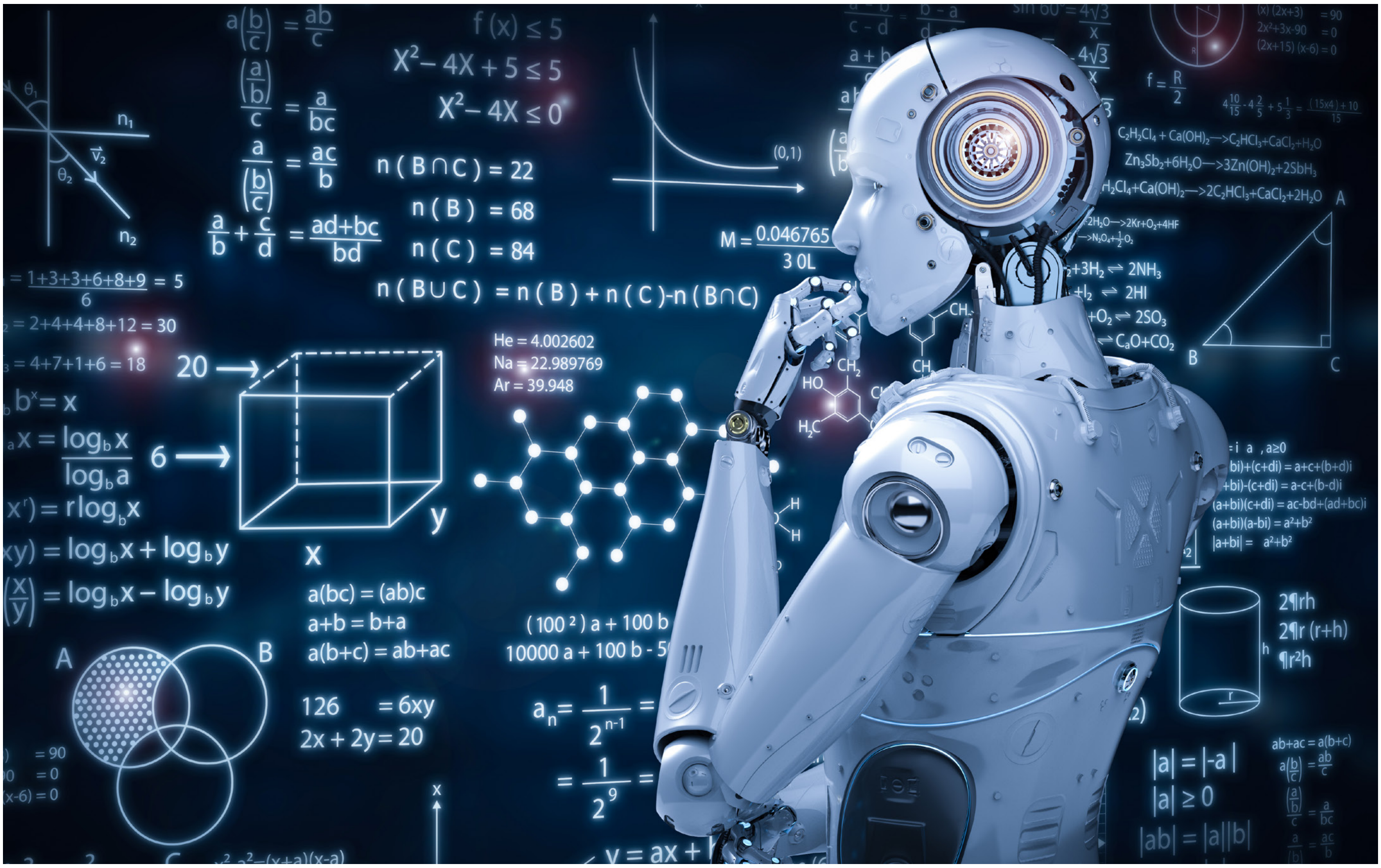

IMAGE: © SCIENCE PHOTO/SHUTTERSTOCK.COM

When robots sleep, do they dream of algorithms?

