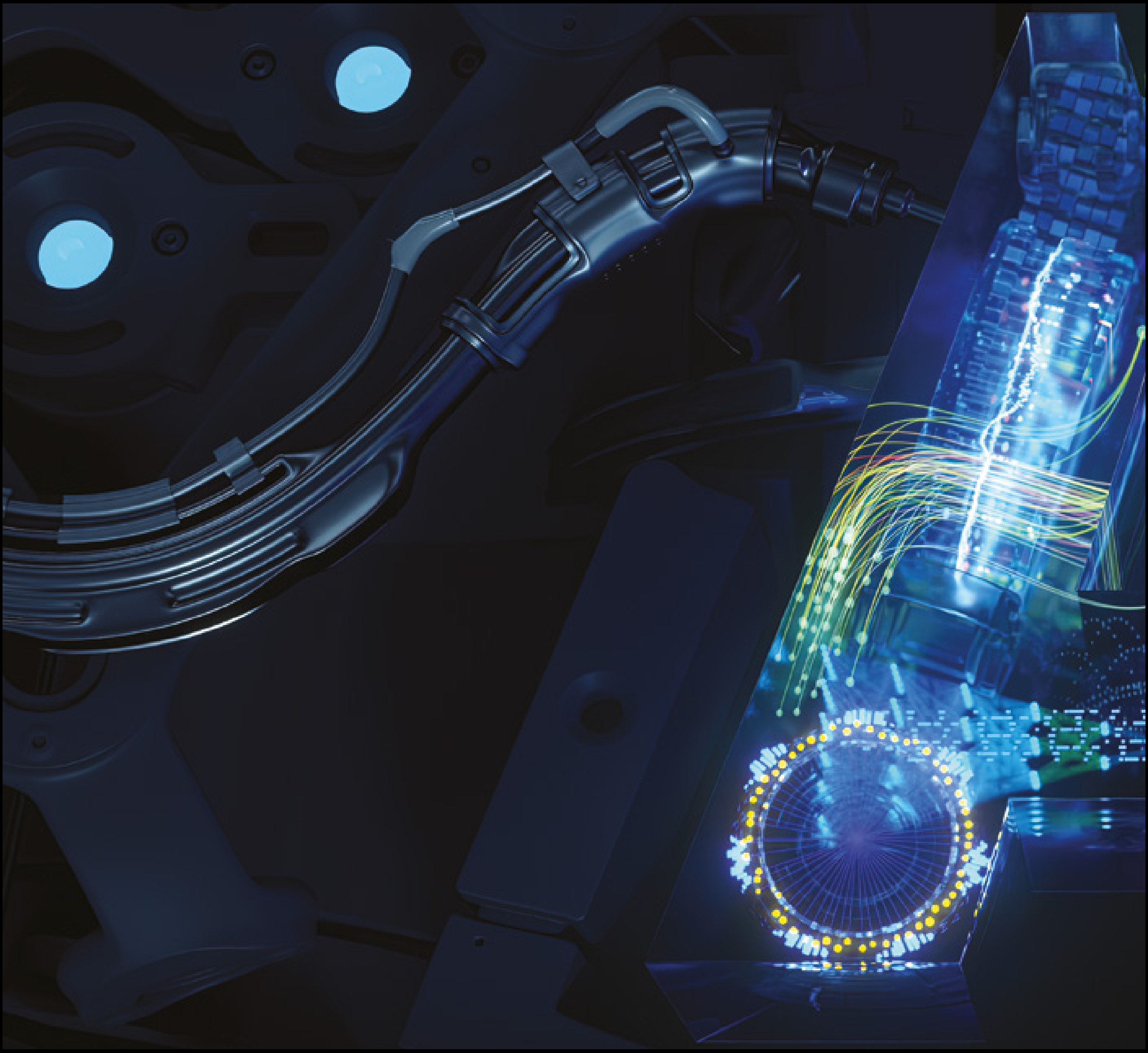
nature
outlook



PETER CROWTHER

Powering up

A revolution in robotics and artificial intelligence

This is a reprint of news features originally published in *Nature*. Read more at https://go.nature.com/robotics-ai

Produced with support from:

FII INSTITUTE
Future Investment Initiative Institute
Impact on Humanity

nature outlook

At the end of the twentieth century, computing was transformed from the preserve of laboratories and industry to a ubiquitous part of everyday life. We are now living through the early stages of a similarly rapid revolution in robotics and artificial intelligence — and the effect on society could be just as enormous.

This collection will be updated continuously online at https://go.nature.com/robotics-ai with stories from journalists and research from across the Nature Portfolio journals. To keep track of the latest and greatest research in robotics and artificial intelligence from across the Nature Portfolio journals, as well as reports from journalists on topics of special interest, sign up to our free newsletter at https://go.nature.com/robotics-signup

We are pleased to acknowledge the financial support of the FII Institute in producing this Outlook. As always, *Nature* retains sole responsibility for all editorial content.

About Nature Outlooks
Nature Outlooks are supplements to *Nature* supported by external funding. They aim to stimulate interest and debate around a subject of particularly strong current interest to the scientific community, in a form that is also accessible to policymakers and the broader public.
Nature has sole responsibility for all editorial content — sponsoring organizations are consulted on the topic of the supplement, but have no influence on reporting thereafter (see go.nature.com/33m79fz). All Nature Outlook supplements are available free online at go.nature.com/outlook

Contact us feedback@nature.com
For information about supporting a future Nature Outlook supplement, visit go.nature.com/partner

Contents

BRYCE VICKMARK/MIT NEWS

Bioinspired robots walk, swim, slither and fly

Engineers look to nature for ideas on how to make robots move through the world.

By Neil Savage

▲ The Mini Cheetah, developed at the Massachusetts Institute of Technology, can run at speeds of up to 3.9 metres per second.

Inspiration can come from anywhere. For Radhika Nagpal, it came from her honeymoon.

Nagpal was snorkelling in the Bahamas when she was approached by a school of colourful striped fish, moving as one. “They come straight at you and check you out and then move off,” says Nagpal, now a mechanical engineer at Princeton University in New Jersey. “I was like, ‘Wow, that is a collective behaviour that I’ve never seen.’”

Her mind returned to those curious fish years later, when she was pondering ways to build swarms of robots that could coordinate their behaviour in challenging environments. The result is a school of robotic fish – called Bluebots – that can coordinate their activity with their fellows[1].

Nagpal’s school is small, only ten fish with limited abilities. The fish are equipped with blue LEDs so that their comrades can spot them underwater. Simple rules in their programming, such as swimming to the left when they see another Bluebot, enable them to synchronize their movement. But Nagpal hopes to eventually build larger collectives with more complex behaviours.

Such robotic schools could be tasked with locating and recording data on coral reefs to help researchers to study the reefs’ health over time. Just as living fish in a school might engage in different behaviours simultaneously – some mating, some caring for young, others finding food – but suddenly move as one when a predator approaches, robotic fish would have to perform individual tasks while communicating to each other when it’s time to do something different.

“The majority of what my lab really looks at is the coordination techniques – what kinds of algorithms have evolved in nature to make systems work well together?” she says.

Many roboticists are looking to biology for inspiration in robot design, particularly in the area of locomotion. Although big industrial robots in vehicle factories, for instance, remain anchored in place, other robots will be more useful if they can move through the world, performing different tasks and coordinating their behaviour.

Some robots can already move on wheels, but wheeled robots cannot climb stairs and are stymied by rough or shifting terrain, such as sand or gravel. By borrowing movement strategies from nature – walking, crawling, swimming, slithering, flying or leaping – robots could gain new functionality. They might perform search-and-rescue operations after an earthquake, or explore caves that are too small or unstable for people to venture into. They could carry out underwater inspections of ships and bridges. And unmanned aerial vehicles (UAVs) could fly more efficiently and better handle turbulence.

“The basic idea is looking to nature to see how things can potentially be done differently, how we can improve our automated systems,” says Michael Tolley, a mechanical engineer who heads the Bioinspired Robotics and Design Lab at the University of California, San Diego.

See Spot run

Perhaps the most obvious strategy for robotic motion is walking, and legged robots do exist. Spot, a low-slung, four-legged machine that looks like a headless yellow dog, can climb uphill and navigate stairs. Its developer, Boston Dynamics in Waltham, Massachusetts, markets the US$74,500 device for mobile inspection of factories, construction sites and hazardous environments. A similar-looking robot, the Mini Cheetah, has been developed at the Massachusetts Institute of Technology (MIT) in Cambridge. “More than 90% of land animals are quadruped,” says Sangbae Kim, a mechanical engineer at MIT who helped to design the Mini Cheetah. “So a natural place to look at is the quadrupedal world. And the cheetah is a king of that world in terms of the speed.”

The Mini Cheetah can already perform backflips, and it runs as fast as 3.9 metres per second – about one-tenth as fast as an actual cheetah, but speedy for a robot. Now Kim is developing control software that he hopes will allow the robot to move smoothly across varying surfaces. This is challenging because the rules for how best to move a limb vary depending on the friction and hardness of the surface. Currently, moving from grass to concrete, or running up a gravelly hill, can cause the robot to stumble. “It runs really ugly and awkward,” Kim says. “It doesn’t fall, but it’s not efficient.”

> “The basic idea is looking to nature to see how things can potentially be done differently, how we can improve our automated systems.”

Nevertheless, quadruped robots are one of the better options for negotiating difficult terrain, says J. Sean Humbert, a mechanical engineer who directs the Bio-Inspired Perception and Robotics Laboratory at the University of Colorado, Boulder. Last year, his group took part in the US Defense Advanced Research Projects Agency’s Subterranean Challenge, in which robots were tasked with navigating tunnels, caves and urban settings to find particular targets; the team took third place, winning $500,000. “The robots that ended up doing really well across the teams were the legged robots,” Humbert says. But faced with a sandy, uphill, rocky landscape, these robots struggled. “Even our Spot robot tipped over and slid around,” he says.

Feel the strain

One possible solution, Humbert says, is to endow robots with animals’ innate ability to sense and respond to mechanosensory information, such as pressure, strain or vibration.

He's been taking that approach with flying machines by embedding strain sensors in the wings of fixed-wing UAVs, as well as in the arms of quadrotor drones, which rely on spinning blades to fly and hover.

The work grew out of studies of honey bees. When Humbert placed bees in a wind tunnel and hit them with sudden gusts of air, their flight would be momentarily disturbed. After a quick change in the pattern of their wing beats, they would right themselves. Honey bees beat their wings 251 times per second, and the animals could make these corrections in just 15 to 20 beats – about 0.08 seconds. "Our conclusion was that [that] had to be mechanosensory information," Humbert says. "Vision is just not fast enough to correct the spins that we're seeing." If a drone could similarly sense a disturbance and automatically correct for it that rapidly, he says, it would be much less likely to crash or be knocked off course.

Fish also respond to mechanosensory stimuli, using a system of sensory organs known as the lateral line. The structure consists of hundreds of tiny sensors spread along the head, trunk and tail fin, and it enables fish to sense changes in the motion and pressure of water caused by obstacles, such as rocks and other animals. "Fish are sensing all of that and are using that, as well as vision, to position themselves relative to each other," Nagpal says. No comparable underwater pressure sensor exists, but her team hopes to develop one to improve the Bluebots' navigation.

In San Diego, Tolley is exploring robots built from polymers or other pliable materials that can more safely interact with humans or squeeze through tight spaces. Squishy, pliable robots could have more flexible motion than hard robots with only a few joints, but getting them to walk on soft legs is a challenge.

Tolley designed a robot with four soft legs, each divided into three chambers[2]. Pressurized air first enters one chamber, then moves to the next. This movement causes the legs to bend, then relax. By alternatively activating opposing pairs of legs, the robot trundles along like a turtle. And because it does not need electronic controls, its design could be useful even in the presence of electromagnetic interference.

Hard or soft, one issue robots struggle with is falling over. If a multimillion-dollar robot trips over a rock on Mars, an entire mission could be jeopardized. Some researchers are looking to insects for solutions, particularly click beetles, which can jump up to 20 times their body length without using their legs[3].

Click beetles use a muscle to compress soft tissue, building up energy; a latch system holds the compressed tissue in place. When the animal releases the latch, producing its characteristic clicking sound, the tissue expands rapidly and the beetle is launched into the air, accelerating at about 530 times the force of gravity. (By comparison, a rider on a roller coaster typically experiences about four times the force of gravity.) If a robot could do that, it would have a mechanism for righting itself after tipping over, says Aimy Wissa, a mechanical and aerospace engineer who runs the Bio-inspired Adaptive Morphology Lab at Princeton.

