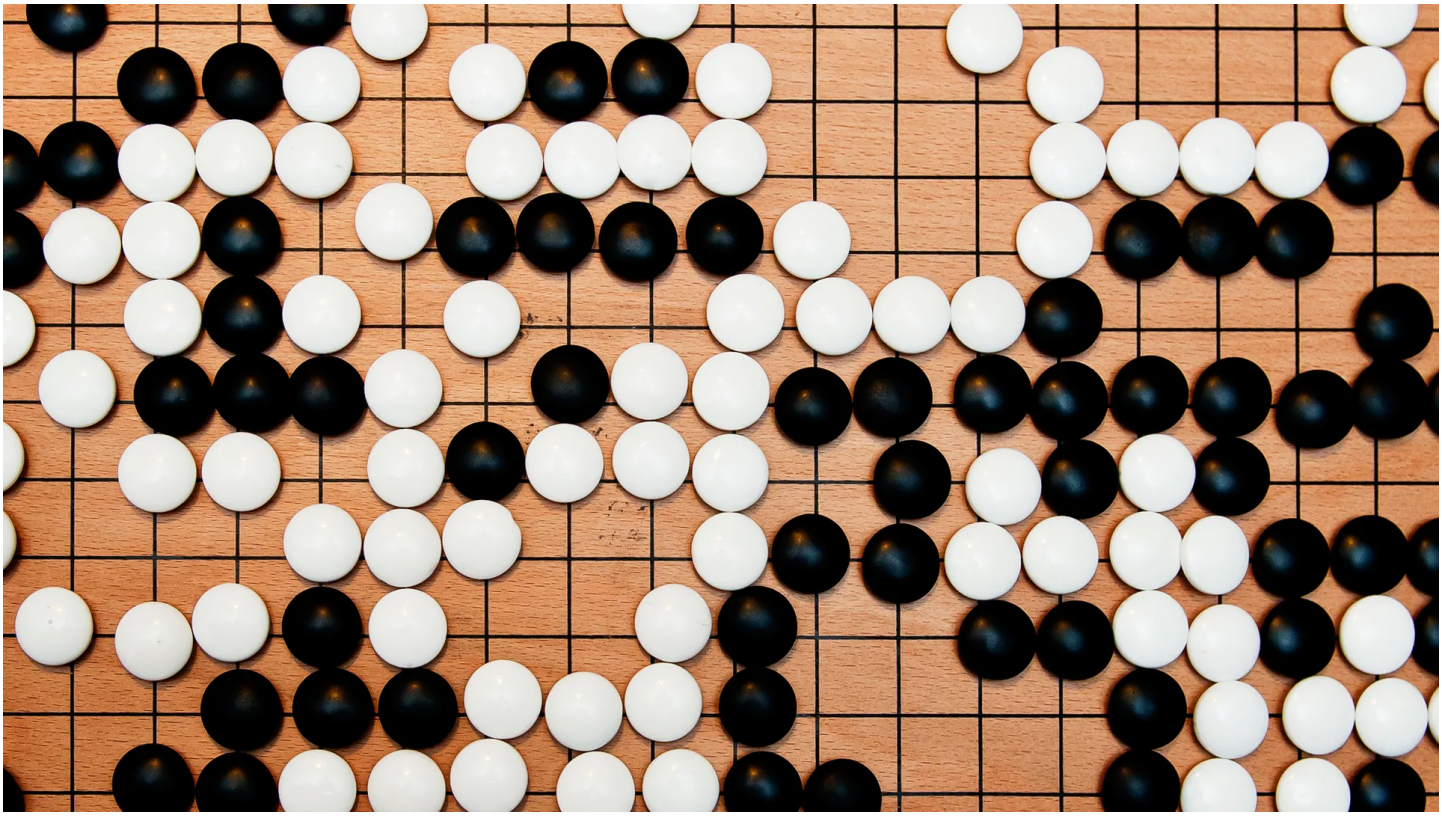
GREGORY BARBER BUSINESS SEP 16, 2019 7:00 AM

Artificial Intelligence Confronts a 'Reproducibility' Crisis

Machine-learning systems are black boxes even to the researchers that build them. That makes it hard for others to assess the results.