As artificial intelligence becomes a standard laboratory tool, scientists are quickly discovering both the promise and perils of algorithmically driven research. **By Alan Dove**

Artificial intelligence (AI) is cropping up everywhere these days, according to major news sources that are themselves increasingly driven by computer algorithms. Marketers use AI to target advertisements, engineers use it to anticipate device failures, and AI-driven social media platforms wield outsize influence on everything from fashion to politics.

While all types of AI—also called machine learning—entail programming a computer to learn from examples and make inferences, practitioners distinguish different forms of it. Within the broader field of AI, a subset of strategies employ artificial neural networks. These mimic biological brains, with elements of a program connecting to each other like neurons. Machine learning algorithms running on neural networks are often called deep learning systems, to distinguish them from other approaches such as statistical correlation.

Today, scientists deploy all types of AI to dig through immense quantities of data, from sources ranging from high-throughput DNA and RNA sequencing to massive collections of electronic medical records. A sampling of these efforts reveals a wide range of strategies and applications, and underscores both the potential and challenges of using AI in research.

The new face of genetics

Some of the software developers now applying machine learning to scientific problems got their start working for social media companies. For example, the creators of the algorithms that now power Facebook's automatic photo-tagging features have spent the past few years focusing on a slightly different image-processing problem: identifying rare genetic disorders from facial features.

"About half of genetic disorders are actually characterized by very unique facial patterns," says Dekel Gelbman, chief executive officer (CEO) of **FDNA,** a phenotyping application company in Boston, Massachusetts. While most people can recognize the distinctive features of a person with

cont.>

Upcoming features

Genomics: CRISPR and Beyond—September 27 ■ Bioengineering: Quantum Technology—November 15 ■ Bioprinting—January 10

Down syndrome, human geneticists with specialized training can pinpoint thousands of other, less frequent conditions from facial appearance as well. This type of diagnosis relies on extensive experience, which is difficult to obtain because of the rarity of many genetic disorders. "A handful of very experienced geneticists, who also sometimes refer to themselves as dysmorphologists, are able to very quickly look at a patient and say, 'I've seen something like this before,'" says Gelbman.

Using carefully curated collections of photos, Gelbman and his colleagues trained a machine learning algorithm to group faces according to diagnostic characteristics. The current iteration of the technology uses a deep learning system, and FDNA has built several smartphone applications atop the same framework for different users. Physicians can take one application into the clinic, where they can photograph a patient with the phone's camera and get diagnostic suggestions from the application immediately. A forum application allows them to discuss those diagnoses with experts, while a library application provides relevant literature. Additional applications allow medical educators and researchers to access the same algorithm.

While most individual genetic diseases are rare, their collective impact is large: An estimated 10% of children are born with a rare genetic disorder serious enough to affect their quality of life. "On average, a rare-disease patient waits for seven and a half years before they get a diagnosis, which is ... just unimaginable," says Gelbman. He hopes that automating the dysmorphologist's job will speed diagnosis.

To do that, though, FDNA must overcome two related hurdles: (1) physicians' reluctance to rely on a technology they don't understand and (2) government regulators' stringent standards for medical diagnostics. Both struggle with the impenetrability of current machine learning systems. "It's really hard to trust AI systems, [because] it's really hard for even the programmers to understand the logic of a result," says Gelbman. Developers train and test an algorithm until it yields correct answers, but the reasoning behind those answers often remains inscrutable.

To address this problem, Gelbman advocates more transparency about how the algorithms are being trained and tested. "In the future, organizations are going to be more forthcoming with data sources and the policies for curating and validating data, the validation, and with making benchmarks available for audit," he says. For its part, the U.S. Food and Drug Administration (FDA) has been boning up on AI, and Gelbman says their understanding of the technology has increased substantially in the past year. That said, FDNA has so far kept its applications out of the regulators' purview, by clearly labeling them as providing advice and references, not definitive diagnoses.

