MS TECH | WILLIAMFALCON VIA GITHUB

ARTIFICIAL INTELLIGENCE

AI is wrestling with a replication crisis

Tech giants dominate research but the line between real breakthrough and product showcase can be fuzzy. Some scientists have had enough.

by **Will Douglas Heaven**

November 12, 2020

Last month Nature published a damning response written by 31 scientists to a study from Google Health that had appeared in the journal earlier this year. Google was describing successful trials of an AI that looked for signs of breast cancer in medical images. But according to its critics, the Google team provided so little information about its code and how it was tested that the study amounted to nothing more than a promotion of proprietary tech.

“We couldn’t take it anymore,” says Benjamin Haibe-Kains, the lead author of the response, who studies computational genomics at the University of Toronto. “It's not about this study in particular—it’s a trend we've been witnessing for multiple years now that has started to really bother us.”

Haibe-Kains and his colleagues are among a growing number of scientists pushing back against a perceived lack of transparency in AI research. “When we saw that paper from Google, we realized that it was yet another example of a very high-profile journal publishing a very exciting study that has nothing to do with science,” he says. “It's more an advertisement for cool technology. We can’t really do anything with it.”

Science is built on a bedrock of trust, which typically involves sharing enough details about how research is carried out to enable others to replicate it, verifying results for themselves. This is how science self-corrects and weeds out results that don’t stand up. Replication also allows others to build on those results, helping to advance the field. Science that can’t be replicated falls by the wayside.

At least, that’s the idea. In practice, few studies are fully replicated because most researchers are more interested in producing new results than reproducing old ones. But in fields like biology and physics—and computer science overall—researchers are typically expected to provide the information needed to rerun experiments, even if those reruns are rare.

Ambitious noob

AI is feeling the heat for several reasons. For a start, it is a newcomer. It has only really become an experimental science in the past decade, says Joelle Pineau, a computer scientist at Facebook AI Research and McGill University, who coauthored the complaint. “It used to be theoretical, but more and more we are running experiments,” she says. “And our dedication to sound methodology is lagging behind the ambition of our experiments.”

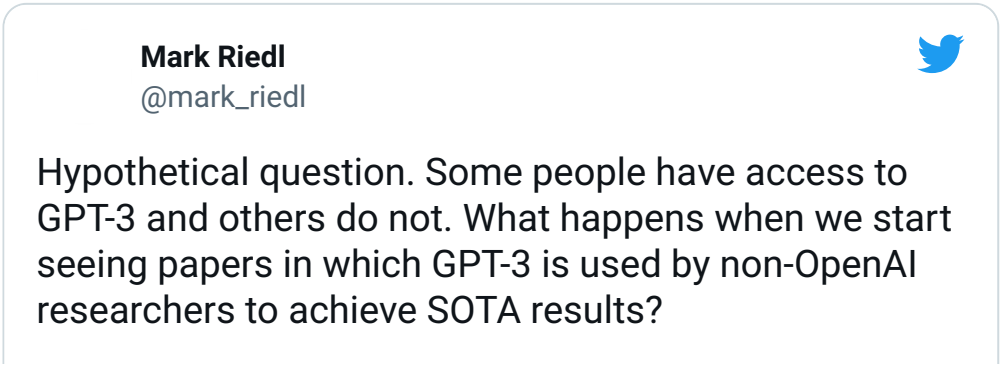
The problem is not simply academic. A lack of transparency prevents new AI models and techniques from being properly assessed for robustness, bias, and safety. AI moves quickly from research labs to real-world applications, with direct impact on people’s lives. But machine-learning models that work well in the lab can fail in the wild—with potentially dangerous consequences. Replication by different researchers in different settings would expose problems sooner, making AI stronger for everyone.

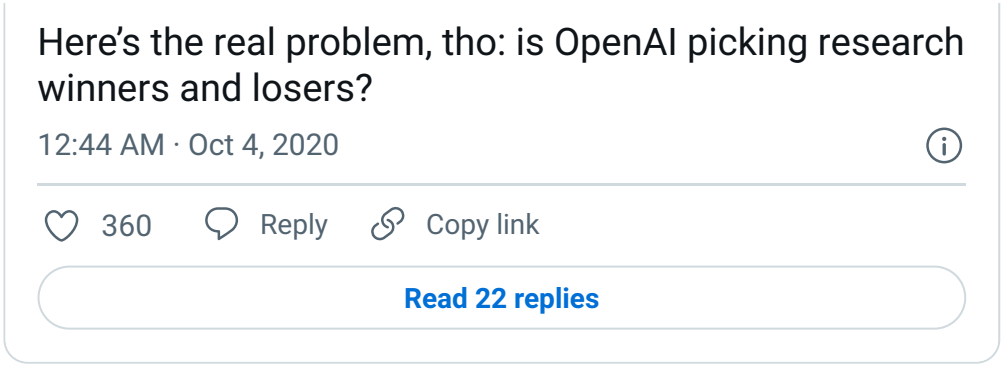
AI already suffers from the black-box problem: it can be impossible to say exactly how or why a machine-learning model produces the results it does. A lack of transparency in research makes things worse. Large models need as many eyes on them as possible, more people testing them and figuring out what makes them tick. This is how we make AI in health care safer, AI in policing more fair, and chatbots less hateful.

