immediately produced a black hole. However, the researchers argue that the merger evolved along a more complicated—and revealing—path that delayed that collapse.

As the neutron stars spiraled into each other, gravitational-wave detectors in the United States and Italy sensed ripples in space generated by the whirling bodies. The waves allowed physicists to peg their combined mass at 2.73 solar masses. Two seconds after the gravitational waves, orbiting telescopes detected a powerful, short gamma ray burst. Telescopes on Earth spotted the event's afterglow, which faded over several days from bright blue to dimmer red.

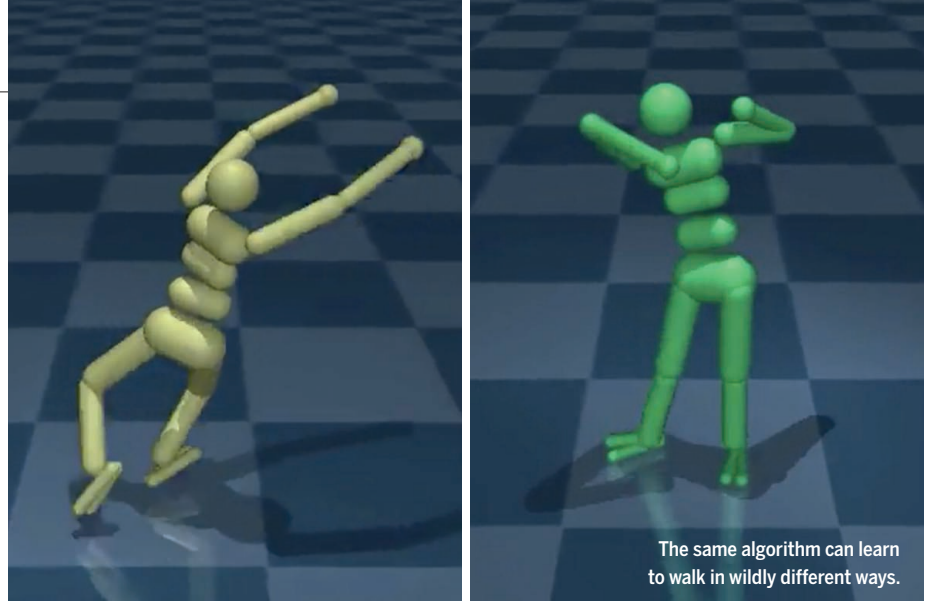
Together, the clues suggest the merger first produced a spinning, overweight neutron star momentarily propped up by centrifugal force. The afterglow shows that the merger spewed between 0.1 and 0.2 solar masses of newly formed radioactive elements into space, more than could have escaped from a black hole. The ejected material's initial blue tint shows that at first, it lacked heavy elements called lanthanides. A flux of particles called neutrinos presumably slowed those elements' formation, and a neutron star radiates copious neutrinos. The short gamma ray burst, the supposed birth cry of a black hole, indicates that the merged neutron star collapsed in seconds.

To derive their mass limits, the teams dove into the details of the spinning neutron star. They generally argue that at first the outer layers of the merged neutron star likely spun faster than its center. Then it flung off material and slowed to form a rigid spinning body whose mass researchers could calculate from the masses of the original neutron stars minus the ejected material. The fact that this spinning neutron star survived only momentarily suggests that its mass was close to the limit for such a spinner.

That last inference is essential, Rezzolla says. Theory suggests that the mass of a rigidly spinning neutron star can exceed that of a stationary one by up to 18%, he says. That scaling allows researchers to infer the maximum mass of a stationary, stable neutron star. The whole argument works because the initial neutron stars weren't so massive that they immediately produced a black hole or so light that they produced a spinning neutron star that lingered longer, Shibata says. "This was a very lucky event," he says.

The analyses are persuasive, Lattimer says, although he quibbles with the precision implied in numbers such as 2.17 solar masses. "If you say 2.2 plus or minus a 10th, I would think it gets the same message across." ■

IMAGES: YUVAL TASSA



The same algorithm can learn to walk in wildly different ways.