If Darwin was a computer scientist

It isn't just medical diagnostic tools that need more transparency. "A lot of these methods in machine learning are black-box approaches, and that's an issue when you're working with biologists who really want to understand how the system's working, not just get the right answer; for them, the question becomes, 'Why is the model picking up on this particular solution?'" says Gary Fogel, CEO of **Natural Selection,** an artificial intelligence consultancy in San Diego, California.

Fogel's company builds AI systems using a type of machine learning that, at least in principle, should appeal to biologists: evolutionary algorithms. In this approach, candidate solutions to a problem are treated like individuals in a population, and a fitness function determines their quality. The system selectively amplifies higher-quality solutions and suppresses or eliminates lower-quality ones, until an optimal solution emerges. Natural Selection has used this approach for everything from analyzing genomic data and screening candidate drug molecules to optimizing industrial processes. However, as mentioned previously, the internal logic of each solution may be as hard to understand as a complex organism.

The company compensates for that by building algorithms that identify salient features in a system. "[We try to find] which features are important to disease or to outcomes and ... try to reduce the features to something that's meaningful, so that biologists…understand the biology of the system," says Fogel.

For some research applications, though, opaque algorithms aren't a problem. That's especially true when investigators are using AI as a tool to identify promising leads, which they then check with lab experiments. "If you're just trying to understand genomics, maybe the need isn't there to have something that's an open box," says Fogel, adding that "if it still accurately predicts where microRNA genes are, you really don't care about why it's doing it correctly, as long as it gets it right."

Nonetheless, even researchers hoping to use AI merely as a laboratory tool need to choose their algorithms carefully. "A lot of people are new to the field and are grabbing whatever open source tools they can," says Fogel, adding that "they don't necessarily know how to tune those algorithms to the problem at hand, and they don't realize that how you represent the problem itself is important." He urges scientists in that position to seek help from computer scientists, many of whom are anxious to apply their algorithm design skills to other fields.

Vector calculus

Such collaborations can arise simply by discussing one's work with colleagues. That's what led Daniel Streicker, senior research fellow at the **University of Glasgow** in Scotland, to apply machine learning to one of the oldest problems in epidemiology: identifying viral vectors and reservoir hosts.

Many of the world's deadliest human viruses are zoonotic, reproducing undetected in animal reservoir hosts most of the time, and only spilling over to humans occasionally. When these infections are carried between hosts by arthropod vectors, epidemiologists may spend decades identifying the relevant nonhuman reservoirs and vectors. In recent years, though, investigators have discovered that RNA viruses, the group most suited to jumping between hosts, optimize diverse features of their genomes, including their amino acid, codon, and dinucleotide use, for the host that they predominately infect. That means there should be clues in a virus's genome sequence that would hint at the identity of its host and vector.

As a biologist, Streicker found that idea tantalizing but didn't know how to pursue it. "My office mate, Simon Babayan, gave an informal seminar within our institute talking about the various projects he was applying machine learning methods to, and it just struck me that this could be the perfect way to address this

Featured participants

Columbia University Irving Medical Center www.cuimc.columbia.edu	**Natural Selection** www.natural-selection.com
FDNA www.fdna.com	**University of Glasgow** www.gla.ac.uk

challenge," says Streicker. The two teamed up with Richard Orton, a bioinformatician at the Medical Research Council-University of Glasgow Centre for Virus Research, and began building algorithms to search for viral hosts and vectors.

The team trained their machine learning system on genome sequences from viruses with well-characterized life cycles, letting it identify correlations between different sequence features and particular host and vector species. "You're really just trying to find some combination of weighting of these features that allows you to effectively map the features of the genome to the host that it comes from," says Streicker.

After the training phase, they tested it on another set of viruses with known hosts to validate its reliability. Finally, they gave the system a set of genomes for viruses with poorly understood etiology and let it predict their transmission patterns.

