IS AI FUELLING A REPRODUCIBILITY CRISIS IN SCIENCE?

'Data leakage' threatens the reliability of machine-learning use across disciplines, researchers warn.

By Elizabeth Gibney

From biomedicine to political sciences, researchers increasingly use machine learning as a tool to make predictions on the basis of patterns in their data. But the claims in many such studies are likely to be overblown, according to a pair of researchers at Princeton University in New Jersey. They want to sound an alarm about what they call a "brewing reproducibility crisis" in machine-learning-based sciences.

Machine learning is being sold as a tool that researchers can learn in a few hours and use by themselves — and many follow that advice, says Sayash Kapoor, a machine-learning researcher at Princeton. "But you wouldn't expect a chemist to be able to learn how to run a lab using an online course," says Kapoor, who has co-authored a preprint on the 'crisis'[1]. Peer reviewers do not have the time to scrutinize these models, so academia currently lacks mechanisms to root out irreproducible papers, he says. Kapoor and his co-author Arvind Narayanan created guidelines for scientists to avoid such pitfalls, including an explicit checklist to submit with each paper.

What is reproducibility?

Kapoor and Narayanan's definition of reproducibility is wide. It says that other teams should be able to replicate the results of a model, given the full details on data, code and conditions — often termed computational reproducibility, something that is already a concern for machine-learning scientists. The pair also define a model as irreproducible when researchers make errors in data analysis that mean that the model is not as predictive as claimed.

Judging such errors is subjective and often requires deep knowledge of the field in which machine learning is being applied. Some researchers whose work has been critiqued by the team disagree that their papers are flawed, or say Kapoor's claims are too strong. In social studies, for example, researchers have developed machine-learning models that aim to predict when a country is likely to slide into civil war. Kapoor and Narayanan claim that, once errors are corrected, these models perform no better than standard statistical techniques. But David Muchlinski, a political scientist at the Georgia Institute of Technology in Atlanta, whose paper[2] was examined by the pair, says that the field of conflict prediction has been unfairly maligned and that follow-up studies back up his work.

Still, the team's rallying cry has struck a chord. More than 1,200 people signed up to what was initially a planned as a small online workshop on reproducibility on 28 July, organized by Kapoor and colleagues, designed to come up with and disseminate solutions.

"I'm somewhat surprised that there hasn't been a crash in the legitimacy of machine learning already."

Unless the crisis is dealt with, machine learning's reputation could take a hit, says Momin Malik, a data scientist at the Mayo Clinic in Rochester, Minnesota, who spoke at the workshop. "I'm somewhat surprised that there hasn't been a crash in the legitimacy of machine learning already. But I think it could be coming very soon."

Kapoor and Narayanan say similar pitfalls occur in the application of machine learning to multiple sciences. The pair analysed 20 reviews in 17 research fields, and counted 329 research papers whose results could not be fully replicated because of problems in how machine learning was applied[1].

Narayanan himself is not immune: a 2015 paper on computer security that he co-authored[3] is among the 329. "It really is a problem that needs to be addressed collectively by this entire community," says Kapoor.

Machine-learning troubles

The failures are not the fault of any individual researcher, he adds. Instead, a combination of hype around AI and inadequate checks and balances is to blame. The most prominent issue that Kapoor and Narayanan highlight is 'data leakage', when information from the data set a model learns on includes data that it is later evaluated on. If these are not entirely separate, the model has effectively already seen the answers, and its predictions seem much better than they really are. The team has identified eight major types of data leakage that researchers can be vigilant against.

Some data leakage is subtle. For example, temporal leakage is when training data include points from later in time than the test data — which is a problem because the future depends on the past. As an example, Malik points to a 2011 paper[4] that claimed that a model analysing Twitter users' moods could predict the stock market's closing value with an accuracy of 87.6%. But because the team had tested the model's predictive power using data from a time period earlier than some of its training set, the algorithm had effectively been allowed to see the future, he says.

Wider issues include training models on data sets that are narrower than the population that they are ultimately intended to reflect, says Malik. For example, an AI that spots

K. H. FUNG/SPL

A CT scan of a tumour in human lungs. Researchers are experimenting with AI algorithms that can spot early signs of the disease.

pneumonia in chest X-rays that was trained only on older people might be less accurate on younger individuals. Another problem is that algorithms often end up relying on shortcuts that don't always hold, says Jessica Hullman, a computer scientist at Northwestern University in Evanston, Illinois, who spoke at the workshop. For example, a computer-vision algorithm might learn to recognize a cow by the grassy background in most cow images, so it would fail when it encounters an image of the animal on a mountain or beach.