COMPUTER SCIENCE

Artificial intelligence faces reproducibility crisis

Unpublished code and sensitivity to training conditions make many claims hard to verify

By **Matthew Hutson**

Last year, computer scientists at the University of Montreal (U of M) in Canada were eager to show off a new speech recognition algorithm, and they wanted to compare it to a benchmark, an algorithm from a well-known scientist. The only problem: The benchmark's source code wasn't published. The researchers had to recreate it from the published description. But they couldn't get their version to match the benchmark's claimed performance, says Nan Rosemary Ke, a Ph.D. student in the U of M lab. "We tried for 2 months and we couldn't get anywhere close."

The booming field of artificial intelligence (AI) is grappling with a replication crisis, much like the ones that have afflicted psychology, medicine, and other fields over the past decade. AI researchers have found it difficult to reproduce many key results, and that is leading to a new conscientiousness about research methods and publication protocols. "I think people outside the field might assume that because we have code, reproducibility is kind of guaranteed," says Nicolas Rougier, a computational neuroscientist at France's National Institute for Research in Computer Science and Automation in Bordeaux. "Far from it." Last week, at a meeting of the Association for the Advancement of Artificial Intelligence (AAAI) in New Orleans, Louisiana, reproducibility was on the agenda, with some teams diagnosing the problem—and one laying out tools to mitigate it.

The most basic problem is that researchers often don't share their source code. At the AAAI meeting, Odd Erik Gundersen, a computer scientist at the Norwegian University of Science and Technology in Trondheim, reported the results of a survey of 400 algorithms presented in papers at two top AI conferences in the past few years. He found that only 6% of the presenters shared the algorithm's code. Only a third shared the data they tested their algorithms on, and just half shared "pseudocode"—a limited summary of an algorithm. (In many cases, code is also absent from AI papers published in journals, including *Science* and *Nature*.)

Researchers say there are many reasons for the missing details: The code might be a work in progress, owned by a company, or held tightly by a researcher eager to stay ahead of the competition. It might be dependent on other code, itself unpublished. Or it might be that the code is simply lost, on a crashed disk or stolen laptop—what Rougier calls the "my dog ate my program" problem.

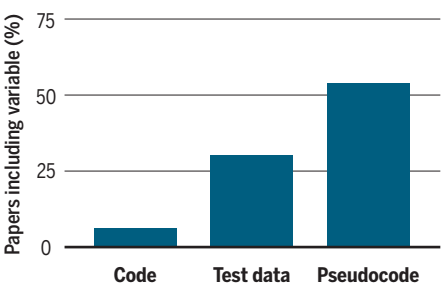
Assuming you can get and run the original code, it still might not do what you expect. In the area of AI called machine learning, in which computers derive expertise from experience, the training data for an algorithm can influence its performance.

Ke suspects that not knowing the training for the speech-recognition benchmark was what tripped up her group. "There's randomness from one run to another," she says. You can get "really, really lucky and have one run with a really good number," she adds. "That's usually what people report."

At the AAAI meeting, Peter Henderson, a computer scientist at McGill University in Montreal, showed that the performance of AIs designed to learn by trial and error is highly sensitive not only to the exact code used, but also to the random numbers generated to kick off training, and to "hyperparameters"—settings that are not core to the algorithm but that affect how quickly it learns. He ran several of these "reinforcement learning" algorithms under different conditions and found wildly different results. For example, a virtual "half-cheetah"—a stick figure used in motion algorithms—could learn to sprint in one test but would flail around on the floor in another. Henderson says researchers should document more of these key details. "We're trying to push the field to have better experimental procedures, better evaluation methods," he says.

Code break

In a survey of 400 artificial intelligence papers presented at major conferences, just 6% included code for the papers' algorithms. Some 30% included test data, whereas 54% included pseudocode, a limited summary of an algorithm.