Even more interesting, Wissa says, is that the beetle can perform this manoeuvre four or five times in rapid succession, without suffering any apparent damage. She's trying to develop models that explain how the energy is rapidly dissipated without harming the insect, which could prove useful in applications involving rapid acceleration and deceleration, such as bulletproof vests. Other creatures also store and release energy to trigger rapid motion, including fruit-fly larvae and Venus flytraps (Dionaea muscipula), and understanding how they do so could lead to more-responsive artificial muscles, Tolley says.

Totally legless

In some places, such as narrow underground passages or on unstable surfaces, legs could require too much space or be too unstable to propel a robot. Howie Choset, a computer scientist at the Robotics Institute of Carnegie Mellon University in Pittsburgh, Pennsylvania, builds snake-like robots with 16 joints that provide a range of motion that could drive everything from surgical instruments wending through the body to reconnaissance robots exploring archaeological sites.

In one early project, Choset took his robo-snakes to the Red Sea, where ancient Egyptians had dug caves to store boats that they'd built for trade with the Land of Punt, thought to be located in modern Somalia. The caves were no longer safe for human explorers, but snake robots seemed well suited to the task – until they didn't. "The truth is, we got stuck," Choset says. "We couldn't go up and down the sandy inclines."

To work out how a real snake would approach the problem, Choset looked to sidewinders, snakes that move by thrusting their bodies sideways in an S-shaped curve, gliding easily over sand4. Because sand is granular, it can behave as either a liquid or a solid, depending on how much force is applied. Choset found that sidewinders can exert the right amount of pushing force so that the sand remains solid underneath them and supports their bodies. "It wasn't until we started looking at the real snakes, the sidewinders, and how they moved on sandy terrains that we were able to understand how to make our robot work on sandy terrains," he says.

As for Wissa, she's trying to build robots that can both swim and fly, using an animal that can do both as inspiration: flying fish5. These creatures use their pelvic fins to skim across the water's surface and then launch into the air, where they can glide up to 400 metres.

Flying fish, Wissa explains, are "actually very good gliders". But when they drop back to the water, they don't submerge. "They actually just dip their caudal fin and they flap it vigorously, and then they can take off again," Wissa says. "You can think of it as a taxiing manoeuvre." She hopes to learn enough about this behaviour to develop a robot that can move through both air and water using the same propulsion mechanisms. "We're very good as engineers in designing things for a single function," Wissa says. "Where nature really can teach us a lot of lessons is this concept of multi-functionality."

For another type of multi-functional locomotion, Wissa focuses on grasshoppers, which can jump and then open their wings to glide. She hopes to understand what makes them such good gliders. Many other insects rely on high-frequency flapping to fly. Perhaps, she says, it has to do with their wing shape.

Wissa also seeks inspiration from birds. She's used aerodynamic testing and structural modelling to investigate covert feathers – small, stiff feathers that overlap other feathers on a bird's wings and tail6. When a bird tries to land in windy conditions, the covert feathers on the wings deploy, either passively in response to air flow or actively under control of a tendon. The covert feathers alter the shape of the wing and give the bird finer control over its interaction with air flow, and don't require as much energy as flapping the whole wing. By learning to understand the physics of these feathers, Wissa hopes to improve the flight of a UAV.

A two-way street

Biology has informed robotics, but the engineering involved can also provide insights into animal kinesiology. "We didn't start by looking at biology," Choset says. Instead, he mathematically modelled the fundamental principles of the motion he was interested in. "And in doing so, something kind of magical happened – we started coming up with ways to explain how biology works. So, is it robot-inspired biology or biologically inspired robots?"

Other engineers have had similar experiences. Nagpal is collaborating with ichthyologist George Lauder at Harvard University in Cambridge to model the hydrodynamics of schooling, to see whether the formation provides living fish with an energy benefit. And designs that make drones fly in a more energy-efficient way might help to explain how birds and insects have evolved to do something similar. Wissa hopes her work, in addition to building flying, swimming robots, will lead to a greater understanding of flying fish. "We're using this model to actually test hypotheses about nature, about why some species of flying fish have enlarged pelvic fins while others don't," Wissa says.

But despite the links between biology and engineering, don't expect bio-inspired robots to ultimately look like the creatures that influenced them. Wissa says that, although many first attempts at mimicking biology resemble the original biological forms, scientists' ultimate aim is to understand the principles behind how the systems operate, and then adapt those to different structures and materials. "We're just copying the physics and the rules for how things work," she says, "and then making engineering systems that serve the same function." ■

CARNEGIE MELLON UNIV.

"It wasn't until we started looking at the real snakes, the sidewinders, and how they moved on sandy terrains that we were able to understand how to make our robot work on sandy terrains,"

▲ This robot, inspired by sidewinding snakes, moves by twisting in an S-shaped curve.

Bioinspired robots walk, swim, slither and fly

Read this article online

REFERENCES

[1] Berlinger, F., Gauci, M. & Nagpal, R. *Sci. Robot.* **6**, eabd8668 (2021).

[2] Drotman, D., Jadhav, S., Sharp, D., Chan, C. & Tolley, M. T. *Sci. Robot.* **6**, eaay2627 (2021).

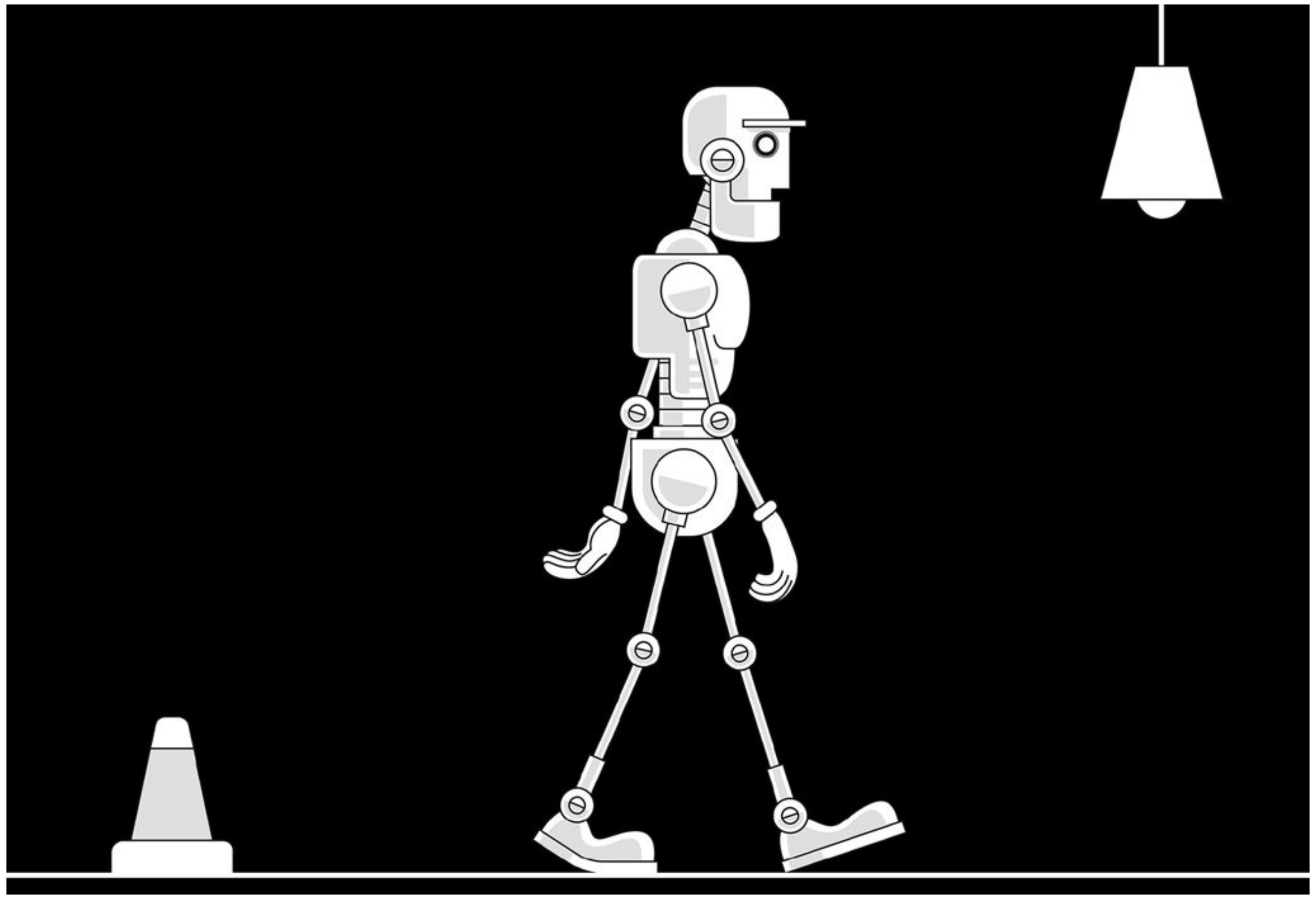
[3] Bolmin, O. *et al. Proc. Natl Acad. Sci. USA* **118**, e2014569118 (2021).

[4] Chaohui Gong, R., Hatton, L. & Choset, H. In *2012 IEEE International Conference on Robotics and Automation* 4222–4227 (2012).

[5] Saro-Cortes, V. *et al. Integr. Comp. Biol.* https://doi.org/10.1093/icb/icac101 (2022).

[6] Duan, C. & Wissa, A. *Bioinspir. Biomim.* **16**, 046020 (2021).

DAVID PARKINS



Learning over a lifetime

Artificial-intelligence researchers turn to lifelong learning in the hopes of making machine intelligence more adaptable. By Neil Savage

▲ The Mini Cheetah, developed at the Massachusetts Institute of Technology, can run at speeds of up to 3.9 metres per second.

Bing Liu was road testing a self-driving car, when suddenly something went wrong. The vehicle had been operating smoothly until it reached a T-junction and refused to move. Liu and the car's other occupants were baffled. The road they were on was deserted, with no pedestrians or other cars in sight. "We looked around, we noticed nothing in the front, or in the back. I mean, there was nothing," says Liu, a computer scientist at the University of Illinois Chicago.

Bing Liu was road testing a self-driving car, when suddenly something went wrong. The vehicle had been operating smoothly until it reached a T-junction and refused to move. Liu and the car's other occupants were baffled. The road they were on was deserted, with no pedestrians or other cars in sight. "We looked around, we noticed nothing in the front, or in the back. I mean, there was nothing," says Liu, a computer scientist at the University of Illinois Chicago.

Stumped, the engineers took over control of the vehicle and drove back to the laboratory to review the trip. They worked out that the car had been stopped by a pebble in the road. It wasn't something a person would even notice, but when it showed up on the car's sensors it registered as an unknown object — something the artificial intelligence (AI) system driving the car had not encountered before.

The problem wasn't with the AI algorithm as such — it performed as intended, stopping short of the unknown object to be on the safe side. The issue was that once the AI had finished its training, using simulations to develop a model that told it the differences between a clear road and an obstacle, it could learn nothing more. When it encountered something that had not been part of its training data, such as the pebble or even a dark spot on the road, the AI did not know how to react. People can build on what they've learnt and adapt as their environment changes; most AI systems are locked into what they already know.

In the real world, of course, unexpected situations inevitably arise. Therefore, Liu argues that any system aiming to perform learnt tasks outside a lab needs to be capable of on-the-job learning — supplementing the model it's already developed with new data that it encounters. The car could, for instance, detect another car driving through a dark patch on the road with no problem, and decide to imitate it, learning in the process that a wet bit of road was not a problem. In the case of the pebble, it could use a voice interface to ask the car's occupant what to do. If the rider said it was safe to continue, it could drive on, and it could then call on that answer for its next pebble encounter. "If the system can continually learn, this problem is easily solved," Liu says.

Such continual learning, also known as lifelong learning, is the next step in the evolution of AI. Much AI relies on neural networks, which take data and pass them through a series of computational units, known as artificial neurons, which perform small mathematical functions on the data. Eventually the network develops a statistical model of the data that it can then match to new inputs. Researchers, who have based these neural networks on the operation of the human brain, are looking to humans again for inspiration on how to make AI systems that can keep learning as they encounter new information. Some groups are trying to make computer neurons more complex so they're more like neurons in living organisms. Others are imitating the growth of new neurons in humans so machines can react to fresh experiences. And some are simulating dream states to overcome a problem of forgetfulness. Lifelong learning is necessary not only for self-driving cars, but for any intelligent system that has to deal with surprises, such as chatbots, which are expected to answer questions about a product or service, and robots that can roam freely and interact with humans. "Pretty much any instance where you deploy AI in the future, you would see the need for lifelong learning," says Dhireesha Kudithipudi, a computer scientist who directs the MATRIX AI Consortium for Human Well-Being at the University of Texas at San Antonio.

