ILLUSTRATION BY ADRIÀ VOLTÀ

IS AI LEADING TO A REPRODUCIBILITY CRISIS IN SCIENCE?

Scientists worry that ill-informed use of artificial intelligence is driving a deluge of unreliable or useless research. **By Philip Ball**

During the COVID-19 pandemic in late 2020, testing kits for the viral infection were scant in some countries. So the idea of diagnosing infection with a medical technique that was already widespread — chest X-rays — sounded appealing. Although the human eye can't reliably discern differences between infected and non-infected individuals, a team in India reported that artificial intelligence (AI) could do it, using machine learning to analyse a set of X-ray images[1].

The paper — one of dozens of studies on the idea — has been cited more than 900 times. But the following September, computer scientists Sanchari Dhar and Lior Shamir at Kansas State University in Manhattan took a closer look[2]. They trained a machine-learning algorithm on the same images, but used only blank background sections that showed no body parts at all. Yet their AI could still pick out COVID-19 cases at well above chance level.

The problem seemed to be that there were consistent differences in the backgrounds of the medical images in the data set. An AI system could pick up on those artefacts to succeed in the diagnostic task, without learning any clinically relevant features — making it medically useless.

Shamir and Dhar found several other cases in which a reportedly successful image classification by AI — from cell types to face recognition — returned similar results from blank or meaningless parts of the images. The algorithms performed better than chance at recognizing faces without faces, and cells without cells. Some of these papers have been cited hundreds of times.

"These examples might be amusing", Shamir says — but in biomedicine, misclassification could be a matter of life and death. "The problem is extremely common — a lot more common than most of my colleagues would want to believe." A separate review in 2021 examined 62 studies using machine learning to diagnose COVID-19 from chest X-rays or computed tomography scans; it concluded that none of the AI models was clinically useful, because of methodological flaws or biases in image data sets[3].

The errors that Shamir and Dhar found are just some of the ways in which machine learning can give rise to misleading claims in research. Computer scientists Sayash Kapoor and Arvind Narayanan at Princeton University in New Jersey reported earlier this year that the problem of data leakage (when there is insufficient separation between the data used to train an AI system and those used to test it) has caused reproducibility issues in 17 fields that they examined, affecting hundreds of papers[4]. They argue that naive use of AI is leading to a reproducibility crisis.

Machine learning (ML) and other types of AI are powerful statistical tools that have advanced almost every area of science by picking out patterns in data that are often invisible to human researchers. At the same time, some researchers worry that ill-informed use of AI software is driving a deluge of papers with claims that cannot be replicated, or that are wrong or useless in practical terms.

There has been no systematic estimate of the extent of the problem, but researchers say that, anecdotally, error-strewn AI papers are everywhere. "This is a widespread issue impacting many communities beginning to adopt machine-learning methods," Kapoor says.

> "I SEE A LOT OF COMMON MISTAKES REPEATED OVER AND OVER."

Aeronautical engineer Lorena Barba at George Washington University in Washington DC agrees that few, if any, fields are exempt from the issue. "I'm confident stating that scientific machine learning in the physical sciences is presenting widespread problems," she says. "And this is not about lots of poor-quality or low-impact papers," she adds. "I have read many articles in prestigious journals and conferences that compare with weak baselines, exaggerate claims, fail to report full computational costs, completely ignore limitations of the work, or otherwise fail to provide sufficient information, data or code to reproduce the results."

"There is a proper way to apply ML to test a scientific hypothesis, and many scientists were never really trained properly to do that because the field is still relatively new," says Casey Bennett at DePaul University in Chicago, Illinois, a specialist in the use of computer methods in health. "I see a lot of common mistakes repeated over and over," he says. For ML tools used in health research, he adds, "it's like the Wild West right now."

How AI goes astray

As with any powerful new statistical technique, AI systems can make it easy for researchers looking for a particular result to fool themselves. "AI provides a tool that allows researchers to 'play' with the data and parameters until the results are aligned with the expectations," says Shamir.

"The incredible flexibility and tunability of AI, and the lack of rigour in developing these models, provide way too much latitude," says computer scientist Benjamin Haibe-Kains at the University of Toronto, Canada, whose lab applies computational methods to cancer research.