Henderson's experiment was conducted in a test bed for reinforcement learning algorithms called Gym, created by OpenAI, a nonprofit based in San Francisco, California. John Schulman, a computer scientist at OpenAI who helped create Gym, says that it helps standardize experiments. "Before Gym, a lot of people were working on reinforcement learning, but everyone kind of cooked up their own environments for their experiments, and that made it hard to compare results across papers," he says.

IBM Research presented another tool at the AAAI meeting to aid replication: a system for recreating unpublished source code automatically, saving researchers days or weeks of effort. It's a neural network—a machine learning algorithm made of layers of small computational units, analogous to neurons—that is designed to recreate other neural networks. It scans an AI research paper looking for a chart or diagram describing a neural net, parses those data into layers and connections, and generates the network in new code. The tool has now reproduced hundreds of published neural networks, and IBM is planning to make them available in an open, online repository.

Joaquin Vanschoren, a computer scientist at Eindhoven University of Technology in the Netherlands, has created another repository for would-be replicators: a website called OpenML. It hosts not only algorithms, but also data sets and more than 8 million experimental runs with all their attendant details. "The exact way that you run your experiments is full of undocumented assumptions and decisions," Vanschoren says. "A lot of this detail never makes it into papers."

Psychology has dealt with its reproducibility crisis in part by creating a culture that favors replication, and AI is starting to do the same. In 2015, Rougier helped start *ReScience*, a computer science journal dedicated to replications. The large Neural Information Processing Systems conference has started linking from its website to papers' source code when available. And Ke is helping organize a "reproducibility challenge," in which researchers are invited to try to replicate papers submitted for an upcoming conference. Ke says nearly 100 replications are in progress, mostly by students, who may receive academic credit for their efforts.

Yet AI researchers say the incentives are still not aligned with reproducibility. They don't have time to test algorithms under every condition, or the space in articles to document every hyperparameter they tried. They feel pressure to publish quickly, given that many papers are posted online to arXiv every day without peer review. And many are reluctant to report failed replications. At *ReScience*, for example, all the published replications have so far been positive. Rougier says he's been told of failed attempts, but young researchers often don't want to be seen as criticizing senior researchers. That's one reason why Ke declined to name the researcher behind the speech recognition algorithm she wanted to use as a benchmark.

Gundersen says the culture needs to change. "It's not about shaming," he says. "It's just about being honest." ■

Matthew Hutson is a journalist in New York City.

CITIZEN SCIENCE

U.K. moms are turning parenting into an experiment

Unusual collaboration studies human milk, baby temperature, and schooling

By **Tania Rabesandratana**

On 21 February, about 160 lactating mothers will head to Charing Cross Hospital in London to donate 25 milliliters of milk each for an unusual scientific study. The freshly pumped samples will be analyzed to determine how the composition of human milk changes with the nursling's age, from 3 months to 4 years old.

It's a matter about which surprisingly little is known. But the experiment is equally remarkable for its origin: A group of mothers came up with the idea for the study and designed it together with breast cancer researcher Natalie Shenker and microbial ecologist Simon Cameron, both at Imperial College London. The mothers recruited the milk donors—in just a few days—and they will be involved in the data analysis and possible write-ups.

This unusual collaboration was made possible by the Parenting Science Gang (PSG), a citizen science project in the United Kingdom funded by the Wellcome Trust. It links parents, gathered in Facebook groups around a specific interest, with scientists who help them design and carry out experiments. The project, which has already initiated multiple lines of research into issues such as schooling and gender stereotypes, is an effort to bring evidence to a realm rife with uncertainty and folk wisdom. "I try to raise my children with science in mind," explains Melissa Branzburg, a PSG member and mother of two.

Several blogs and publications have recently sprung up to address the growing hunger for evidence among science-minded parents, and some academics are dispensing advice as well. But PSG allows parents to take matters into their own hands. It was born from a smaller-scale project in which mothers studied which detergents were best to clean cloth diapers, or nap-