> "AI, to date, is really not intelligent. If it's a neural network, you train it in advance, you give it a data set and that's all. It does not have the ability to improve with time."

Continual learning will be necessary if AI is to truly live up to its name. "AI, to date, is really not intelligent," says Hava Siegelmann, a computer scientist at the University of Massachusetts Amherst who created the Lifelong Learning Machines research-funding initiative for the US Defense Advanced Research Projects Agency. "If it's a neural network, you train it in advance, you give it a data set and that's all. It does not have the ability to improve with time."

Model making

In the past decade, computers have become adept at tasks such as classifying cats or tumours in images, identifying sentiment in written language, and winning at chess. Researchers might, for instance, feed the computer photos that have been labelled by humans as containing cats. The computer receives the photos, which it interprets as numerical descriptions of pixels with various colour and brightness values, and runs them through layers of artificial neurons. Each neuron has a randomly chosen weight, a value by which it multiplies the value of the input data. The computer runs the input data through the layers of neurons and checks the output data against validation data to see how accurate the results are. It then repeats the process, altering the weights in each iteration until the output reaches a high accuracy. The process produces a statistical model of the values and the placement of pixels that define a cat. The network can then analyse a new photo and decide whether it matches the model — that is, whether there's a cat in the picture. But that cat model, once developed, is pretty much set in stone.

One way to get the computer to learn to identify many objects would be to develop lots of models. You could train one neural network to recognize cats and another to recognize dogs. That would require two data sets, one for each animal, and would double the time and computing power needed to develop each model. But suppose you wanted the computer to distinguish between pictures of cats and dogs. You would have to train a third network, either using all the original data or comparing the two existing models. Add other animals into the mix and yet more models must be developed.

Training and storing more models requires greater resources, and this can quickly become a problem. Training a neural network can take reams of data and weeks of time. For instance, an AI system called GPT-3, which learnt to produce text that sounds as if it was written by a human, required almost 15 days of training on 10,000 high-end computer processors[1]. The ImageNet data set, which is often used to train neural networks in object recognition, contains more than 14 million images. Depending on the subset of the total number of images that is used, it can take from a few minutes to more than a day and a half to download. Any machine that has to spend days re-learning a task each time it encounters new information will essentially grind to a halt.

doi: https://doi.org/10.1038/d41586-022-01962-y |

| doi: https://doi.org/10.1038/d41586-022-01962-y

One system that could make the generation of multiple models more efficient is Self-Net[2], created by Rolando Estrada, a computer scientist at Georgia State University in Atlanta, and his students Jaya Mandivarapu and Blake Camp. Self-Net compresses the models, to prevent a system with a lot of different animal models from growing too unwieldy.

The system uses an autoencoder, a separate neural network that learns which parameters – such as clusters of pixels in the case of image-recognition tasks – the original neural network focused on when building its model. One layer of neurons in the middle of the autoencoder forces the machine to pick a tiny subset of the most important weights of the model. There might be 10,000 numerical values going into the model and another 10,000 coming out, but in the middle layer the autoencoder reduces that to just 10 numbers. So the system has to find the ten weights that will allow it to get the most accurate output, Estrada says.

The process is similar to compressing a large TIFF image file down to a smaller JPEG, he says; there's a small loss of fidelity, but what is left is good enough. The system tosses out most of the original input data, and then saves the ten best weights. It can then use those to perform the same cat-identification task with almost the same accuracy, without having to store enormous amounts of data.

To streamline the creation of models, computer scientists often use pre-training. Models that are trained to perform similar tasks have to learn similar parameters, at least in the early stages. Any neural network learning to recognize objects in images, for instance, first needs to learn to identify diagonal and vertical lines. There's no need to start from scratch each time, so newer models can be pre-trained with the weights that already recognize those basic features. To make models that can recognize cows or pigs or kangaroos, Estrada can pre-train other neural networks with the parameters from his autoencoder. Because all animals share some of the same facial features, even if the details of size or shape are different, such pre-training allows new models to be generated more efficiently.

The system is not a perfect way to get networks to learn on the job, Estrada says. A human still has to tell the machine when to switch tasks; for example, when to start looking for horses instead of cows. That requires a human to stay in the loop, and it might not always be obvious to a person that it's time for the machine to do something different. But Estrada hopes to find a way to automate task switching so the computer can learn to identify characteristics of the input data and use that to decide which model it should use, so it can keep operating without interruption.

TEJ PANDIT

▲ Computer scientist Dhireesha Kudithipudi (right) and her student Nicholas Soures discuss factors that affect continual learning.

Out with the old

It might seem that the obvious course is not to make multiple models but rather to grow a network. Instead of developing two networks for recognizing cats and horses respectively, for instance, it might appear easier to teach the cat-savvy network to also recognize horses. This approach, however, forces AI designers to confront one of the main issues in lifelong learning, a phenomenon known as catastrophic forgetting. A network trained to recognize cats will develop a set of weights across its artificial neurons that are specific to that task. If it is then asked to start identifying horses, it will start readjusting the weights to make it more accurate for horses. The model will no longer contain the right weights for cats, causing it to essentially forget what a cat looks like. "The memory is in the weights. When you train it with new information, you write on the same weights," says Siegelmann. "You can have a billion examples of a car driving, and now you teach it 200 examples related to some accident that you don't want to happen, and it may know these 200 cases and forget the billion."

One method of overcoming catastrophic forgetting uses replay – that is, taking data from a previously learnt task and interweaving them with new training data. This approach, however, runs head-on into the resource problem. "Replay mechanisms are very memory hungry and computationally hungry, so we do not have models that can solve these problems in a resource-efficient way," Kudithipudi says. There might also be reasons not to store data, such as concerns about privacy or security, or because they belong to someone unwilling to share them indefinitely.

Siegelmann says replay is roughly analogous to what the human brain does when it dreams. Many neuroscientists think that the brain consolidates memories and learns things by replaying experiences during sleep. Similarly, replay in neural networks can reinforce weights that might otherwise be overwritten. But the brain doesn't actually review a moment-by-moment rerun of its experiences, Siegelmann says. Rather, it reduces those experiences to a handful of characteristic features and patterns – a process known as abstraction – and replays just those parts. Her brain-inspired replay tries to do something similar; instead of reviewing mountains of stored data, it selects certain facets of what it has learnt to replay. Each layer in a neural network, Siegelmann says, moves the learning to a higher level of abstraction, from the specific input data in the bottom layer to mathematical relationships in the data at higher layers. In this way, the system sorts specific examples of objects into classes. She lets the network select the most important of the abstractions in the top couple of layers and replay those. This technique keeps the learnt weights reasonably stable – although not perfectly so – without having to store any previously used data at all.

Because such brain-inspired replay focuses on the most salient points that the network has learnt, the network can find associations between new and old data more easily. The method also helps the network to distinguish between pieces of data that it might not have separated easily before – finding the differences between a pair of identical twins, for example. If you're down to only a handful of parameters in each set, instead of millions, it's easier to spot the similarities. "Now, when we replay one with the other, we start looking at the differences," Siegelmann says. "It forces you to find the separation, the contrast, the associations."

Focusing on high-level abstractions rather than specifics is useful for continual learning because it allows the computer to make comparisons and draw analogies between different scenarios. For example, if your self-driving car has to work out how to handle driving on ice in Massachusetts, Siegelmann says, it might use data that it has about driving on ice in Michigan. Those examples won't exactly match the new conditions, because they're from different roads. But the car also has knowledge about driving on snow in Massachusetts, where it is familiar with the roads. So if the car can identify only the most important differences and similarities between snow and ice, Massachusetts and Michigan, instead of getting bogged down in minor details, it might come up with a solution to the specific, new situation of driving on ice in Massachusetts.

A modular approach

Looking at how the brain handles these issues can inspire ideas, even if they don't replicate what's going on biologically. To deal with the need for a neural network that can learn tasks without overwriting the old, scientists take a cue from neurogenesis – the process by which neurons are formed in the brain. A machine can't grow parts the way a body can, but computer scientists can replicate new neurons in software by generating connections in parts of the system. Although the mature neurons have learnt to react to only certain data inputs, these 'baby neurons' can respond to all the input. "They can react to new samples that are fed into the model," Kudithipudi says. In other words, they can learn from new information while the already-trained neurons retain what they've learnt.

Adding more neurons is just one way to enable a system to learn new things. Estrada has come up with another approach, on the basis of the fact that a neural network is only a loose approximation of a human brain. "We call the nodes in a neural network 'neurons'. But if you see what they're actually doing, they're basically computing a weighted sum. It's an incredibly simplified view of real, biological neurons, which perform all sorts of complex nonlinear signal processing."

In an effort to mimic some of the complicated behaviours of real neurons more successfully, Estrada and his students developed what he calls deep artificial neurons (DANs)[3]. A DAN is a small neural network that is treated as a single neuron in a larger neural network.

DANs can be trained for one particular task – for instance, Estrada might develop one for identifying handwritten numbers. The model in the DAN is then fixed, so it can't be changed and will always provide the same output to other neurons in the still-trainable network layers surrounding it. That larger network can go on to learn a related task, such as identifying numbers written by someone else – but the original model is not forgotten.

Learning over a lifetime

Read this article online

"You end up with this general-purpose module that you can reuse for similar tasks in the future," Estrada says. "These modules allow the system to learn to perform the new tasks in a similar way to the old tasks, so that the features are more compatible with each other over time. So that means that the features are more stable and it forgets less."

So far, Estrada and his colleagues have shown that this technique works on fairly simple tasks, such as number recognition. But they're trying to adapt it to more challenging problems, including learning how to play old video games such as Space Invaders. "And then, if that's successful, we could use it for more sophisticated things," says Estrada. It might, for instance, prove useful in autonomous drones, which are sent out with basic programming but have to adapt to new data in the environment, and will have to do any on-the-fly learning within tight power and processing constraints.

There's a long way to go before AI can function as people do, dealing with an endless variety of ever-changing scenarios. But if computer scientists can develop the techniques to allow machines to make the continual adaptations that living creatures are capable of, it could go a long way towards making AI systems more versatile, more accurate and more recognizably intelligent ■

REFERENCES

1. Patterson, D. *et al.* Preprint at https://arxiv.org/abs/2104.10350 (2021).
2. Mandivarapu, J. K., Camp, B. & Estrada, R. *Front. Artif. Intell.* **3**, 19 (2020).
3. Camp, B., Mandivarapu, J. K. & Estrada, R. Preprint at https://arxiv.org/abs/2011.07035 (2020).

doi: https://doi.org/10.1038/d41586-022-01962-y |

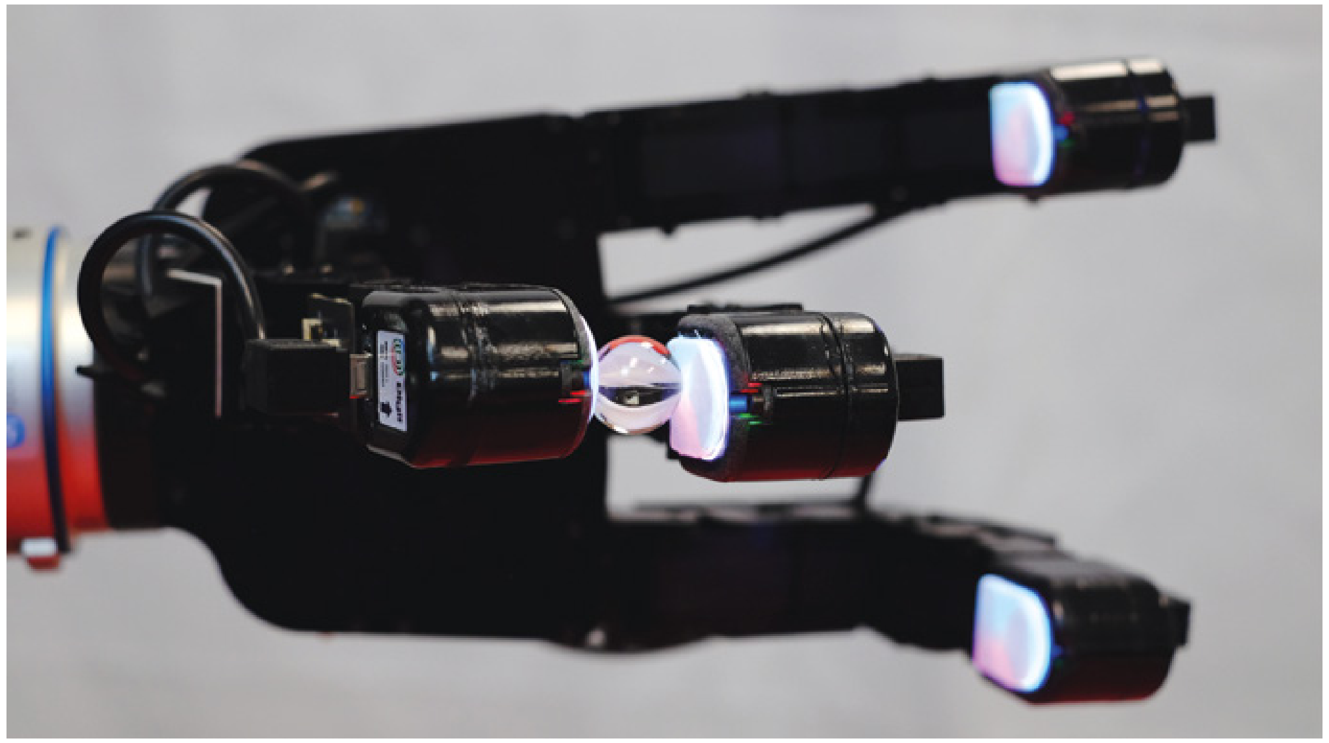
| doi: https://doi.org/10.1038/d41586-022-01962-y

Teaching robots to touch

Robots have become increasingly adept at interacting with the world around them. But to fulfil their potential, they also need a sense of touch.