Data leakage seems to be particularly common, according to Kapoor and Narayanan, who have laid out a taxonomy of such problems[4]. ML algorithms are trained on data until they can reliably produce the right outputs for each input — to correctly classify an image, say. Their performance is then evaluated on an unseen (test) data set. As ML experts know, it is essential to keep the training set separate from the test set. But some researchers apparently don't know how to ensure this.

The issue can be subtle: if a random subset of test data is taken from the same pool as the training data, that could lead to leakage. And if medical data from the same individual (or same scientific instrument) are split between training and test sets, the AI might learn to identify features associated with that individual or that instrument, rather than a specific medical ailment — a problem identified, for example, in one use of AI to analyse histopathology images[5]. That's why it is essential to run 'control' trials on blank backgrounds of images, Shamir says, to see if what the algorithm is generating makes logical sense.

Kapoor and Narayanan also raise the problem of when the test set doesn't reflect real-world data. In this case, a method might give reliable and valid results on its test data, but that can't be reproduced in the real world.

"There is way more variation in the real world than in the lab, and the AI models are often not tested for it until we deploy them," Haibe-Kains says.

In one example, an AI developed by researchers at Google Health in Palo Alto, California, was used to analyse retinal images for signs of diabetic retinopathy, which can cause blindness. When others in the Google Health team trialled it in clinics in Thailand, it rejected many images taken under suboptimal conditions, because the system had been trained on high-quality scans. The high rejection rate created a need for more follow-up appointments with patients — an unnecessary workload[6].

Efforts to correct training or test data sets can lead to their own problems. If the data are imbalanced — that is, they don't sample the real-world distribution evenly — researchers might apply rebalancing algorithms, such as the Synthetic Minority Oversampling Technique (SMOTE)[7], which generates synthetic data for under-sampled regions.

However, Bennett says, "in situations when the data is heavily imbalanced, SMOTE will lead to overly optimistic estimates of performance, because you are essentially creating lots of 'fake data' based on an untestable assumption about the underlying data distribution". In other words, SMOTE ends up not so much balancing as manufacturing the data set, which is then pervaded with the same biases

that are inherent in the original data.

Even experts can find it hard to escape these problems. In 2022, for instance, data scientist Gaël Varoquaux at the French National Institute for Research in Digital Science and Technology (INRIA) in Paris and his colleagues ran an international challenge for teams to develop algorithms that could make accurate diagnoses of autism spectrum disorder from brain-structure data obtained by magnetic resonance imaging (MRI)[8].

The challenge garnered 589 submissions from 61 teams, and the 10 best algorithms (mostly using ML) seemed to perform better using MRI data compared with the existing method of diagnosis, which uses genotypes. But those algorithms did not generalize well to another data set that had been kept private from the public data given to teams to train and test their models. "The best predictions on the public dataset were too good to be true, and did not carry over to the unseen, private dataset," the researchers wrote[8]. In essence, this is because developing and testing a method on a small data set, even when trying to avoid data leakage, will always end up overfitting to those data, Varoquaux says – that is, being too closely focused on aligning to the particular patterns in the data so that the method loses generality.

Overcoming the problem

This August, Kapoor, Narayanan and their co-workers proposed a way to tackle the issue with a checklist of standards for reporting AI-based science[9], which runs to 32 questions on factors such as data quality, details of modelling and risks of data leakage. They say their list "provides a cross-disciplinary bar for reporting standards in ML-based science". Other checklists have been created for specific fields, such as for the life sciences[10] and chemistry[11].

Many argue that research papers using AI should make their methods and data fully open. A 2019 study by data scientist Edward Raff at the Virginia-based analytics firm Booz Allen Hamilton found that only 63.5% of 255 papers using AI methods could be reproduced as reported[12], but computer scientist Joelle Pineau at McGill University in Montreal, Canada (who is also vice-president of AI research at Meta) and others later stated that reproducibility rises to 85% if the original authors help with those efforts by actively supplying data and code[13]. With that in mind, Pineau and her colleagues proposed a protocol for papers that use AI methods, which specifies that the source code be included with the submission and that – as with Kapoor and Narayan's recommendations – it be assessed against a standardized ML reproducibility checklist[13].