Facebook researchers said they found it "very difficult, if not impossible" to reproduce DeepMind's AlphaGo program. PHOTOGRAPH: GETTY IMAGES

The AI Database →

APPLICATION: ETHICS END USER: RESEARCH SECTOR: IT, RESEARCH

TECHNOLOGY: MACHINE LEARNING, NEURAL NETWORK

A FEW YEARS ago, Joelle Pineau, a computer science professor at McGill, was helping her students design a new algorithm when they fell into a rut. Her lab studies reinforcement learning, a type of artificial intelligence that's used, among

studies reinforcement learning, a type of artificial intelligence that's used, among other things, to help virtual characters ("half cheetah" and "ant" are popular)

teach themselves how to move about in virtual worlds. It's a prerequisite to building autonomous robots and cars. Pineau's students hoped to improve on another lab's system. But first they had to rebuild it, and their design, for reasons unknown, was falling short of its promised results. Until, that is, the students tried some "creative manipulations" that didn't appear in the other lab's paper.

Lo and behold, the system began performing as advertised. The lucky break was a symptom of a troubling trend, according to Pineau. Neural networks, the technique that's given us Go-mastering bots and text generators that craft classical Chinese poetry, are often called black boxes because of the mysteries of how they work. Getting them to perform well can be like an art, involving subtle tweaks that go unreported in publications. The networks also are growing larger and more complex, with huge data sets and massive computing arrays that make replicating and studying those models expensive, if not impossible for all but the best-funded labs.

"Is that even research anymore?" asks Anna Rogers, a machine-learning researcher at the University of Massachusetts. "It's not clear if you're demonstrating the superiority of your model or your budget."

Pineau is trying to change the standards. She's the reproducibility chair for NeurIPS, a premier artificial intelligence conference. Under her watch, the conference now asks researchers to submit a "reproducibility checklist" including items often omitted from papers, like the number of models trained before the "best" one was selected, the computing power used, and links to code and datasets. That's a change for a field where prestige rests on leaderboards—rankings that determine whose system is the "state of the art" for a particular task—and offers great incentive to gloss over the tribulations that led to those spectacular results.

The idea, Pineau says, is to encourage researchers to offer a road map for others to replicate their work. It's one thing to marvel at the eloquence of a new text generator or the "superhuman" agility of a videogame-playing bot. But even the most sophisticated researchers have little sense of how they work. Replicating those AI models is important not just for identifying new avenues of research, but also as a way to investigate algorithms as they augment, and in some cases supplant, human decision-making—everything from who stays in jail and for

supplant, human decision-making—everything from who stays in jail and for how long to who is approved for a mortgage.

LEARN MORE

The WIRED Guide to Artificial Intelligence

Others are also attacking the problem. Researchers at Google have proposed so-called "model cards" to detail how machine-learning systems have been tested, including results that point out potential bias. Others have sought to show how fragile the term "state of the art" is when systems, optimized for the data sets used in rankings, are set loose in other contexts. Last week, researchers at the Allen Institute for Artificial Intelligence, or AI2, released a paper that aims to expand Pineau's reproducibility checklist to other parts of the experimental process. They call it "Show Your Work."

"Starting where someone left off is such a pain because we never fully describe the experimental setup," says Jesse Dodge, an AI2 researcher who coauthored the research. "People can't reproduce what we did if we don't talk about what we did." It's a surprise, he adds, when people report even basic details about how a system was built. A survey of reinforcement learning papers last year found only about half included code.

Sometimes, basic information is missing because it's proprietary—an issue especially for industry labs. But it's more often a sign of the field's failure to keep up with changing methods, Dodge says. A decade ago, it was more straightforward to see what a researcher changed to improve their results. Neural networks, by comparison, are finicky; getting the best results often involves tuning thousands of little knobs, what Dodge calls a form of "black magic."

tuning thousands of little knobs, what Dodge calls a form of "black magic."