By Marcus Woo

▼ DIGIT is a camera-based fingertip-like sensor manufactured by companies GelSight and Meta AI.



M. LAMBETA ET AL. IEEE ROBOT. AUTOM. LETT. 5, 3838–3845 (2020)

Fork in hand, a robot arm skewers a strawberry from above and delivers it to Tyler Schrenk's mouth. Sitting in his wheelchair, Schrenk nudges his neck forward to take a bite. Next, the arm goes for a slice of banana, then a carrot. Each motion it performs by itself, on Schrenk's spoken command.

For Schrenk, who became paralysed from the neck down after a diving accident in 2012, such a device would make a huge difference in his daily life if it were in his home. "Getting used to someone else feeding me was one of the strangest things I had to transition to," he says. "It would definitely help with my well-being and my mental health."

His home is already fitted with voice-activated power switches and door openers, enabling him to be independent for about 10 hours a day without a caregiver. "I've been able to figure most of this out," he says. "But feeding on my own is not something I can do." Which is why he wanted to test the feeding robot, dubbed ADA (short for assistive dexterous arm). Cameras located above the fork enable ADA to see what to pick up. But knowing how forcefully to stick a fork into a soft banana or a crunchy carrot, and how tightly to grip the utensil, requires a sense that humans take for granted: "Touch is key," says Tapomayukh Bhattacharjee, a roboticist at Cornell University in Ithaca, New York, who led the design of ADA while at the University of Washington in Seattle. The robot's two fingers are equipped with sensors that measure the sideways (or shear) force when holding the fork[1]. The system is just one example of a growing effort to endow robots with a sense of touch.

"The really important things involve manipulation, involve the robot reaching out and changing something about the world," says Ted Adelson, a computer-vision specialist at the Massachusetts Institute of Technology (MIT) in Cambridge. Only with tactile feedback can a robot adjust its grip to handle objects of different sizes, shapes and textures. With touch, robots can help people with limited mobility, pick up soft objects such as fruit, handle hazardous materials and even assist in surgery. Tactile sensing also has the potential to improve prosthetics, help people to literally stay in touch from afar, and even has a part to play in fulfilling the fantasy of the all-purpose household robot that will take care of the laundry and dishes. "If we want robots in our home to help us out, then we'd want them to be able to use their hands," Adelson says. "And if you're using your hands, you really need a sense of touch."

With this goal in mind, and buoyed by advances in machine learning, researchers around the world are developing myriad tactile sensors, from finger-shaped devices to electronic skins. The idea isn't new, says Veronica Santos, a roboticist at the University of California, Los Angeles. But advances in hardware, computational power and algorithmic knowhow have energized the field. "There is a new sense of excitement about tactile sensing and how to integrate it with robots," Santos says.

Feel by sight

One of the most promising sensors relies on well-established technology: cameras. Today's cameras are inexpensive yet powerful, and combined with sophisticated computer vision algorithms, they've led to a variety of tactile sensors. Different designs use slightly different techniques, but they all interpret touch by visually capturing how a material deforms on contact.

ADA uses a popular camera-based sensor called GelSight, the first prototype of which was designed by Adelson and his team more than a decade ago[2]. A light and a camera sit behind a piece of soft rubbery material, which deforms when something presses against it. The camera then captures the deformation with super-human sensitivity, discerning bumps as small as one micrometre. GelSight can also estimate forces, including shear forces, by tracking the motion of a pattern of dots printed on the rubbery material as it deforms[2].

> **"There is a new sense of excitement about tactile sensing and how to integrate it with robots."**

GelSight is not the first or the only camera-based sensor (ADA was tested with another one, called FingerVision). However, its relatively simple and easy-to-manufacture design has so far set it apart, says Roberto Calandra, a research scientist at Meta AI (formerly Facebook AI) in Menlo Park, California, who has collaborated with Adelson. In 2011, Adelson co-founded a company, also called GelSight, based on the technology he had developed. The firm, which is based in Waltham, Massachusetts, has focused its efforts on industries such as aerospace, using the sensor technology to inspect for cracks and defects on surfaces.

One of the latest camera-based sensors is called Insight, documented this year by Huanbo Sun, Katherine Kuchenbecker and Georg Martius at the Max Planck Institute for Intelligent Systems in Stuttgart, Germany[3]. The finger-like device consists of a soft, opaque, tent-like dome held up with thin struts, hiding a camera inside.

It's not as sensitive as GelSight, but it offers other advantages. GelSight is limited to sensing contact on a small, flat patch, whereas Insight detects touch all around its finger in 3D, Kuchenbecker says. Insight's silicone surface is also easier to fabricate, and it determines forces more precisely. Kuchenbecker says that Insight's bumpy interior surface makes forces easier to see, and unlike GelSight's method of first determining the geometry of the deformed rubber surface and then calculating the forces involved, Insight determines forces directly from how light hits its camera. Kuchenbecker thinks this makes Insight a better option for a robot that needs to grab and manipulate objects; Insight was designed to form the tips of a three-digit robot gripper called TriFinger.

Skin solutions

Camera-based sensors are not perfect. For example, they cannot sense invisible forces, such as the magnitude of tension of a taut rope or wire. A camera's frame-rate might also not be quick enough to capture fleeting sensations, such as a slipping grip, Santos says. And squeezing a relatively bulky camera-based sensor into a robot finger or hand, which might already be crowded with other sensors or actuators (the components that allow the hand to move) can also pose a challenge.

This is one reason other researchers are designing flat and flexible devices that can wrap around a robot appendage. Zhenan Bao, a chemical engineer at Stanford University in California, is designing skins that incorporate flexible electronics and replicate the body's ability to sense touch. In 2018, for example, her group created a skin that detects the direction of shear forces by mimicking the bumpy structure of a below-surface layer of human skin called the spinosum[4].

When a gentle touch presses the outer layer of human skin against the dome-like bumps of the spinosum, receptors in the bumps feel the pressure. A firmer touch activates deeper-lying receptors found below the bumps, distinguishing a hard touch from a soft one. And a sideways force is felt as pressure pushing on the side of the bumps.

Bao's electronic skin similarly features a bumpy structure that senses the intensity and direction of forces. Each one-millimetre bump is covered with 25 capacitors, which store electrical energy and act as individual sensors. When the layers are pressed together, the amount of stored energy changes. Because the sensors are so small, Bao says, a patch of electronic skin can pack in a lot of them, enabling the skin to sense forces accurately and aiding a robot to perform complex manipulations of an object.

To test the skin, the researchers attached a patch to the fingertip of a rubber glove worn by a robot hand. The hand could pat the top of a raspberry and pick up a ping-pong ball without crushing either.

Although other electronic skins might not be as sensor-dense, they

doi: https://doi.org/10.1038/d41586-022-01401-y |

| doi: https://doi.org/10.1038/d41586-022-01401-y

tend to be easier to fabricate. In 2020, Benjamin Tee, a former student of Bao who now leads his own laboratory at the National University of Singapore, developed a sponge-like polymer that can sense shear forces[5]. Moreover, similar to human skin, it is self-healing: after being torn or cut, it fuses back together when heated and stays stretchy, which is useful for dealing with wear and tear.

The material, dubbed AiFoam, is embedded with flexible copper wire electrodes, roughly emulating how nerves are distributed in human skin. When touched, the foam deforms and the electrodes squeeze together, which changes the electrical current travelling through it. This allows both the strength and direction of forces to be measured. AiFoam can even sense a person's presence just before they make contact – when their finger comes within a few centimetres, it lowers the electric field between the foam's electrodes.

Last November, researchers at Meta AI and Carnegie Mellon University in Pittsburgh, Pennsylvania, announced a touch-sensitive skin comprising a rubbery material embedded with magnetic particles[6]. Dubbed ReSkin, when it deforms the particles move along with it, changing the magnetic field. It is designed to be easily replaced – it can be peeled off and a fresh skin installed without requiring complex recalibration – and 100 sensors can be produced for less than US$6.

Rather than being universal tools, different skins and sensors will probably lend themselves to particular purposes. Bhattacharjee and his colleagues, for example, have created a stretchable sleeve that fits over a robot arm and is useful for sensing incidental contact between a robotic arm and its environment[7]. The sheet is made from layered fabric that detects changes in electrical resistance when pressure is applied to it. It can't detect shear forces, but it can cover a broad area and wrap around a robot's joints.

Bhattacharjee is using the sleeve to identify not just when a robotic arm comes into contact with something as it moves through a cluttered environment, but also what it bumps up against. If a helper robot in a home brushed against a curtain while reaching for an object, it might be fine for it to continue, but contact with a fragile wine glass would require evasive action.

Other approaches use air to provide a sense of touch. Some robots use suction grippers to pick up and move objects in warehouses or in the oceans. In these cases, Hannah Stuart, a mechanical engineer at the University of California, Berkeley, is hoping that measuring suction airflow can provide tactile feedback to a robot. Her group has shown that the rate of airflow can determine the strength of the suction gripper's hold and even the roughness of the surface it is suckered on to[8]. And underwater, it can reveal how an object moves while being held by a suction-aided robot hand[9].

Processing feelings

Today's tactile technologies are diverse, Kuchenbecker says. "There are multiple feasible options, and people can build on the work of others," she says. But designing and building sensors is only the start. Researchers then have to integrate them into a robot, which must then work out how to use a sensor's information to execute a task. "That's actually going to be the hardest part," Adelson says.

For electronic skins that contain a multitude of sensors, processing and analysing data from them all would be computationally and energy intensive. To handle so many data, researchers such as Bao are taking inspiration from the human nervous system, which processes a constant flood of signals with ease. Computer scientists have been trying to mimic the nervous system with neuromorphic computers for more than 30 years. But Bao's goal is to combine a neuromorphic approach with a flexible skin that could integrate with the body seamlessly – for example, on a bionic arm.

ALEXIS E. BLOCK

▲ Alexis Block, a postdoc at the University of California, Los Angeles, experiences a hug from a HuggieBot, a robot she helped to create that can feel when someone pats or squeezes it.

Unlike in other tactile sensors, Bao's skins deliver sensory signals as electrical pulses, such as those in biological nerves. Information is stored not in the intensity of the pulses, which can wane as a signal travels, but instead in their frequency. As a result, the signal won't lose much information as the range increases, she explains.

Pulses from multiple sensors would meet at devices called synaptic transistors, which combine the signals into a pattern of pulses – similar to what happens when nerves meet at synaptic junctions. Then, instead of processing signals from every sensor, a machine-learning algorithm needs only to analyse the signals from several synaptic junctions, learning whether those patterns correspond to, say, the fuzz of a sweater or the grip of a ball.

In 2018, Bao's lab built this capability into a simple, flexible, artificial nerve system that could identify Braille characters[10]. When attached to a cockroach's leg, the device could stimulate the insect's nerves – demonstrating the potential for a prosthetic device that could integrate with a living creature's nervous system.