But researchers note that providing enough details for full reproducibility is hard in any computational science, let alone in AI.

And checklists can only achieve so much. Reproducibility doesn't guarantee that the model is giving correct results, but only self-consistent ones, warns computer scientist Joaquin Vanschoren at the Eindhoven University of Technology in the Netherlands. He also points out that "a lot of the really high-impact AI models are created by big companies, who seldom make their codes available, at least immediately." And, he says, sometimes people are reluctant to release their own code because they don't think it is ready for public scrutiny.

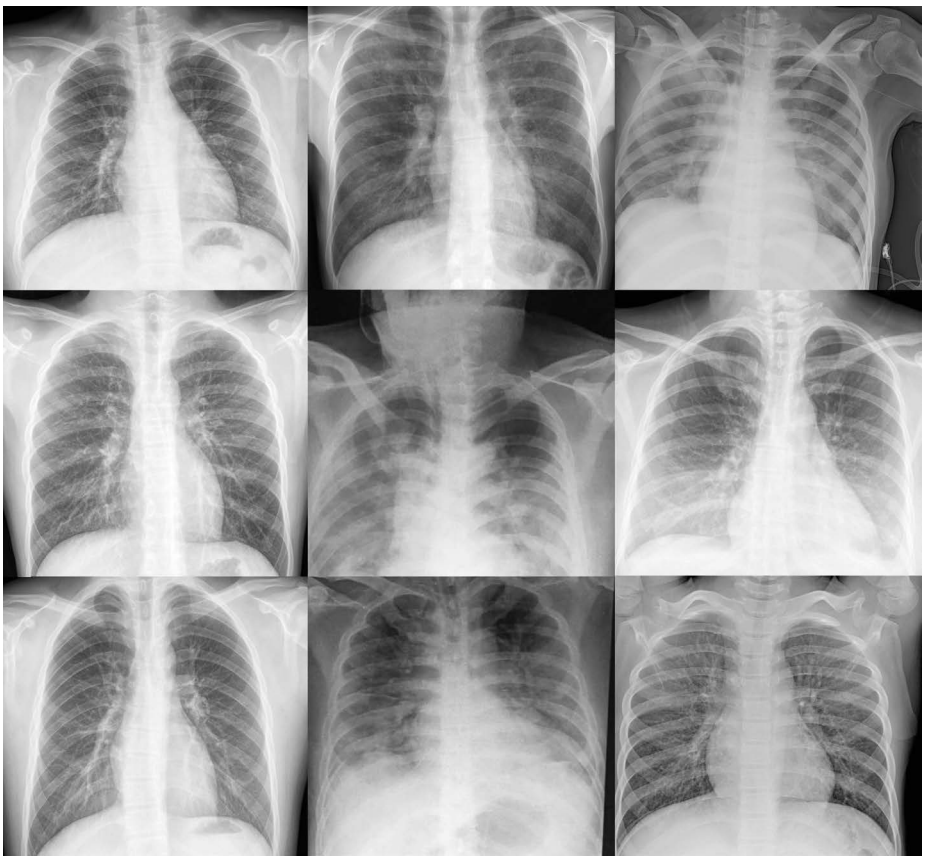
> "WE COULD SEE A GREATER AMOUNT OF INTEGRITY ISSUES IN SCIENCE."

Although some computer-science conferences require that code be made available to have a peer-reviewed proceedings paper published, this is not yet universal. "The most important conferences are more serious about it, but it's a mixed bag," says Vanschoren.

Part of the problem could be that there simply are not enough data available to properly test the models. "If there aren't enough public data sets, then researchers can't evaluate their models correctly and end up publishing low-quality results that show great performance," says Joseph Cohen, a scientist at Amazon AWS Health AI, who also directs the US-based non-profit Institute for Reproducible Research. "This issue is very bad in medical research."

The pitfalls might be all the more hazardous for generative AI systems such as large language models (LLMs), which can create new data, including text and images, using models derived from their training data. Researchers can use such algorithms to enhance the resolution of images, for instance. But unless they take great care, they could end up introducing artefacts, says Viren Jain, a research scientist at Google in Mountain View, California, who works on developing AI for visualizing and manipulating large data sets.

