5 Conclusion

The Rise of AI Scientists

In late October 2025, I logged onto Zoom to watch a first-of-its-kind event. What I was about to see had all the trappings of a small scientific conference: Over the previous several months, young researchers had submitted papers to the event for consideration, and some of those researchers would soon have the opportunity to present their findings to the virtual audience. Under the surface, however, this event was entirely unlike any scientific conference I had ever attended. Many of the key players, reviewer and researcher alike, were AI systems.

The event, called Agents4Science, was devised by Stanford computer scientist James Zou and his team to test just how much science AI systems could now do.[1] Each project submitted to the conference was conducted primarily by AI, though human researchers were permitted to provide guidance, advice, and high-level ideas. (Humans also acted as presenters at the event.) Other AI systems reviewed each submission and determined whether or not it should be included in the conference, although it was human reviewers who decided which submissions were worthy of prizes.

The result was certainly interesting, though not groundbreaking. Much of the value of the event was in the insights that participants offered about the experiences of

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 10:44:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/5453C6AF35983DD157B1E6E0B2647683

collaborating with AI, as none of the individual AI-driven papers seemed likely to make waves in their respective fields. But the fact that the event worked at all suggests that science might be in for yet another revolution. Today, AI systems can write scientific papers. Soon they'll be able to write good ones.

It has been less than a decade since DeepMind released the first version of AlphaFold, and in that time AI has become nearly unrecognizable. Large language model-based tools such as Claude and ChatGPT do much more than predict protein structures: They scan the scientific literature for new ideas, write the code to test them, and then draft articles to communicate the results. Rather than making predictions in only a single domain, these language model systems can act as general-purpose scientific assistants. At their cores, they still run on deep neural networks, but the breadth of their capabilities makes them categorically different from specialized tools such as AlphaFold and AlexNet.

Nevertheless, studying how AI has changed science over the past decade may still be the best tool we have for anticipating what effects it may have in the future. Where specialized neural networks have started to peel prediction and understanding apart from each other, general-purpose AI systems have the potential to take that process much farther – or even separate prediction from understanding entirely.

Specialized networks have had a marked impact on fields such as protein folding and neuroscience. But unlike general-purpose systems, they won't ever be able to fully dissociate prediction and understanding. That's partly because

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 10:44:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/5453C6AF35983DD157B1E6E0B2647683

their scope is limited: Only some scientific domains offer enough data to effectively train deep neural networks. The more important reason, however, is that networks such as AlphaFold can't conduct entire scientific studies on their own. It's humans who have to decide when and why to run the networks and how to interpret the results – and to do that, they need a solid understanding of the scientific domain in which they are operating.

Large language model-based systems are different. They can identify current research gaps, propose projects to fill them, and execute those projects – even calling on resources such as AlphaFold when they are needed. If such systems ever become capable enough to rival human scientists, human understanding could become entirely unnecessary for achieving practical goals in science, because the AI systems could get there on their own.

We are certainly nowhere near that point yet. But AI companies are investing heavily in developing language model-based tools that can assist scientists in their work. In February of 2025, for example, Google introduced a system that it calls an "AI co-scientist," which is designed to generate hypotheses and detailed research plans in response to a scientific goal specified by the user. Some companies are also working to build systems that can conduct research independent of any human support or supervision. Sakana AI, a Tokyo-based startup, has built an "AI scientist" that can independently take a research project all the way from conceptualization to execution and publication. In 2025, a paper written by Sakana's AI scientist was accepted by a workshop at a major machine learning conference – not an

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 10:44:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/5453C6AF35983DD157B1E6E0B2647683

earth-shattering accomplishment, as the workshop had a 70 percent acceptance rate, but certainly a sign that the system's work was competent and coherent.

Scientists who try to use AI tools as scientific collaborators are often impressed. Gary Peltz, a Stanford geneticist, compared the experience of using Google's AI co-scientist to "consulting the oracle of Delphi" in a *Nature Medicine* article.[2] Matthew Schwartz, a physicist at Harvard, wrote in early 2026 of his experience shepherding Claude through a full research project and concluded that the tool was about as effective as a second-year PhD student – still in need of substantial guidance, but able to carry out meaningful scientific work independently.[3] Based on his observations, he recommended that today's human graduate students focus on learning to run experiments in the physical world, so that they won't be displaced from their jobs by machines – at least not until AI gets good at pipetting as well.