Ultimately, to make sense of sensor data, a robot must rely on machine learning. Conventionally, processing a sensor's raw data was tedious and difficult, Calandra says. To understand the raw data and convert them into physically meaningful numbers such as force, roboticists had to calibrate and characterize the sensor. With machine learning, roboticists can skip these laborious steps. The algorithms enable a computer to sift through a huge amount of raw data and identify meaningful patterns by itself. These patterns – which can represent a sufficiently tight grip or a rough texture – can be learnt from training data or from computer simulations of its intended task, and then applied in real-life scenarios.

> "My hope is by open-sourcing this ecosystem, we're lowering the entry bar for new researchers who want to approach the problem. This is really the beginning."

"We've really just begun to explore artificial intelligence for touch sensing," Calandra says. "We are nowhere near the maturity of other fields like computer vision or natural language processing." Computer-vision data are based on a two-dimensional array of pixels, an approach that computer scientists have exploited to develop better algorithms, he says. But researchers still don't fully know what a comparable structure might be for tactile data. Understanding the structure for those data, and learning how to take advantage of them to create better algorithms, will be one of the biggest challenges of the next decade.

Barrier removal

The boom in machine learning and the variety of emerging hardware bodes well for the future of tactile sensing. But the plethora of technologies is also a challenge, researchers say. Because so many labs have their own prototype hardware, software and even data formats, scientists have a difficult time comparing devices and building on one another's work. And if roboticists want to incorporate touch sensing into their work for the first time, they would have to build their own sensors from scratch – an often expensive task, and not necessarily in their area of expertise.

This is why, last November, GelSight and Meta AI announced a partnership to manufacture a camera-based fingertip-like sensor called DIGIT. With a listed price of $300, the device is designed to be a standard, relatively cheap, off-the-shelf sensor that can be used in any robot. "It definitely helps the robotics community, because the community has been hindered by the high cost of hardware," Santos says.

Depending on the task, however, you don't always need such advanced hardware. In a paper published in 2019, a group at MIT led by Subramanian Sundaram built sensors by sandwiching a few layers of material together, which change electrical resistance when under pressure[11]. These sensors were then incorporated into gloves, at a total material cost of just $10. When aided by machine learning, even a tool as simple as this can help roboticists to better understand the nuances of grip, Sundaram says.

Not every roboticist is a machine-learning specialist, either. To aid with this, Meta AI has released open source software for researchers to use. "My hope is by open-sourcing this ecosystem, we're lowering the entry bar for new researchers who want to approach the problem," Calandra says. "This is really the beginning."

Although grip and dexterity continue to be a focus of robotics, that's not all tactile sensing is useful for. A soft, slithering robot, might need to feel its way around to navigate rubble as part of search and rescue operations, for instance. Or a robot might simply need to feel a pat on the back: Kuchenbecker and her student Alexis Block have built a robot with torque sensors in its arms and a pressure sensor and microphone inside a soft, inflatable body that can give a comfortable and pleasant hug, and then release when you let go. That kind of human-like touch is essential to many robots that will interact with people, including prosthetics, domestic helpers and remote avatars. These are the areas in which tactile sensing might be most important, Santos says. "It's really going to be the human–robot interaction that's going to drive it."

So far, robotic touch is confined mainly to research labs. "There's a need for it, but the market isn't quite there," Santos says. But some of those who have been given a taste of what might be achievable are already impressed. Schrenk's tests of ADA, the feeding robot, provided a tantalizing glimpse of independence. "It was just really cool," he says. "It was a look into the future for what might be possible for me." ■

Teaching robots to touch

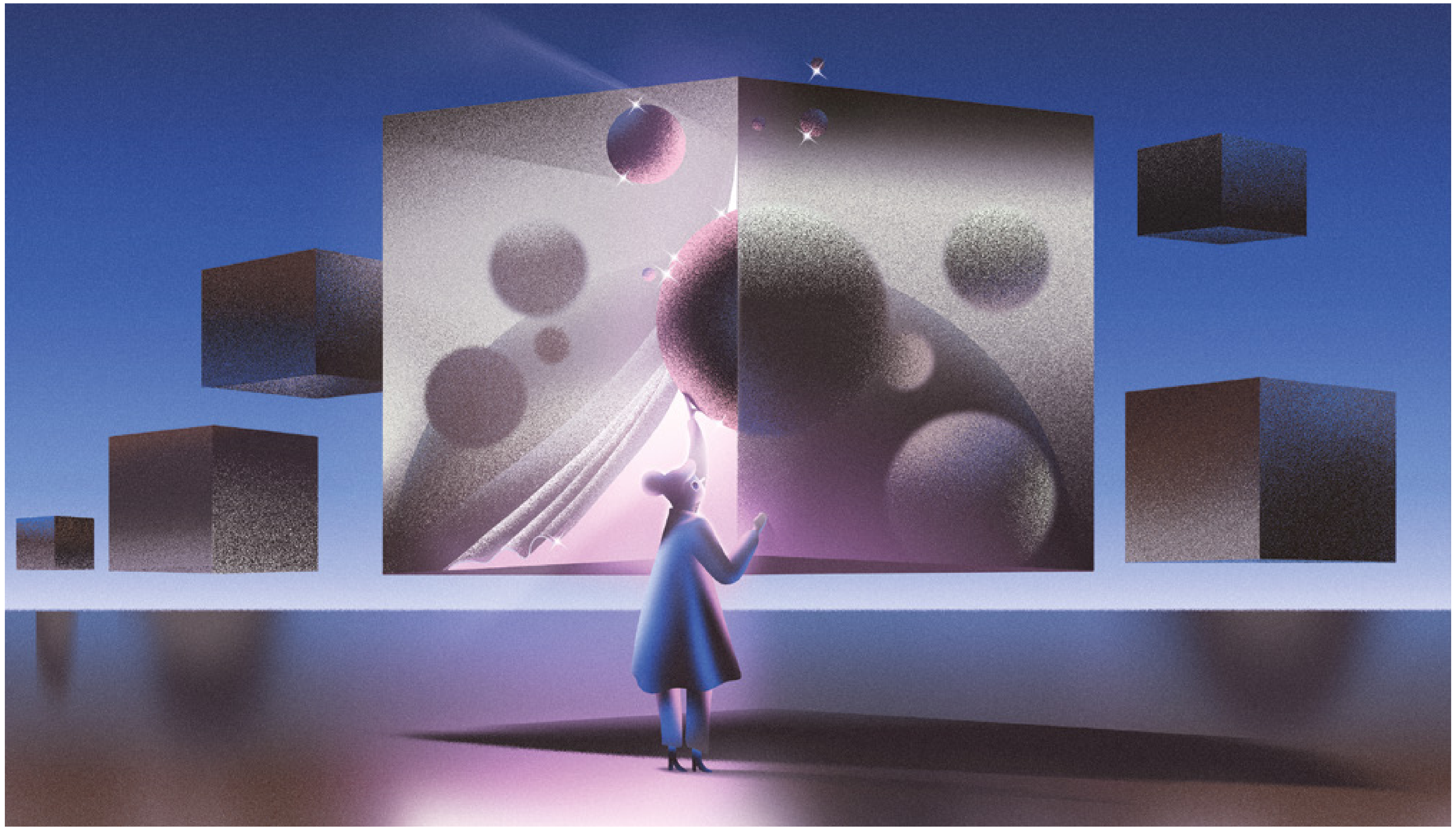
Read this article online

REFERENCES

1. Song, H., Bhattacharjee, T. & Srinivasa, S. S. *2019 International Conference on Robotics and Automation* 8367–8373 (IEEE, 2019).
2. Yuan, W., Dong, S. & Adelson, E. H. *Sensors* **17**, 2762 (2017).
3. Sun, H., Kuchenbecker, K. J. & Martius, G. *Nature Mach. Intell.* **4**, 135–145 (2022).
4. Boutry, C. M. *et al. Sci. Robot.* **3**, aau6914 (2018).
5. Guo, H. *et al. Nature Commun.* **11**, 5747 (2020).
6. Bhirangi, R., Hellebrekers, T., Majidi, C. & Gupta, A. Preprint at http://arxiv.org/abs/2111.00071 (2021).
7. Wade, J., Bhattacharjee, T., Williams, R. D. & Kemp, C. C. *Robot. Auton. Syst.* **96**, 1–14 (2017).
8. Huh, T. M. *et al. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems* 1786–1793 (IEEE, 2021).
9. Nadeau, P., Abbott, M., Melville, D. & Stuart, H. S. *2020 IEEE International Conference on Robotics and Automation* 3701–3707 (IEEE, 2020).
10. Kim, Y. *et al. Science* **360**, 998–1003 (2018).
11. Sundaram, S. *et al. Nature* **569**, 698–702 (2019).

Breaking into the black box of artificial intelligence

Scientists are finding ways to explain the inner workings of complex machine-learning models. By Neil Savage



SANDRO RYBAK

In February 2020, with COVID-19 spreading rapidly around the globe and antigen tests hard to come by, some physicians turned to artificial intelligence (AI) to try to diagnose cases[1]. Some researchers tasked deep neural networks — complex systems that are adept at finding subtle patterns in images — with looking at X-rays and chest computed tomography (CT) scans to quickly distinguish between people with COVID-based pneumonia and those without[2]. "Early in the COVID-19 pandemic, there was a race to build tools, especially AI tools, to help out," says Alex DeGrave, a computer engineer at the University of Washington in Seattle. But in that rush, researchers did not notice that many of the AI models had decided to take a few shortcuts.

The AI systems honed their skills by analysing X-rays that had been labelled as either COVID-positive or COVID-negative. They would then use the differences they had spotted between the images to make inferences about new, unlabelled X-rays. But there was a problem. "There wasn't a lot of data available at the time," says DeGrave.

Radiographs of people with COVID-19 were being released by a number of hospitals, he explains. Scans of people without COVID-19, meanwhile, came mainly from a repository of lung images held by the US National Institutes of Health, put together before the pandemic. As a result, the data sets had characteristic differences that had nothing to do with whether a person had the disease. For instance, many X-rays use the letter R to label a person's right side, so a radiologist looking at the image can orient it properly. However, the appearance of these markers differs from one hospital to another. With most of the COVID-negative images coming from a single source, some of the AI systems trained in this way based their diagnoses not just on the biology on display, but on the style and placement of the letter R on the X-ray.

DeGrave and Joseph Janizek, both members of computer scientist Su-In Lee's Lab of Explainable AI for Biological and Medical Sciences in Seattle, published a paper[3] in Nature Machine Intelligence in May 2021 reporting the problem. The decision-making process of a machine-learning model is often referred to as a black box — researchers and users typically know the inputs and outputs, but it is hard to see what's going on inside. But DeGrave and Janizek were able to prise open these boxes, using techniques designed to test AI systems and explain why they do what they do.

There are many advantages to building explainable AI, sometimes known as XAI. In a medical setting, understanding why a system made a certain diagnosis can help to convince a pathologist that it is legitimate. In some cases, explanations are required by law: when a system makes a decision on loan eligibility, for example, both the United States and the European Union require evidence that if credit is denied it is not for reasons barred by law, such as race or sex. Insight into an AI system's inner workings can also help computer scientists to improve and refine the models they create — and might even lead to fresh ideas about how to approach certain problems. However, the benefits of XAI can only be achieved if the explanations it gives are themselves understandable and verifiable — and if the people building the models see it as a worthwhile endeavour.

A neuron by any other name

The deep neural networks that DeGrave and Janizek investigated have become popular for their uncanny ability to learn about what's in a photograph, the meaning of spoken language and much more, just through exposure. These networks work in a similar way to the human brain. Just as certain living nerve cells fire in a pattern in response to external stimuli — the sight of a cat, for instance, will trigger a different pattern from the sight of a tree — the artificial neurons in a neural network produce a characteristic response on the basis of the input they receive.

The neurons in this case are mathematical functions. Input data comes into the system in numerical form, describing, for instance, the colour of a pixel in a photograph. The neurons then perform a calculation on that data. In the human body, neurons fire off a signal only if the stimulus they receive surpasses a certain electrical threshold. Similarly, each mathematical neuron in an artificial neural network is weighted with a threshold value. If the result of the calculation surpasses that threshold, it is passed to another layer of neurons for further calculations. Eventually, the system learns statistical patterns about how the data coming out relates to the data going in. Images that have been labelled as having a cat in them will have systematic differences from those labelled as not having a cat, and these telltale signs can then be looked for in other images to ascertain the probability of a cat being present.