Many of the results confirmed existing theories, but the system also revealed some surprises. For example, virologists have thought that Crimean–Congo hemorrhagic fever virus spreads mainly through a tick vector, but the computer predicted that direct transmission between livestock animals could also be a major route of infection. The algorithm also predicted that in addition to bats, nonhuman primates could be important reservoir hosts for Ebola viruses (*1*).

To prioritize their research, Streicker's group now hopes to apply the same approach to the deluge of new viral genome sequences coming from metagenomics projects. "We are thinking about how we can use similar approaches to try to predict whether or not humans will be infected by a virus," says Streicker, adding that "this is obviously a question that's quite relevant to surveillance and public health, because there's so much virus discovery going on now." While their initial work focused exclusively on single-stranded RNA viruses, they also hope to expand the project to encompass other types of viral genomes.

Do all the studies, HAL

While genome sequences have become one major focus for algorithm-driven research, other massive data sets are also ripe for machine learning. Over the past few years, for example, researchers at **Columbia University Irving Medical Center** (CUIMC) in New York City have been using various computational approaches to analyze immense troves of electronic medical records, and also to study the biomedical literature itself.

The latter effort has shone a spotlight on what many have called the reproducibility crisis, in which different studies with seemingly valid designs reach opposite conclusions. Observational studies, where researchers take existing medical records and classify patients into control and experimental groups retroactively, are especially problematic. In recent years such studies have yielded results showing that, for example, antidepressants either increase or decrease the risk of suicide, depending on which study one believes. "No two groups pick the same variables to correct for, then they insist that you have to pick exactly the right variables," says George Hripcsak, chair of biomedical informatics at CUIMC.

A related problem is that journals favor papers showing positive results, often based on an arbitrary statistical standard. Hripcsak's own analysis of the literature shows that bias dramatically, with a sharp cutoff in published *p* (probability) values—a measure of statistical significance—at 0.05. Researchers therefore face heavy pressure to pick variables and statistical techniques that will yield publishable *p* values, which can bias their analyses.

To tackle this issue, Hripcsak and his colleagues have turned the job of study design over to a computer. In one recent project, they tapped into multiple databases encompassing hundreds of millions of individual patient medical records, and used an algorithm to design and perform all reasonable observational studies on the data simultaneously. Focusing on depression, the algorithm identified 6,000 potential research hypotheses and over 55,000 control hypotheses, covering 17 treatments, 272 pairs of combined treatments, and 22 outcomes. The algorithm ran for about a month on a powerful computer and generated 5,984 estimates of effects from different treatments. Each of the results meets current methodological standards for publication as a paper in a top peer-reviewed journal. However, the team saw a reassuring distribution of both positive and negative results, indicating they'd avoided the usual publication bias (*2*).

Eliminating human biases doesn't automatically solve the problem, though. "When we do research in this new area we don't want to ... be guilty of the same thing we're trying to prevent, so that's where we're looking at the special things that AI possibly brings in that cause bias," says Hripcsak. Like others in the field, he worries that the opacity of many machine learning algorithms can conceal troubling errors. For example, "economic factors or other things [could] make it so some racial group doesn't do well on a treatment, and then the system recommends not giving them that treatment, when in fact it had nothing to do with their race," says Hripcsak.

Despite the hurdles, though, he and others in the field are optimistic about AI's future in research. "I see a revolution happening, which is great," says Fogel.

References

1. S. A. Babayan, R. J. Orton, D. G. Streicker, *Science* **362,** 577–580 (2018), https://doi.org/10.1126/science.aap9072.
2. M. J. Schuemie, P. B. Ryan, G. Hripcsak, D. Madigan, M. A. Suchard, *Philos. Trans. Royal Soc. A* **376,** 20170356 (2018), https://doi.org/10.1098/rsta.2017.0356.

Alan Dove is a science writer and editor based in Massachusetts.