Picking the best model often requires a large number of experiments. The magic gets expensive, fast.

Even the big industrial labs, with the resources to design the largest, most complex systems, have signaled alarm. When Facebook attempted to replicate AlphaGo, the system developed by Alphabet's DeepMind to master the ancient game of Go, the researchers appeared exhausted by the task. The vast computational requirements—millions of experiments running on thousands of devices over days—combined with unavailable code, made the system "very difficult, if not impossible, to reproduce, study, improve upon, and extend," they wrote in a paper published in May. (The Facebook team ultimately succeeded.)

The AI2 research proposes a solution to that problem. The idea is to provide more data about the experiments that took place. You can still report the best model you obtained after, say, 100 experiments—the result that might be declared "state of the art"—but you also would report the range of performance you would expect if you only had the budget to try it 10 times, or just once.

The point of reproducibility, according to Dodge, isn't to replicate the results exactly. That would be nearly impossible given the natural randomness in neural networks and variations in hardware and code. Instead, the idea is to offer a road map to reach the same conclusions as the original research, especially when that involves deciding which machine-learning system is best for a particular task.

Keep Reading

That could help research become more efficient, Dodge explains. When his team rebuilt some popular machine-learning systems, they found that for some budgets, more antiquated methods made more sense than flashier ones. The idea is to help smaller academic labs by outlining how to get the best bang for their buck. A side benefit, he adds, is that the approach could encourage greener research, given that training large models can require as much energy as the lifetime emissions of a car.

Pineau says she's heartened to see others trying to "open up the models," but she's unsure whether most labs would take advantage of those cost-saving benefits. Many researchers would still feel pressure to use more computers to stay at the cutting edge, and then tackle efficiency later. It's also tricky to generalize how researchers should report their results, she adds. It's possible AI2's "show your work" approach could mask complexities in how researchers select the best models.

Those variations in methods are partly why the NeurIPS reproducibility checklist is voluntary. One stumbling block, especially for industrial labs, is proprietary code and data. If, say, Facebook is doing research with your Instagram photos, there's an issue with sharing that data publicly. Clinical research involving health data is another sticking point. "We don't want to move toward cutting off researchers from the community," she says.

It's difficult, in other words, to develop reproducibility standards that work without constraining researchers, especially as methods rapidly evolve. But

Pineau is optimistic. Another component of the NeurIPS reproducibility effort is a challenge that involves asking other researchers to replicate accepted papers. Compared with other fields, like the life sciences, where old methods die hard, the field is more open to putting researchers in those kinds of sensitive situations. "It's young both in terms of its people and its technology," she says. "There's less inertia to fight."

More Great WIRED Stories

- What is Wi-Fi 6, and when will I get it?
- These hallucinatory landscape photographs will blow your mind
- One scientist's quest to bring DNA sequencing to every sick kid
- We can be heroes: How nerds are reinventing pop culture
- Supermicro bug could let "virtual USBs" take over corporate servers
- 👁️ How do machines learn? Plus, read the latest news on artificial intelligence
- 🎧 Things not sounding right? Check out our favorite wireless headphones, soundbars, and bluetooth speakers

Gregory Barber is a staff writer at WIRED covering energy and the environment. He graduated from Columbia University with a bachelor's degree in computer science and English literature and now lives in San Francisco.

STAFF WRITER

TOPICS ARTIFICIAL INTELLIGENCE MACHINE LEARNING NEURAL NETWORKS

MORE FROM WIRED

Nick Thompson in conversation with Geoffrey Hinton

In this rare interview since (jointly) winning the 2018 Turing Award for his work on neural networks, hear about the conceptual and engineering breakthroughs that have made deep neural networks a critical element of computing. Their research has allowed artificial intelligence technologies to progress at a rate that was not possible in the past and has reinvented the way technology is built.