There are variations in the design of neural networks, as well as other machine-learning techniques. The more layers of calculation a model applies to an input, the more challenging it becomes to explain what it is doing. Simple models such as small decision trees — which weigh up a handful of competing choices that lead to different answers — are not really black boxes, says Kate Saenko, a computer scientist at Boston University in Massachusetts. Small decision trees are "basically a set of rules where a human can easily understand what that model is doing, so it's inherently interpretable", she says. A deep neural network, however, is typically too complex for us to wrap our heads around easily. "A neural network is doing a computation that involves millions, or more likely now billions, of numbers," Saenko says.

Mapping activity

In general, attempts to explain the mysterious workings of a deep neural network involve finding out what characteristics of the input data are affecting the results, and using that to infer what's happening inside the black box. One tool that helped DeGrave and Janizek to work out that the orientation markers on chest X-rays were affecting diagnoses was saliency maps – colour-coded charts that show which part of an image the computer paid the most attention to when making its call.

Saenko and her colleagues developed a technique called D-RISE (detector randomized input sampling for explanation) to produce such maps[4]. The researchers take a photo – for instance, of a vase full of flowers – and systematically block out different parts of the image before showing it to an AI tasked with identifying a particular object, such as the vase. They then record how obscuring each cluster of pixels affects the accuracy of the results, as well as telling the system to colour code the whole photo according to how important each part was to the recognition process.

Unsurprisingly, in a picture of a flower-filled vase, the vase itself is lit up in bright reds and yellows – its presence is important. But it is not the only area of the picture that is highlighted. "The saliency extends all the way up to the bouquet of flowers," Saenko says. "They're not labelled as part of the vase, but the model learns that if you see flowers, it's much more likely that this object is a vase."

D-RISE highlights the factors that, if removed, would cause the AI model to change its results. "It's useful for understanding what mistakes they might be making, or if they're doing something for the wrong reason," says Saenko, whose work in this area was partly funded by a now-completed XAI programme run by the US Defense Advanced Research Projects Agency.

Altering input data to identify important features is a basic approach to many types of AI model. But the task becomes more challenging in more complex neural networks, says Anupam Datta, a computer scientist at Carnegie Mellon University in Pittsburgh, Pennsylvania. In those complex cases, scientists want to tease out not just which features play a part in the decision-making and how big that role is, but also how the importance of a feature alters in relation to changes in other features. "The causality element still carries over because we are trying to still figure out which features have the highest causal effect on the model's prediction," Datta says. "But the mechanism for measuring it changes a little bit." As with Saenko's saliency maps, he systematically blocks out individual pixels in images. A mathematical value can then be assigned to that portion of the image, representing the magnitude of the change that results from obscuring that part. Seeing which pixels are most important tells Datta which neurons in the hidden layers have the greatest role in the outcome, helping him to map the model's internal structure and draw conclusions about the concepts it has learnt[5].

> "We were only focused on the promoters for gene regulation, but the AI found clues in sequences in genes that the researchers would have ignored."

Advances from explanation

Another way DeGrave and Janizek measured saliency relied on a complex type of neural network known as a generative adversarial network (GAN). A typical GAN consists of a pair of networks. One generates data – an image of a street, for instance – and the other tries to determine whether the output is real or fake.

The two networks continue to interact in this way until the first network is reliably creating images that can fool the other. In their case, the Washington researchers asked a GAN to turn COVID-positive X-rays into COVID-negative images[3]. By seeing which aspects of the X-rays it altered, the researchers could see what part of the image the computer considered important to its diagnosis.

Although the basic principle of a GAN is straightforward, the subtle dynamics of the pair of networks is not well understood. "The way that a GAN generates images is quite mysterious," says Antonio Torralba, a computer scientist at the Massachusetts Institute of Technology in Cambridge, who is trying to solve this enigma. Given a random input of numbers, it eventually outputs a picture that looks real. This approach has been used to create photos of faces that don't exist and produce news stories that read as if they were written by a person.

Torralba and his team decided to dissect a GAN and look at what the individual neurons were doing. Just like Datta, they found some neurons focused on specific concepts[6]. "We found groups of units that were responsible for drawing trees, other groups responsible for drawing buildings, and some units drawing doors and windows," he says. And just as Saenko's models had learnt that flowers suggest a vase, units in his GAN also learnt from context. One developed a detector for beds to decide whether a scene was a bedroom, and another learnt that doors don't usually exist in trees.

Being able to recognize which neurons are identifying or producing which objects opens up the possibility of being able to refine a neural network without having to show it thousands of new photographs, Torralba says. If a model has been trained to recognize cars, but all the images it trained on were of cars on a paved surface, it might fail when shown a picture of a car on snow. But a computer scientist who understands the model's internal connections might be able to tweak the model to recognize a layer of snow as equivalent to a paved surface. Similarly, a computer special-effects designer who might want to automate the creation of an impossible scene could re-engineer the model by hand to accomplish that.

Another value of explainability is that the way a machine performs a task might provide the people watching it with some insight into how they could do things differently or better themselves. Computational biologist Laura-Jayne Gardiner trained an AI to predict which genes were at work in regulating circadian clocks, internal molecular timers that govern a range of biological processes[7]. Gardiner and her colleagues at IBM Research Europe and the Earlham Institute, a life-sciences research group in Norwich, UK, also made the computer highlight the features that it used to decide whether a gene was likely to play a part in circadian rhythm. Its approach was surprising. "We were only focused on the promoters for gene regulation," Gardiner says, but the AI found clues in sequences in genes that the researchers would have ignored. "You end up with this ranked list of the features," Gardiner explains; the team can use this in its lab-based research to further refine its understanding of the biology.

Accuracy and trust

Coming up with explanations is a start, but there should also be a way to quantify their accuracy, says Pradeep Ravikumar, a computer scientist at Carnegie Mellon University who is working on ways to automate such evaluation[8]. Explanations that seem to make sense to a human could in fact prove to have little relation to what the model is actually doing.

"The question of how to objectively evaluate explanations is still in its early stages," Ravikumar says. "We need to get better explanations and also better ways to evaluate explanations." One way to test the veracity of an explanation is to make small changes to the features that it says are important. If they truly are, these minor changes in the input should lead to big changes in the output. Similarly, large alterations to irrelevant features – say, removing a bus from a picture of a cat – should not affect the results. If the evaluation system goes one step further and predicts not just which features are important, but also how the model's answer would change if small changes were made to those features, this can also be tested. "If an explanation was actually explaining the model, then it would have a better sense of how the model would behave with these small changes," Ravikumar says.

The search for explanations can sometimes seem like so much work that many computer scientists might be tempted to skip it, and take the AI's results at face value. But at least some level of explainability is relatively simple – saliency maps, for instance, can now be generated quickly and inexpensively, Janizek says. By contrast, training and using a GAN is more complex and time-consuming. "You definitely have to have pretty good familiarity with deep-learning stuff, and a nice machine with some graphics processing units to get it to work," Janizek says. A third method his group tried – altering a few hundred images manually with photo-editing software to identify whether a feature was important – was even more labour intensive.

Saenko says many researchers in the machine-learning community have also tended to see a trade-off between explainability and accuracy. They think that the level of detail and the number of calculations that make neural networks more accurate than smaller decision trees also put them out of reach of all human comprehension. But some are questioning whether that trade-off is real, Janizek says. "It could end up being the case that a more interpretable model is a more useful model and a more accurate model."

It's also beginning to look as if some of the patterns that neural networks can pick out that are imperceptible to people might not be as important as computer scientists once thought, he adds. "How often are they something that's truly predictive in a way that's going to generalize across environments? And how often are they some weird kind of source-specific noise?"

However big or small the challenge of explainability might be, a good explanation is not always going to be enough to convince users to rely on a system, Ravikumar says. Knowing why an AI assistant, such as Amazon's Alexa, answered a question in a certain way might not foster trust among users as much as, say, laws that prohibit the misuse of recordings of private conversations. Perhaps physicians will need clinical evidence that a computer's diagnoses have proved right over time, and a verified biological reason why the factors the computer is looking at should be relevant. And policymakers might require that some protections regarding the use of such systems be written into law. "These are broader questions that I think the community hasn't really thought too deeply about," Ravikumar says.

However, in the area of explanations, AI researchers have been making strides. Although there might still be specifics to be worked out to cover the variety of machine-learning models in use, the problem will be cracked, probably in a year or two, says Torralba. People, he says, "always talk about this black box, and we don't think that neural networks are black boxes. If they are working really well, then if you look inside, what they do makes sense." ■

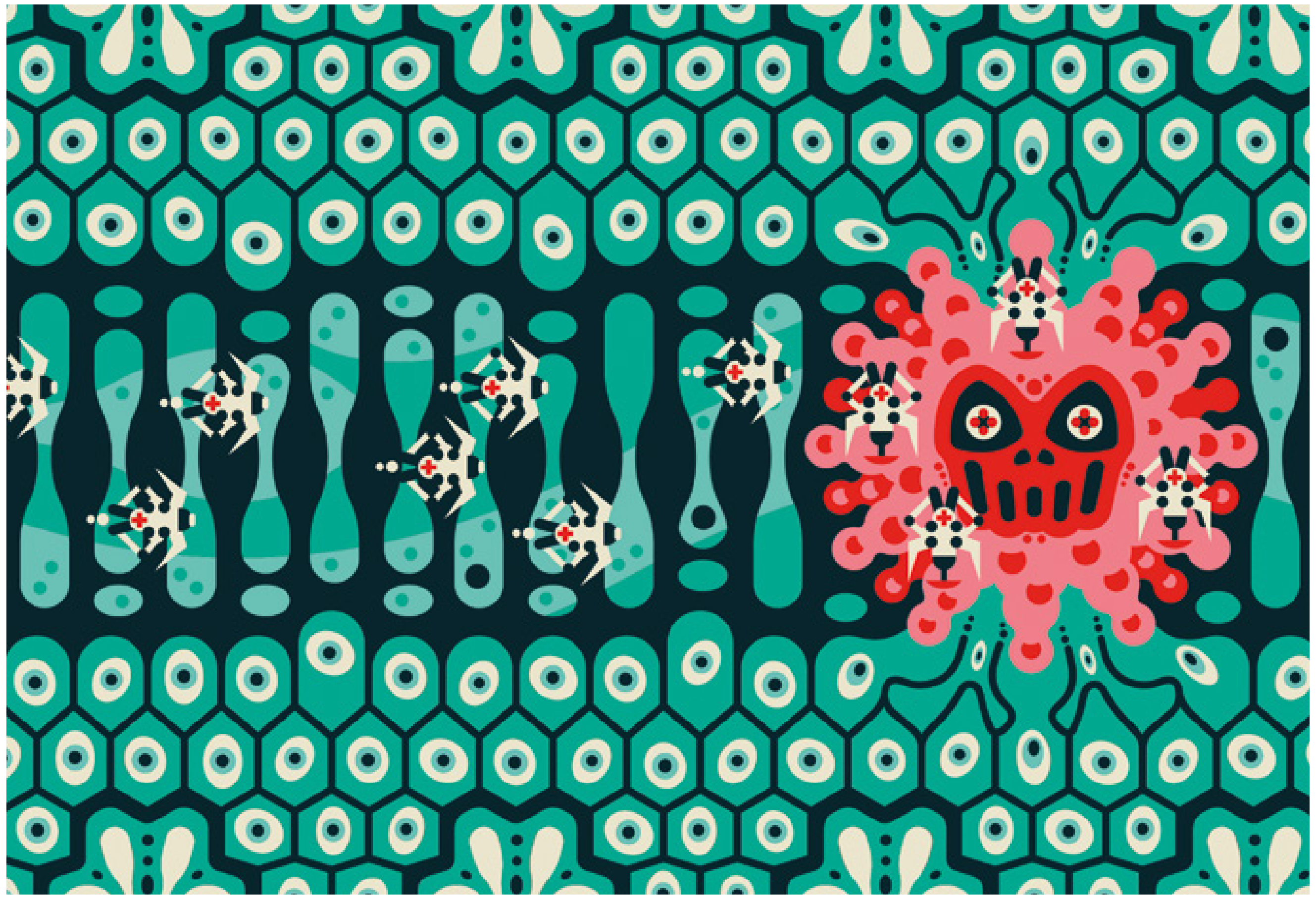
Breaking into the black box of artificial intelligence

Read this article online

REFERENCES

1. Wang, L. & Wong, A. Preprint at https://arxiv.org/abs/2003.09871 (2020).
2. Ai, T. *et al. Radiology* **296**, E32–E40 (2020).
3. DeGrave, A. J., Janizek, J. D. & Lee, S.-I. *Nature Mach*. Intell. **3**, 610–619 (2021).
4. Petsiuk, V. *et al*. Preprint at https://arxiv.org/abs/2006.03204 (2020).
5. Leino, K., Sen, S., Datta, A., Fredrikson, M. & Li, L. Preprint at https://arxiv.org/abs/1802.03788 (2018).
6. Bau, D. *et al*. Preprint at https://arxiv.org/abs/1811.10597 (2018).
7. Gardiner, L.-J. *et al. Proc. Natl Acad. Sci. USA* **118**, e2103070118 (2021).
8. Yeh, C.-K. & Ravikumar, P. *Appl. AI Lett*. **2**, e57 (2021).

