are in a landscape of an evolving technology that is both very powerful and has known limitations and biases. ■

Published Online: February 7, 2024.
doi:10.1001/jama.2023.22981

Conflict of Interest Disclosures: Dr Ghassemi reported receiving funding from CIFAR, Quanta Computing, Microsoft Research, Helmsley Trust, Wellcome Trust, J-Clinic, IBM, Moore Foundation, Janssen Research and Development, VW Foundation, and Takeda.

Note: Source references are available through embedded hyperlinks in the article text online.