"There has been a lot of interest in the microscopy world to improve the quality of images, like removing noise," he says. "But I wouldn't say these things are foolproof, and they could be introducing artefacts." He has seen such dangers in his own work on images of brain tissue. "If we weren't careful to take the proper steps to validate things, we could have easily



HEALTHY AND PNEUMONIA: D. KERMANY *ET AL./CELL* (CC BY 4.0); COVID-19: E. M. EDRADA *ET AL./TROP. MED. HEALTH* (CC BY 4.0).

Chest X-ray images of healthy people (left); those with COVID-19 (centre); and those with pneumonia (right).

done something that ended up inadvertently prompting an incorrect scientific conclusion."

Jain is also concerned about the possibility of deliberate misuse of generative AI as an easy way to create genuine-seeming scientific images. "It's hard to avoid the concern that we could see a greater amount of integrity issues in science," he says.

Culture shift

Some researchers think that the problems will only be truly addressed by changing cultural norms about how data are presented and reported. Haibe-Kains is not very optimistic that such a change will be easy to engineer. In 2020, he and his colleagues criticized a prominent study on the potential of ML for detecting breast cancer in mammograms, authored by a team that included researchers at Google Health[14]. Haibe-Kains and his co-authors wrote that "the absence of sufficiently documented methods and computer code underlying the study effectively undermines its scientific value"[15] — in other words, the work could not be examined because there wasn't enough information to reproduce it.

The authors of that study said in a published response, however, that they were not at liberty to share all the information, because some of it came from a US hospital that had privacy concerns with making it available. They added that they "strove to document all relevant machine learning methods while keeping the paper accessible to a clinical and general scientific audience"[16].

More widely, Varoquaux and computer scientist Veronika Cheplygina at the IT University of Copenhagen have argued that current publishing incentives, especially the pressure to generate attention-grabbing headlines, act against the reliability of AI-based findings[17]. Haibe-Kains adds that authors do not always "play the game in good faith" by complying with data-transparency guidelines, and that journal editors often don't push back enough against this.

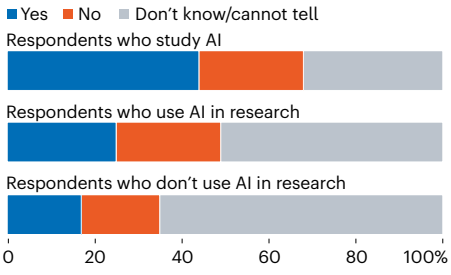
The problem is not so much that editors waive rules about transparency, Haibe-Kains argues, but that editors and reviewers might be "poorly educated on the real versus fictitious obstacles for sharing data, code and so on, so they tend to be content with very shallow, unreasonable justifications [for not sharing such information]". Indeed, authors might simply not understand what is required of them to ensure the reliability and reproducibility of their work. "It's hard to be completely transparent if you don't fully understand what you are doing," says Bennett.

In a *Nature* survey this year that asked more than 1,600 researchers about AI, views on the adequacy of peer review for AI-related journal articles were split. Among the scientists who used AI for their work, one-quarter thought reviews were adequate, one-quarter felt they were not and around half said they didn't know (see 'Quality of AI review in research papers' and *Nature* **621**, 672–675; 2023).

QUALITY OF AI REVIEW IN RESEARCH PAPERS

A *Nature* survey of more than 1,600 scientists found split opinions on the quality of peer-review of research papers that use AI.

Q: Do you think that journal editors and peer-reviewers, in general, can adequately review papers in your field that use AI?



SOURCE: *NATURE* **621**, 672–675 (2023).

Although plenty of potential problems have been raised about individual papers, they rarely seem to get resolved. Individual cases tend to get bogged down in counterclaims and disputes about fine details. For example, in some of the case studies investigated by Kapoor and Narayanan, involving uses of ML to predict outbreaks of civil war, some of their claims that the results were distorted by data leakage were met with public rebuttals by the authors (see *Nature* **608**, 250–251; 2022). And the authors of the study on COVID-19 identification from chest X-rays[1] critiqued by Dhar and Shamir told *Nature* that they do not accept the criticisms.