JAN KALLWEJT



Miniature medical robots step out from sci-fi

Tiny machines that deliver therapeutic payloads to precise locations in the body are the stuff of science fiction. But some researchers are trying to turn them into a clinical reality. By Anthony King

Cancer drugs usually take a scattergun approach. Chemotherapies inevitably hit healthy bystander cells while blasting tumours, sparking a slew of side effects. It is also a big ask for an anticancer drug to find and destroy an entire tumour – some are difficult to reach, or hard to penetrate once located.

A long-dreamed-of alternative is to inject a battalion of tiny robots into a person with cancer. These miniature machines could navigate directly to a tumour and smartly deploy a therapeutic payload right where it is needed. “It is very difficult for drugs to penetrate through biological barriers, such as the blood–brain barrier or mucus of the gut, but a microrobot can do that,” says Wei Gao, a medical engineer at the California Institute of Technology in Pasadena.

Among his inspirations is the 1966 film Fantastic Voyage, in which a miniaturized submarine goes on a mission to remove a blood clot in a scientist’s brain, piloted through the bloodstream by a similarly shrunken crew. Although most of the film remains firmly in the realm of science fiction, progress on miniature medical machines in the past ten years has seen experiments move into animals for the first time.

There are now numerous micrometre- and nanometre-scale robots that can propel themselves through biological media, such as the matrix between cells and the contents of the gastrointestinal tract. Some are moved and steered by outside forces, such as magnetic fields and ultrasound. Others are driven by onboard chemical engines, and some are even built on top of bacteria and human cells to take advantage of those cells’ inbuilt ability to get around. Whatever the source of propulsion, it is hoped that these tiny robots will be able to deliver therapies to places that a drug alone might not be able to reach, such as into the centre of solid tumours. However, even as those working on medical nano- and microrobots begin to collaborate more closely with clinicians, it is clear that the technology still has a long way to go on its fantastic journey towards the clinic.

Poetry in motion

One of the key challenges for a robot operating inside the human body is getting around. In *Fantastic Voyage*, the crew uses blood vessels to move through the body. However, it is here that reality must immediately diverge from fiction. “I love the movie,” says roboticist Bradley Nelson, gesturing to a copy of it in his office at the Swiss Federal Institute of Technology (ETH) Zurich in Switzerland. “But the physics are terrible.” Tiny robots would have severe difficulty swimming against the flow of blood, he says. Instead, they will initially be administered locally, then move towards their targets over short distances.

When it comes to design, size matters. “Propulsion through biological media becomes a lot easier as you get smaller, as below a micron bots slip between the network of macromolecules,” says Peer Fischer, a robotics researcher at the Max Planck Institute for Intelligent Systems in Stuttgart, Germany. Bots are therefore typically no more than 1–2 micrometres across. However, most do not fall below 300 nanometres. Beyond that size, it becomes more challenging to detect and track them in biological media, as well as more difficult to generate sufficient force to move them.

Scientists have several choices for how to get their bots moving. Some opt to provide power externally. For instance, in 2009, Fischer – who was working at Harvard University in Cambridge, Massachusetts, at the time, alongside fellow nanoroboticist Ambarish Ghosh – devised a glass propeller, just 1–2 micrometres in length, that could be rotated by a magnetic field[1]. This allowed the structure to move through water, and by adjusting the magnetic field, it could be steered with micrometre precision. In a 2018 study[2], Fischer launched a swarm of micropropellers into a pig’s eye *in vitro*, and had them travel over centimetre distances through the gel-like vitreous humour into the retina – a rare demonstration of propulsion through real tissue. The swarm was able to slip through the network of biopolymers within the vitreous humour thanks in part to a silicone oil and fluorocarbon coating applied to each propeller. Inspired by the slippery surface that the carnivorous pitcher plant *Nepenthes* uses to catch insects, this minimized interactions between the micropropellers and biopolymers.

> “We could load anticancer drugs efficiently into the head of the sperm, into the DNA. Then the sperm can fuse with other cells.”

Another way to provide propulsion from outside the body is to use ultrasound. One group placed magnetic cores inside the membranes of red blood cells[3], which also carried photoreactive compounds and oxygen. The cells’ distinctive biconcave shape and greater density than other blood components allowed them to be propelled using ultrasonic energy, with an external magnetic field acting on the metallic core to provide steering. Once the bots are in position, light can excite the photosensitive compound, which transfers energy to the oxygen and generates reactive oxygen species to damage cancer cells.

This hijacking of cells is proving to have therapeutic merits in other research projects. Some of the most promising strategies aimed at treating solid tumours involve human cells and other single-celled organisms jazzed up with synthetic parts. In Germany, a group led by Oliver Schmidt, a nanoscientist at Chemnitz University of Technology, has designed a biohybrid robot based on sperm cells[4]. These are some of the fastest motile cells, capable of hitting speeds of 5 millimetres per minute, Schmidt says. The hope is that these powerful swimmers can be harnessed to deliver drugs to tumours in the female reproductive tract, guided by magnetic fields. Already, it has been shown that they can be magnetically guided to a model tumour in a dish.

“We could load anticancer drugs efficiently into the head of the sperm, into the DNA,” says Schmidt. “Then the sperm can fuse with other cells when it pushes against them.” At the Chinese University of Hong Kong, meanwhile, nanoroboticist Li Zhang led the creation of microswimmers from *Spirulina* microalgae cloaked in the mineral magnetite. The team then tracked a swarm of them inside rodent stomachs using magnetic resonance imaging5. The biohybrids were shown to selectively target cancer cells. They also gradually degrade, reducing unwanted toxicity.

Another way to get micro- and nanobots moving is to fit them with a chemical engine: a catalyst drives a chemical reaction, creating a gradient on one side of the machine to generate propulsion. Samuel Sánchez, a chemist at the Institute for Bioengineering of Catalonia in Barcelona, Spain, is developing nanomotors driven by chemical reactions for use in treating bladder cancer. Some early devices relied on hydrogen peroxide as a fuel. Its breakdown, promoted by platinum, generated water and oxygen gas bubbles for propulsion. But hydrogen peroxide is toxic to cells even in minuscule amounts, so Sánchez has transitioned towards safer materials. His latest nanomotors are made up of

doi: https://doi.org/10.1038/d41586-022-00859-0 |

| doi: https://doi.org/10.1038/d41586-022-00859-0

honeycombed silica nanoparticles, tiny gold particles and the enzyme urease[6]. These 300–400-nm bots are driven forwards by the chemical breakdown of urea in the bladder into carbon dioxide and ammonia, and have been tested in the bladders of mice. "We can now move them and see them inside a living system," says Sánchez.

Breaking through

A standard treatment for bladder cancer is surgery, followed by immunotherapy in the form of an infusion of a weakened strain of *Mycobacterium bovis* bacteria into the bladder, to prevent recurrence. The bacterium activates the person's immune system, and is also the basis of the BCG vaccine for tuberculosis. "The clinicians tell us that this is one of the few things that has not changed over the past 60 years," says Sánchez. There is a need to improve on BCG in oncology, according to his collaborator, urologic oncologist Antoni Vilaseca at the Hospital Clinic of Barcelona. Current treatments reduce recurrences and progression, "but we have not improved survival", Vilaseca says. "Our patients are still dying."

The nanobot approach that Sánchez is trying promises precision delivery. He plans to insert his bots into the bladder (or intravenously), to motor towards the cancer with their cargo of therapeutic agents to target cancer cells, using abundant urea as a fuel. He might use a magnetic field for guidance, if needed, but a more straightforward replacement of BCG with bots that do not require external control, perhaps using an antibody to bind a tumour marker, would please clinicians most. "If we can deliver our treatment to the tumour cells only, then we can reduce side effects and increase activity," says Vilaseca.

Miniature medical robots step out from sci-fi

Read this article online

Not all cancers can be reached by swimming through liquid, however. Natural physiological barriers can block efficient drug delivery. The gut wall, for example, allows absorption of nutrients into the bloodstream, and offers an avenue for getting therapies into bodies. "The gastrointestinal tract is the gateway to our body," says Joseph Wang, a nanoengineer at the University of California, San Diego. However, a combination of cells, microbes and mucus stops many particles from accessing the rest of the body. To deliver some therapies, simply being in the intestine isn't enough – they also need to be able to burrow through its defences to reach the bloodstream, and a nanomachine could help with this.

In 2015, Wang and his colleagues, including Gao, reported the first self-propelled robot in vivo, inside a mouse stomach[7]. Their zinc-based nanomotor dissolved in the harsh stomach acids, producing hydrogen bubbles that rocketed the robot forwards. In the lower gastrointestinal tract, they instead use magnesium. "Magnesium reacts with water to give a hydrogen bubble," says Wang. In either case, the metal micromotors are encapsulated in a coating that dissolves at the right location, freeing the micromotor to propel the bot into the mucous wall.

Some bacteria have already worked out their own ways to sneak through the gut wall. *Helicobacter pylori*, which causes inflammation in the stomach, excretes urease enzymes to generate ammonia and liquefy the thick mucous that lines the stomach wall. Fischer envisages future micro- and nanorobots borrowing this approach to deliver drugs through the gut.

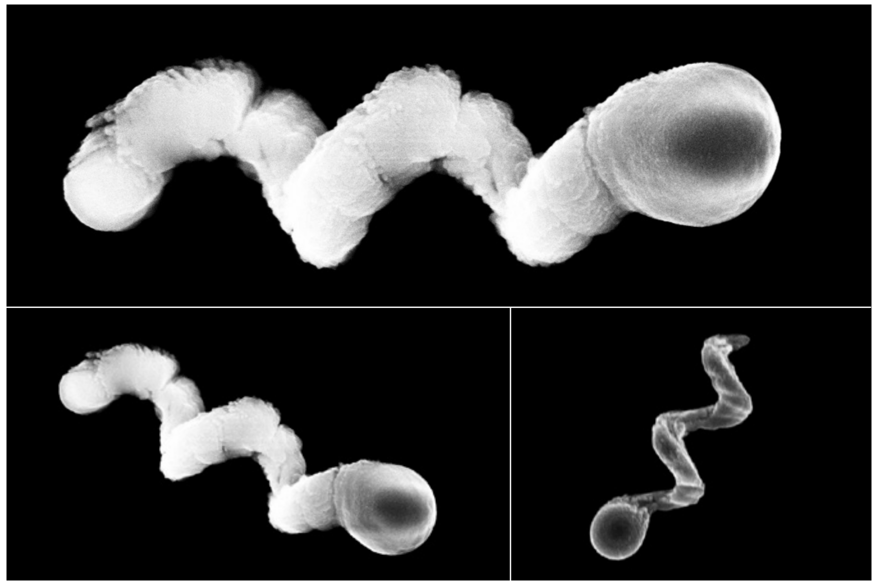
Solid tumours are another difficult place to deliver a drug. As these malignancies develop, a ravenous hunger for oxygen promotes an outside surface covered with blood vessels, while an oxygen-deprived core builds up within. Low oxygen levels force cells deep inside to switch to anaerobic metabolism and churn out lactic acid, creating acidic conditions. As the oxygen gradient builds, the tumour becomes increasingly difficult to penetrate. Nanoparticle drugs lack a force with which to muscle through a tumour's fortifications, and typically less than 2% of them will make it inside8. Proponents of nanomachines think that they can do better.

Sylvain Martel, a nanoroboticist at Montreal Polytechnic in Canada, is trying to break into solid tumours using bacteria that naturally contain a chain of magnetic iron-oxide nanocrystals. In nature, these *Magnetococcus* species seek regions that have low oxygen. Martel has engineered such a bacterium to target active cancer cells deep inside tumours[8]. "We guide them with a magnetic field towards the tumour," explains Martel, taking advantage of the magnetic crystals that the bacteria typically use like a compass for orientation. The precise locations of low-oxygen regions are uncertain even with imaging, but once these bacteria reach the right location, their autonomous capability kicks in and they motor towards low-oxygen regions. In a mouse, more than half the bacteria injected close to tumour grafts broke into this tumour region, each laden with dozens of drug-loaded liposomes. Martel cautions, however, that there is still some way to go before the technology is proven safe and effective for treating people with cancer.

