The international journal of science / 14 March 2024

nature

Why scientists trust AI too much — and what to do about it

Some researchers see superhuman qualities in artificial intelligence. All scientists need to be alert to the risks this creates.

Scientists of all stripes are embracing artificial intelligence (AI) — from developing 'self-driving' laboratories, in which robots and algorithms work together to devise and conduct experiments, to replacing human participants in social-science experiments with bots[1].

Many downsides of AI systems have been discussed. For example, generative AI such as ChatGPT tends to make things up, or 'hallucinate' — and the workings of machine-learning systems are opaque.

In a Perspective article[2] published in *Nature* this week, social scientists say that AI systems pose a further risk: that researchers envision such tools as possessed of superhuman abilities when it comes to objectivity, productivity and understanding complex concepts. The authors argue that this put researchers in danger of overlooking the tools' limitations, such as the potential to narrow the focus of science or to lure users into thinking they understand a concept better than they actually do.

Scientists planning to use AI "must evaluate these risks now, while AI applications are still nascent, because they will be much more difficult to address if AI tools become deeply embedded in the research pipeline", write co-authors Lisa Messeri, an anthropologist at Yale University in New Haven, Connecticut, and Molly Crockett, a cognitive scientist at Princeton University in New Jersey.

The peer-reviewed article is a timely and disturbing warning about what could be lost if scientists embrace AI systems without thoroughly considering such hazards. It needs to be heeded by researchers and by those who set the direction and scope of research, including funders and journal editors. There are ways to mitigate the risks. But these require that the entire scientific community views AI systems with eyes wide open.

To inform their article, Messeri and Crockett examined around 100 peer-reviewed papers, preprints, conference proceedings and books, published mainly over the past five years. From these, they put together a picture of the ways in which scientists see AI systems as enhancing human capabilities.

In one 'vision', which they call AI as Oracle, researchers see AI tools as able to tirelessly read and digest scientific papers, and so survey the scientific literature more exhaustively than people can. In both Oracle and another vision, called AI as Arbiter, systems are perceived as evaluating scientific findings more objectively than do people, because they are less likely to cherry-pick the literature to support a desired hypothesis or to show favouritism in peer review. In a third vision, AI as Quant, AI tools seem to surpass the limits of the human mind in analysing vast and complex data sets. In the fourth, AI as Surrogate, AI tools simulate data that are too difficult or complex to obtain.

> "Researchers are in danger of overlooking AI tools' limitations."

Informed by anthropology and cognitive science, Messeri and Crockett predict risks that arise from these visions. One is the illusion of explanatory depth[3], in which people relying on another person — or, in this case, an algorithm — for knowledge have a tendency to mistake that knowledge for their own and think their understanding is deeper than it actually is.

Another risk is that research becomes skewed towards studying the kinds of thing that AI systems can test — the researchers call this the illusion of exploratory breadth. For example, in social science, the vision of AI as Surrogate could encourage experiments involving human behaviours that can be simulated by an AI — and discourage those on behaviours that cannot, such as anything that requires being embodied physically.

There's also the illusion of objectivity, in which researchers see AI systems as representing all possible viewpoints or not having a viewpoint. In fact, these tools reflect only the viewpoints found in the data they have been trained on, and are known to adopt the biases found in those data. "There's a risk that we forget that there are certain questions we just can't answer about human beings using AI tools," says Crockett. The illusion of objectivity is particularly worrying given the benefits of including diverse viewpoints in research.

Avoid the traps

If you're a scientist planning to use AI, you can reduce these dangers through a number of strategies. One is to map your proposed use to one of the visions, and consider which traps you are most likely to fall into. Another approach is to be deliberate about how you use AI. Deploying AI tools to save time on something your team already has expertise in is less risky than using them to provide expertise you just don't have, says Crockett.

Journal editors receiving submissions in which use of AI systems has been declared need to consider the risks posed by these visions of AI, too. So should funders reviewing grant applications, and institutions that want their researchers to use AI. Journals and funders should also keep tabs on the balance of research they are publishing and paying for — and ensure that, in the face of myriad AI possibilities, their portfolios remain broad in terms of the questions asked, the methods used and the viewpoints encompassed.

All members of the scientific community must view AI use not as inevitable for any particular task, nor as a panacea, but rather as a choice with risks and benefits that must be carefully weighed. For decades, and long before AI was a reality for most people, social scientists have studied AI. Everyone — including researchers of all kinds — must now listen.

[1] Grossmann, I. *et al. Science* **380**, 1108–1109 (2023).
[2] Messeri, L. & Crockett, M. J. *Nature* **627**, 49–58 (2024).
[3] Rozenblit, L. & Keil, F. *Cogn. Sci.* **26**, 521–562 (2002).