Learning to fly

Not everyone thinks there is an AI crisis looming. "In my experience, I have not seen the application of AI resulting in an increase in irreproducible results," says neuroscientist Lucas Stetzik at Aiforia Technologies, a Helsinki-based consultancy for AI-based medical imaging. Indeed, he thinks that, carefully applied, AI techniques can help to eliminate the cognitive biases that often leak into researchers' work. "I was drawn to AI specifically because I was frustrated by the irreproducibility of many methods and the ease with which some irresponsible researchers can bias or cherry-pick results."

Although concerns about the validity or reliability of many published findings on the uses of AI are widespread, it is not clear that faulty or unreliable findings based on AI in the scientific literature are yet creating real dangers of, say, misdiagnosis in clinical practice. "I think that has the potential to happen, and I would not be shocked to find out it is already happening, but I haven't seen any such reports yet," says Bennett.

Cohen also feels that the issues might resolve themselves, just as teething problems with other new scientific methods have. "I think that things will just naturally work out in the end," he says. "Authors who publish poor-quality papers will be regarded poorly by the research community and not get future jobs. Journals that publish these papers will be regarded as untrustworthy and good authors won't want to publish in them."

Bioengineer Alex Trevino at the bioinformatics company Enable Medicine in Menlo Park, California, says that one key aspect of making AI-based research more reliable is to ensure that it is done in interdisciplinary teams. For example, computer scientists who understand how to curate and handle data sets should work with biologists who understand the experimental complexities of how the data were obtained.

Bennett thinks that, in a decade or two, researchers will have a more sophisticated understanding of what AI can offer and how to use it, much as it took biologists that long to better understand how to relate genetic analyses to complex diseases. And Jain says that, at least for generative AI, reproducibility might improve when there is greater consistency in the models being used. "People are increasingly converging around foundation models: very general models that do lots of things, like OpenAI's GPT-3 and GPT-4," he says. That is much more likely to give rise to reproducible results than some bespoke model trained in-house. "So you could imagine reproducibility getting a bit better if everyone is using the same systems."

Vanschoren draws a hopeful analogy with the aerospace industry. "In the early days it was very dangerous, and it took decades of engineering to make airplanes trustworthy." He thinks that AI will develop in a similar way: "The field will become more mature and, over time, we will learn which systems we can trust." The question is whether the research community can contain the problems in the meantime.

Philip Ball is a science writer in London.

1. Khan, A. I., Shah, J. L. & Bhat, M. M. *Comput. Methods Prog. Biomed.* **196**, 105581 (2020).
2. Dhar, S. & Shamir, L. *Vis. Inform.* **5**, 92–101 (2021).
3. Roberts, M *et al. Nature Mach. Intell.* **3**, 199–217 (2021).
4. Kapoor, S. & Narayanan, A. *Patterns* **4**, 100804 (2023).
5. Oner, M. U., Cheng, Y.-C., Lee, H.K. & Sung, W.-K. Preprint at medRxiv https://doi.org/10.1101/2020.04.23.20076406 (2020).
6. Beede, E. *et al. in Proc. 2020 CHI Conf. Human Factors Comput. Syst.* https://doi.org/10.1145/3313831.3376718 (2020).
7. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. *J. Artif. Intell. Res.* 16, 321–357 (2002).
8. Traut, N. *et al. NeuroImage* **255**, 119171 (2022).
9. Kapoor, S. *et al.* Preprint at https://arxiv.org/abs/2308.07832 (2023).
10. Heil, B. J. *et al. Nature Methods* **18**, 1132–1135 (2021).
11. Artrith, N. *et al. Nature Chem.* **13**, 505–508 (2021).
12. Raff, E. Preprint at https://arxiv.org/abs/1909.06674 (2019).
13. Pineau, J. *et al. J. Mach. Learn. Res.* **22**, 7459–7478 (2021).
14. McKinney, S. M. *et al. Nature* **577**, 89–94 (2020).
15. Haibe-Kains, B. *et al. Nature* **586**, E14–E16 (2020).
16. McKinney, S. M. *et al. Nature* **586**, E17–E18 (2020).
17. Varoquaux, G. & Cheplygina, V. *npj Digit. Med.* **5**, 48 (2022).