One thing that stops AI experiments from being replicated is a lack of access to the code. According to the 2020 State of AI report, a well-vetted annual analysis of the field by investors Nathan Benaich and Ian Hogarth, only 15% of AI studies share their code. Industry researchers are bigger offenders than those affiliated with universities. In particular, the report calls out OpenAI and DeepMind for keeping code under wraps.

Then there's the growing gulf between the haves and have-nots when it comes to the two pillars of AI, data and hardware. Data is often proprietary, such as the information Facebook collects on its users, or sensitive, as in the case of personal medical records. And tech giants carry out more and more research on enormous, expensive clusters of computers that few universities or smaller companies have the resources to access.

To take one example, training the language generator GPT-3 is estimated to have cost OpenAI $10 to $12 million—and that's just the final model, not including the cost of developing and training its prototypes. "You could probably multiply that figure by at least one or two orders of magnitude," says Benaich, who is founder of Air Street Capital, a VC firm that invests in AI startups. Only a tiny handful of big tech firms can afford to do that kind of work, he says: "Nobody else can just throw vast budgets at these experiments."





The rate of progress is dizzying, with thousands of papers published every year. But unless researchers know which ones to trust, it is hard for the field to move forward. Replication lets other researchers check that results have not been cherry-picked and that new AI techniques really do work as described. "It's getting harder and harder to tell which are reliable results and which are not," says Pineau.

What can be done? Like many AI researchers, Pineau divides her time between university and corporate labs. For the last few years, she has been the driving force behind a change in how AI research is published. For example, last year she helped introduce a checklist of things that researchers must provide, including code and detailed descriptions of experiments, when they submit papers to NeurIPS, one of the biggest AI conferences.

Replication is its own reward

Pineau has also helped launch a handful of reproducibility challenges, in which researchers try to replicate the results of published studies. Participants select papers that have been accepted to a conference and compete to rerun the experiments using the information provided. But the only prize is kudos.

This lack of incentive is a barrier to such efforts throughout the sciences, not just in AI. Replication is essential, but it isn't rewarded. One solution is to get students to do the work. For the last couple of years, Rosemary Ke, a PhD student at Mila, a research institute in Montreal founded by Yoshua Bengio, has organized a reproducibility challenge where students try to replicate studies submitted to NeurIPS as part of their machine-learning course. In turn, some successful replications are peer-reviewed and published in the journal ReScience.

"It takes quite a lot of effort to reproduce another paper from scratch," says Ke. "The reproducibility challenge recognizes this effort and gives credit to people who do a good job." Ke and others are also spreading the word at AI conferences via workshops set up to encourage researchers to make their work more transparent. This year Pineau and Ke extended the reproducibility challenge to seven of the top AI conferences, including ICML and ICLR.

Another push for transparency is the Papers with Code project, set up by AI researcher Robert Stojnic when he was at the University of Cambridge. (Stojnic is now a colleague of Pineau's at Facebook.) Launched as a stand-alone website where researchers could link a study to the code that went with it, this year Papers with Code started a collaboration with arXiv, a popular

preprint server. Since October, all machine-learning papers on arXiv have come with a Papers with Code section that links directly to code that authors wish to make available. The aim is to make sharing the norm.

Do such efforts make a difference? Pineau found that last year, when the checklist was introduced, the number of researchers including code with papers submitted to NeurIPS jumped from less than 50% to around 75%. Thousands of reviewers say they used the code to assess the submissions. And the number of participants in the reproducibility challenges is increasing.

Sweating the details

But it is only a start. Haibe-Kains points out that code alone is often not enough to rerun an experiment. Building AI models involves making many small changes—adding parameters here, adjusting values there. Any one of these can make the difference between a model working and not working. Without metadata describing how the models are trained and tuned, the code can be useless. “The devil really is in the detail,” he says.

It’s also not always clear exactly what code to share in the first place. Many labs use special software to run their models; sometimes this is proprietary. It is hard to know how much of that support code needs to be shared as well, says Haibe-Kains.

Pineau isn’t too worried about such obstacles. “We should have really high expectations for sharing code,” she says. Sharing data is trickier, but there are solutions here too. If researchers cannot share their data, they might give directions so that others can build similar data sets. Or you could have a process where a small number of independent auditors were given access to the data, verifying results for everybody else, says Haibe-Kains.

