AI's errors may be impossible to eliminate – what that means for its use in health care

Carlos Gershenson, Professor of Innovation, Binghamton University, State University of New York

Published: December 11, 2025 8:37am EDT

Federal legislation introduced in early 2025 proposed allowing AI to prescribe medication.
Wladimir Bulgar/Science Photo Library via Getty Images

In the past decade, AI's success has led to uncurbed enthusiasm and bold claims – even though users frequently experience errors that AI makes. An AI-powered digital assistant can misunderstand someone's speech in embarrassing ways, a chatbot could hallucinate facts, or, as I experienced, an AI-based navigation tool might even guide drivers through a corn field – all without registering the errors.

People tolerate these mistakes because the technology makes certain tasks more efficient. Increasingly, however, proponents are advocating the use of AI – sometimes with limited human supervision – in fields where mistakes have high cost, such as health care. For example, a bill introduced in the U.S. House of Representatives in early 2025 would allow AI systems to prescribe medications autonomously. Health researchers as well as lawmakers since then have debated whether such prescribing would be feasible or advisable.

How exactly such prescribing would work if this or similar legislation passes remains to be seen. But it raises the stakes for how many errors AI developers can allow their tools to make and what the consequences would be if those tools led to negative outcomes – even patient deaths.

As a researcher studying complex systems, I investigate how different components of a system interact to produce unpredictable outcomes. Part of my work focuses on exploring the limits of science – and, more specifically, of AI.

Over the past 25 years I have worked on projects including traffic light coordination, improving bureaucracies and tax evasion detection. Even when these systems can be highly effective, they are never perfect.

For AI in particular, errors might be an inescapable consequence of how the systems work. My lab's research suggests that particular properties of the data used to train AI models play a role. This is unlikely to change, regardless of how much time, effort and funding researchers direct at improving AI models.

Nobody – and nothing, not even AI – is perfect

As Alan Turing, considered the father of computer science, once said: "If a machine is expected to be infallible, it cannot also be intelligent." This is because learning is an essential part of intelligence, and people usually learn from mistakes. I see this tug-of-war between intelligence and infallibility at play in my research.

In a study published in July 2025, my colleagues and I showed that perfectly organizing certain datasets into clear categories may be impossible. In other words, there may be a minimum amount of errors that a given dataset produces, simply because of the fact that elements of many categories overlap. For some datasets – the core underpinning of many AI systems – AI will not perform better than chance.

Features of different dog breeds may overlap, making it hard for some AI models to differentiate them.
Miras Wonderland/iStock via Getty Images Plus

For example, a model trained on a dataset of millions of dogs that logs only their age, weight and height will probably distinguish Chihuahuas from Great Danes with perfect accuracy. But it may make mistakes in telling apart an Alaskan malamute and a Doberman pinscher, since different individuals of different species might fall within the same age, weight and height ranges.

This categorizing is called classifiability, and my students and I started studying it in 2021. Using data from more than half a million students who attended the Universidad Nacional Autónoma de México between 2008 and 2020, we wanted to solve a seemingly simple problem. Could we use an AI algorithm to predict which students would finish their university degrees on time – that is, within three, four or five years of starting their studies, depending on the major?

We tested several popular algorithms that are used for classification in AI and also developed our own. No algorithm was perfect; the best ones – even one we developed specifically for this task – achieved an accuracy rate of about 80%, meaning that at least 1 in 5 students were misclassified. We realized that many students were identical in terms of grades, age, gender, socioeconomic status and other features – yet some would finish on time, and some would not. Under these circumstances, no algorithm would be able to make perfect predictions.

You might think that more data would improve predictability, but this usually comes with diminishing returns. This means that, for example, for each increase in accuracy of 1%, you might need 100 times the data. Thus, we would never have enough students to significantly improve our model's performance.

Additionally, many unpredictable turns in lives of students and their families – unemployment, death, pregnancy – might occur after their first year at university, likely affecting whether they finish on time. So even with an infinite number of students, our predictions would still give errors.

The limits of prediction

To put it more generally, what limits prediction is complexity. The word complexity comes from the Latin plexus, which means intertwined. The components that make up a complex system are intertwined, and it's the interactions between them that determine what happens to them and how they behave.

Thus, studying elements of the system in isolation would probably yield misleading insights about them – as well as about the system as a whole.

Take, for example, a car traveling in a city. Knowing the speed at which it drives, it's theoretically possible to predict where it will end up at a particular time. But in real traffic, its speed will depend on interactions with other vehicles on the road. Since the details of these interactions emerge in the moment and cannot be known in advance, precisely predicting what happens to the the car is possible only a few minutes into the future.

Not with my health

These same principles apply to prescribing medications. Different conditions and diseases can have the same symptoms, and people with the same condition or disease may exhibit different symptoms. For example, fever can be caused by a respiratory illness or a digestive one. And a cold might cause cough, but not always.

This means that health care datasets have significant overlaps that would prevent AI from being error-free.

Certainly, humans also make errors. But when AI misdiagnoses a patient, as it surely will, the situation falls into a legal limbo. It's not clear who or what would be responsible if a patient were hurt. Pharmaceutical companies? Software developers? Insurance agencies? Pharmacies?

In many contexts, neither humans nor machines are the best option for a given task. "Centaurs," or "hybrid intelligence" – that is, a combination of humans and machines – tend to be better than each on their own. A doctor could certainly use AI to decide potential drugs to use for different patients, depending on their medical history, physiological details and genetic makeup. Researchers are already exploring this approach in precision medicine.

But common sense and the precautionary principle
suggest that it is too early for AI to prescribe drugs without human oversight. And the fact that mistakes may be baked into the technology could mean that where human health is at stake, human supervision will always be necessary.

Carlos Gershenson does not work for, consult, own shares in or receive funding from any company or organization that would benefit from this article, and has disclosed no relevant affiliations beyond their academic appointment.