The high accuracy of predictions in tests often fools people into thinking the models are picking up on the "true structure of the problem" in a human-like way, she says. The situation is similar to the replication crisis in psychology, in which people put too much trust in statistical methods, she adds.

Hype about machine learning's capabilities has played a part in making researchers accept their results too readily, says Kapoor. The word 'prediction' itself is problematic, says Malik, as most prediction is in fact tested retrospectively and has nothing to do with foretelling the future.

Fixing data leakage

Kapoor and Narayanan's solution to tackle data leakage is for researchers to include with their manuscripts evidence that their models don't have each of the eight types of leakage. The authors suggest a template for such documentation, which they call 'model info' sheets.

In the past three years, biomedicine has come far with a similar approach, says Xiao Liu, a clinical ophthalmologist at the University of Birmingham, UK. In 2019, Liu and her colleagues found that only 5% of more than 20,000 papers using AI for medical imaging were described in enough detail to discern whether they would work in a clinical environment[5].

Collaboration can also help, says Malik. He suggests studies involve both specialists in the relevant discipline and researchers in machine learning, statistics and survey sampling.

Fields in which machine learning finds leads for follow-up — such as drug discovery — are likely to benefit hugely from the technology, says Kapoor. But other areas will need more work to show it will be useful, he adds. Although machine learning is still relatively new to many fields, researchers must avoid the kind of crisis in confidence that followed the replication crisis in psychology a decade ago, he says.

1. Kapoor, S. & Narayanan, A. Preprint at https://arxiv.org/abs/2207.07048 (2022).
2. Muchlinski, D., Siroky, D., He, J. & Kocher, M. *Political Anal.* **24**, 87–103 (2016).
3. Caliskan-Islam, A. *et al.* In *Proc. 24th USENIX Security Symposium* 255–270 (USENIX Association, 2015).
4. Bollen, J., Mao, H. & Zeng, X. *J. Comp. Sci.* **2**, 1–8 (2011).
5. Liu, X. *et al. Lancet Digit. Health* **1**, e271–e297 (2019).

ANGEL GARCIA/BLOOMBERG/GETTY

Dairy farming emerged in Europe before humans were able to drink milk without ill effects.

HOW HUMANS EVOLVED THE ABILITY TO DIGEST MILK

Landmark study shows that famine and disease shaped lactose tolerance.

By Ewen Callaway

The dawn of dairy farming in Europe occurred thousands of years before most people there evolved the ability to drink milk as adults without becoming ill. Now researchers think they know why: lactose tolerance was beneficial enough to influence evolution only during occasional episodes of famine and disease, so it took thousands of years for the trait to become widespread (R. P. Evershed *et al. Nature* https://doi.org/h6fn; 2022).

The theory — backed by an analysis of thousands of pottery shards and hundreds of ancient human genomes, as well as sophisticated modelling — explains how the ability to digest milk became so common in modern Europeans, despite being almost non-existent in early dairy farmers. This ability, known as lactase persistence, comes from an enzyme that breaks down milk sugar and usually shuts down after young children are weaned.

The study, published in *Nature* on 27 July, is the first major effort to quantify the forces that have shaped lactase persistence, says Shevan Wilkin, a molecular archaeologist at the University of Zurich, Switzerland. "Lactase-persistence evolution was much more complicated than we ever thought."

The ability to digest milk evolved independently in ancient populations around the world. Researchers have mapped the trait to gene variants that instruct cells to produce high levels of lactase. A variant that most people of European ancestry carry is one of the strongest examples of natural selection on the human genome.

Yet scientists have struggled to explain the forces underlying the high prevalence of lactase persistence in Europe. Many had presumed that the variation proved beneficial only after ancient peoples started routinely consuming dairy products. Another influential theory held that cows, goats and sheep — domesticated around 10,000–12,000 years ago — were at first kept mainly for their meat, and that milk consumption followed millennia later.

But Richard Evershed, a biogeochemist at the University of Bristol, UK, who co-led the latest study, and his team have found milk-fat residues on potsherds dating from the dawn of animal domestication. Ancient-genomics studies showed that the early animal farmers were lactose intolerant, and that tolerance for milk did not become common in Europe until after the Bronze Age, 5,000–4,000 years ago.

To determine the probable forces behind