Hardware is the biggest problem. But DeepMind claims that big-ticket research like AlphaGo or GPT-3 has a trickle-down effect, where money spent by rich labs eventually leads to results

that benefit everyone. AI that is inaccessible to other researchers in its early stages, because it requires a lot of computing power, is often made more efficient—and thus more accessible—as it is developed. “AlphaGo Zero surpassed the original AlphaGo using far less computational resources,” says Koray Kavukcuoglu, vice president of research at DeepMind.

In theory, this means that even if replication is delayed, at least it is still possible. Kavukcuoglu notes that Gian-Carlo Pascutto, a Belgian coder at Mozilla who writes chess and Go software in his free time, was able to re-create a version of AlphaGo Zero called Leela Zero, using algorithms outlined by DeepMind in its papers. Pineau also thinks that flagship research like AlphaGo and GPT-3 is rare. The majority of AI research is run on computers that are available to the average lab, she says. And the problem is not unique to AI. Pineau and Benaich both point to particle physics, where some experiments can only be done on expensive pieces of equipment such as the Large Hadron Collider.

In physics, however, university labs run joint experiments on the LHC. Big AI experiments are typically carried out on hardware that is owned and controlled by companies. But even that is changing, says Pineau. For example, a group called Compute Canada is putting together computing clusters to let universities run large AI experiments. Some companies, including Facebook, also give universities limited access to their hardware. “It’s not completely there,” she says. “But some doors are opening.”

Haibe-Kains is less convinced. When he asked the Google Health team to share the code for its cancer-screening AI, he was told that it needed more testing. The team repeats this justification in a formal reply to Haibe-Kains's criticisms, also published in Nature: "We intend to subject our software to extensive testing before its use in a clinical environment, working alongside patients, providers and regulators to ensure efficacy and safety." The researchers also said they did not have permission to share all the medical data they were using.

It's not good enough, says Haibe-Kains: "If they want to build a product out of it, then I completely understand they won't disclose all the information." But he thinks that if you publish in a scientific journal or conference, you have a duty to release code that others can run. Sometimes that might mean sharing a version that is trained on less data or uses less expensive hardware. It might give worse results, but people will be able to tinker with it. "The boundaries between building a product versus doing research are getting fuzzier by the minute," says Haibe-Kains. "I think as a field we are going to lose."

Research habits die hard

If companies are going to be criticized for publishing, why do it at all? There's a degree of public relations, of course. But the main reason is that the best corporate labs are filled with researchers from universities. To some extent the culture at places like Facebook AI Research, DeepMind, and OpenAI is shaped by traditional academic habits. Tech companies also win by participating in the wider research community. All big AI projects at private labs are built on layers and layers of public research. And few AI researchers haven't made use of open-source machine-learning tools like Facebook's PyTorch or Google's TensorFlow.

As more research is done in house at giant tech companies, certain trade-offs between the competing demands of business and research will become inevitable. The question is how researchers navigate them. Haibe-Kains would like to see journals like Nature split what they publish into separate streams: reproducible studies on one hand and tech showcases on the other.

But Pineau is more optimistic. “I would not be working at Facebook if it did not have an open approach to research,” she says.

Other large corporate labs stress their commitment to transparency too. “Scientific work requires scrutiny and replication by others in the field,” says Kavukcuoglu. “This is a critical part of our approach to research at DeepMind.”

“OpenAI has grown into something very different from a traditional laboratory,” says Kayla Wood, a spokesperson for the company. “Naturally that raises some questions.” She notes that OpenAI works with more than 80 industry and academic organizations in the Partnership on AI to think about long-term publication norms for research.

Pineau believes there’s something to that. She thinks AI companies are demonstrating a third way to do research, somewhere between Haibe-Kains’s two streams. She contrasts the intellectual output of private AI labs with that of pharmaceutical companies, for example, which invest billions in drugs and keep much of the work behind closed doors.

The long-term impact of the practices introduced by Pineau and others remains to be seen. Will habits be changed for good? What difference will it make to AI’s uptake outside research? A lot hangs on the direction AI takes. The trend for ever larger models and data sets—favored by OpenAI, for example—will continue to make the cutting edge of AI inaccessible to most researchers. On the other hand, new techniques, such as model compression and few-shot learning, could reverse this trend and allow more researchers to work with smaller, more efficient AI.

Either way, AI research will still be dominated by large companies. If it’s done right, that doesn’t have to be a bad thing, says Pineau: “AI is changing the conversation about how industry research labs operate.” The key will be making sure the wider field gets the chance to participate. Because the trustworthiness of AI, on which so much depends, begins at the cutting edge.

by Will Douglas Heaven