That said, most experts agree that AI scientists don't yet rival humans who are at the top of their fields. On the whole, AI systems can't yet match human "research taste," or the ability to identify promising questions and research directions – something that any good scientist needs. Research taste is typically something that scientists acquire over the course of their training: Graduate students are often told by their supervisors which problems they should pursue, but, as students gain more experience and independence, they also learn how to make those decisions for themselves. If Schwartz is right, AI systems may now be at that early graduate student phase – capable of carrying out research, but not choosing good research problems. But given how rapidly AI has progressed

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 10:44:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/5453C6AF35983DD157B1E6E0B2647683

over the past few years, it's entirely possible that, like talented students, they will soon learn to make astute choices about which scientific questions are the most promising.

If research taste does remain an insurmountable barrier for AI systems, then human scientists will still need to understand the phenomena they are studying in enough detail to ask AI the right questions – even if all they care about is serving practical goals. But if AI scientists do eventually rival human scientists for research taste, then we could theoretically be heading for a future in which the most cutting-edge and practically useful science is done by trillions of large language model-based systems humming away in remote data centers, with no need for sleep, food, or a vacation. In such a future, human scientific understanding would become entirely unnecessary, according to the practical view of science.

But superhuman AI scientists need not necessarily herald the end of scientific understanding. Even if we humans might not be pushing the boundaries of science ourselves, we still might be able to find ways to deepen our understanding of the universe. Perhaps superhuman scientific AI systems could act as teachers and communicate their new discoveries to us in some way – that's the hope at the heart of Been Kim's work with AlphaZero. Or perhaps we humans could study the new advancements and technologies that emerge from AI science ourselves, and, in doing so, learn indirectly from AI scientists.

This second possibility was the focus of a short story published in 2000 from the prescient science-fiction writer Ted Chiang, who also penned "Story of Your Life."[4] Titled

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 10:44:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/5453C6AF35983DD157B1E6E0B2647683

"Catching Crumbs from the Table," the story appeared in the scientific journal *Nature* and was written in the style of an editorial from some future version of that publication. In that future, gene editing and neurotechnology have advanced to the point that some humans, termed "metahumans," are incalculably more intelligent than their unedited peers. Metahumans make all the genuine scientific advancements, and human scientific journals are left to publish research that seeks to puzzle out what's going on in metahuman science. The scientists who study metahuman research aren't making novel discoveries but are instead only rediscovering what the metahumans have already found out – they are "catching crumbs from the table." But, much like it would be far easier for an early twentieth-century scientist to figure out how to build a computer if they were somehow able to study one in detail, the humans in Chiang's story probably understand the universe better than they would if it weren't for the metahumans. The existence of metahumans is arguably a net positive for human scientific understanding. Superhuman AI scientists could be as well.

It's impossible to say at this point whether scientific AI systems will ever achieve anything close to the intellectual heights of Chiang's metahumans, or even the heights of normal human scientists. But whether AI progress stalls here or advances far beyond what we can currently imagine, we will still need answers to tricky questions about the value of human scientific understanding. In some disciplines, such as protein folding, prediction and understanding are already well on their way to a full separation – and if scientific AI continues to advance, that will become a far more common state of affairs.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 10:44:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/5453C6AF35983DD157B1E6E0B2647683

So it is well worth starting to confront difficult questions about what understanding looks like without prediction. What is the value of understanding, when it is disconnected from practical goals? How much are we willing to invest in advancing scientific understanding for its own sake? And what if the resources spent on scientific understanding could instead go toward accelerating purely practical research? These questions don't have obvious answers, and people will surely disagree about how best to respond to them. But they are questions that we must confront if science is to maintain its humanity in the age of AI.

Downloaded from https://www.cambridge.org/core. UNIVERSIDAD AUTONOMA METROPOLITANA, on 05 Nov 2026 at 10:44:45, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://www.cambridge.org/core/product/5453C6AF35983DD157B1E6E0B2647683