In the Netherlands, chemist Daniela Wilson at Radboud University in Nijmegen and colleagues have developed enzyme-driven nanomotors powered by DNA that might similarly be able to autonomously home in on tumour cells[9]. The motors navigate towards areas that are richer in DNA, such as tumour cells that undergoing apoptosis. "We want to create systems that are able to sense gradients by different endogenous fuels in the body," Wilson says, suggesting that the higher levels of lactic acid or glucose typically found in tumours could also be used for targeting. Once in place, the autonomous bots seem to be picked up by cells more easily than passive particles are – perhaps because the bots push against cells.

Fiction versus reality

Inspirational though Fantastic Voyage might have been for many working in the field of medical nanorobotics, there are some who think the film has become a burden. "People think of this as science fiction, which excites people, but on the other hand they don't take it so seriously," says Martel. Fischer is similarly jaded by movie-inspired hype. "People sometimes write very liberally as if nanobots for cancer treatment are almost here," he says. "But this is not even in clinical trials right now."



CORNELIA MIKSCH, MPI-IS

"We could load anticancer drugs efficiently into the head of the sperm, into the DNA. Then the sperm can fuse with other cells."

▲ An electron microscope image of a glass nanopropeller.

Nonetheless, advances in the past ten years have raised expectations of what is possible with current technology. "There's nothing more fun than building a machine and watching it move. It's a blast," says Nelson. But having something wiggling under a microscope no longer has the same draw, without medical context. "You start thinking, 'how could this benefit society?'" he says.

With this in mind, many researchers creating nanorobots for medical purposes are working more closely with clinicians than ever before. "You find a lot of young doctors who are really interested in what the new technologies can do," Nelson says. Neurologist Philipp Gruber, who works with stroke patients at Aarau Cantonal Hospital in Switzerland, began a collaboration with Nelson two years ago after contacting ETH Zurich. The pair share an ambition to use steerable microbots to dissolve clots in people's brains after ischaemic stroke – either mechanically, or by delivering a drug. "Brad knows everything about engineering," says Gruber, "but we can advise about the problems we face in the clinic and the limitations of current treatment options."

Sánchez tells a similar story: while he began talking to physicians around a decade ago, their interest has warmed considerably since his experiments in animals began three to four years ago. "We are still in the lab, but at least we are working with human cells and human organoids, which is a step forward," says his collaborator Vilaseca.

As these seedlings of clinical collaborations take root, it is likely that oncology applications will be the earliest movers – particularly those that resemble current treatments, such as infusing microbots instead of BCG into cancerous bladders. But even these therapeutic uses are probably at least 7–10 years away. In the nearer term, there might be simpler tasks that nanobots can be used to accomplish, according to those who follow the field closely.

For example, Martin Pumera, a nanoroboticist at the University of Chemistry and Technology in Prague, is interested in improving dental care by landing nanobots beneath titanium tooth implants[10]. The tiny gap between the metal implants and gum tissue is an ideal niche for bacterial biofilms to form, triggering infection and inflammation. When this happens, the implant must often be removed, the area cleaned, and a new implant installed – an expensive and painful procedure. He is collaborating with dental surgeon Karel Klíma at Charles University in Prague.

Another problem the two are tackling is oral bacteria gaining access to tissue during surgery of the jaws and face. "A biofilm can establish very quickly, and that can mean removing titanium plates and screws after surgery, even before a fracture heals," says Klíma. A titanium oxide robot could be administered to implants using a syringe, then activated chemically or with light to generate active oxygen species to kill the bacteria. Examples a few micrometres in length have so far been constructed, but much smaller bots – only a few hundred nanometres in length – are the ultimate aim.

Clearly, this is a long way from parachuting bots into hard-to-reach tumours deep inside a person. But the rising tide of *in vivo* experiments and the increasing involvement of clinicians suggests that microrobots might just be leaving port on their long journey towards the clinic. ■

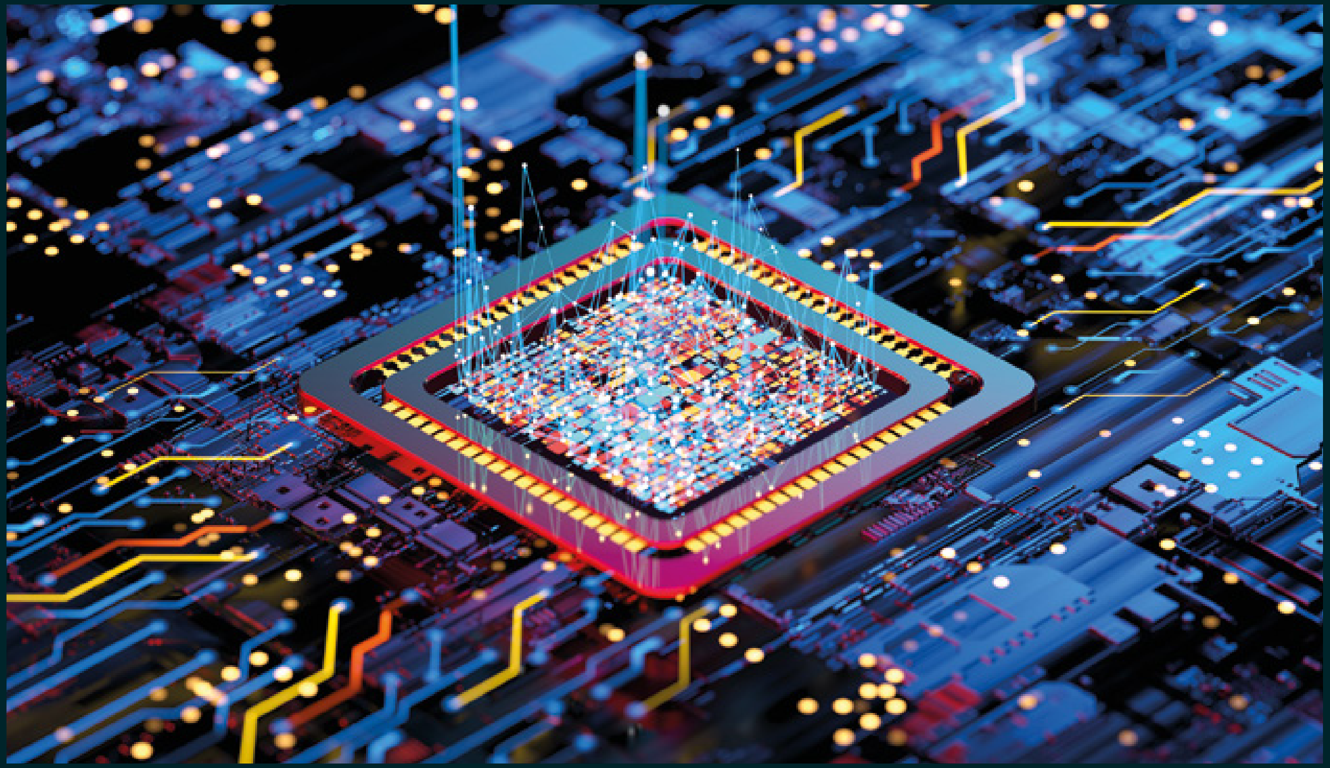
REFERENCES

1. Ghosh, A. & Fischer, P. *Nano Lett*. **9**, 2243–2245 (2009).
2. Wu, Z. *et al. Sci. Adv*. **4**, eaat4388 (2018).
3. Gao, C. *et al. ACS Appl. Mater. Interfaces* **11**, 23392–23400 (2019).
4. Xu, H. *et al. ACS Nano* **12**, 327–337 (2018).
5. Yan, X. *et al. Sci. Robot*. **2**, eaaq1155 (2017).
6. Hortelao, A. C. *et al. Sci. Robot*. **6**, eabd2823 (2021).
7. Gao, W. *et al. ACS Nano* **9**, 117–123 (2015).
8. Felfoul, O. *et al. Nature Nanotechnol*. **11**, 941–947 (2016).
9. *Ye, Y.* et al. Nano Lett. **21**, 8086–8094 (2021).
10. Villa, K. *et al. Cell Rep. Phys. Sci.* **1**, 100181 (2020).

doi: https://doi.org/10.1038/d41586-022-00859-0 |

| doi: https://doi.org/10.1038/d41586-022-00859-0

FII INSTITUTE Future Investment Initiative Institute Impact on Humanity


MF3D/GETTY

The challenge of making moral machines

Artificial intelligence has the potential to improve industries, markets and lives – but only if we can trust the algorithms.

▲ As applications for AIs proliferate, so are questions about ethical development and embedded bias.

In the waning days of 2020, Timnit Gebru, an artificial intelligence (AI) ethicist at Google, submitted a draft of an academic paper to her employer. Gebru and her collaborators had analysed natural language processing (NLP), and specifically the data-intensive approach of training NLP artificial intelligences (AIs). Such AIs can accurately interpret documents produced by humans, and respond naturally to human commands or queries.

In their study, the team found the process of training a NLP AI requires immense resources and creates a considerable risk of embedding significant bias into the AI. That bias can lead to inappropriate or even harmful responses. Google was skeptical of the paper's conclusions, and was displeased that Gebru had submitted it to a prominent conference. The company asked Gebru either to retract the paper or remove any mention of Google affiliations. Gebru refused the terms. Within a day, she learned that she no longer had a job.

Gebru's sudden ouster raised serious questions about the transparency, accountability and safety of AI development, particularly in private companies. It also crystalized concerns about AI algorithms that had been bubbling along for years.

Whether embedded in a natural-language processor or a medical diagnostic, AI algorithms can carry unintentional biases, and those biases can have real-world consequences. The manipulation of the Facebook algorithm to impact the 2016 United States presidential election is one frequently cited example. As another, Aimee van Wynsberghe, an AI ethicist at the University of Bonn in Germany, cites an abortive effort by Amazon to use an AI-based recruiting tool. The tool, which was tested between 2014 and 2017, drew the wrong lessons from the company's past hiring patterns.

"When they put it in practice, they found that the algorithm would not select women for the higher-level positions, only for lower-level ones," says van Wynsberghe.

Yet the development of AI continues to accelerate. The market for AI software is expected to reach US$63 billion in 2022, according to Gartner Research, and that is on top of 20% growth in 2021. Already commonplace in online tools such as recommendation or optimization engines and translation services, higher impact AI applications are on the horizon, particularly in large sectors like energy, include those in transportation, healthcare, manufacturing, drug development and sustainability.

Given the size and number of opportunities, the enthusiasm for AI solutions can obscure risks associated with them. As Gebru found, AIs have the potential to cause real harm. If humans can't trust the very machines meant to help them, the true promise of the technology may never be fulfilled.

Smarter by the day

Although many AIs are programmed directly by humans, most modern implementations are built on artificial neural networks. The algorithms analyse data to identify and extract patterns, essentially 'learning' about the world as they go. The interpretations of these data guide the next step of analysis, or inform decisions made by the algorithm.

Artificial neural networks analyse data collaboratively in a manner roughly analogous to the neurons in the human brain, explains Jürgen Schmidhuber, director of KAUST in Saudi Arabia. He developed a foundational neural network framework known as 'long short-term memory' (LSTM) in the late 1990s.

"In the beginning, the learning machine knows nothing – all the connections are random," he says. "But then over time, it makes some of the connections stronger and some of them weaker, until the whole thing can do interesting things."

Such training is a characteristic of LSTM and other approaches to neural networks, and it's a reason those AIs have become so popular. An AI that learns to learn has the potential to develop novel solutions to extremely difficult problems. The FII Institute THINK initiative, for example, is pursuing a multi-pronged roadmap for AI development to explore healthcare applications such as drug discovery and epidemic control, as well as sustainability-oriented efforts to monitor and protect forest and marine ecosystems – all of which lend themselves to AI applications.

> "We can't just build robots that are 'ethical' – you have to ask ethical for whom, where and when."

But training can build bad habits as easily as good ones. As Gebru found with NLP AIs, very large and improperly curated data sets can amplify rather than rectify human biases in an AI's decision-making process. Sandra Wachter, a researcher specializing in data ethics at the University of Oxford in the United Kingdom, highlights the example of diagnostic software tools designed to detect signs of skin cancer through image analysis, which fare poorly on black- or brown-skinned individuals because they were primarily trained on data from Caucasian patients. "It might be misdiagnosing you in a way that could actually have harmful consequences for your health and might even be lethal," she says.

Similar training data problems have plagued IBM's AI-driven Watson Health platform, and the company recently moved to divest itself of this technology after years of struggling with poor diagnostic performance and ill-advised treatment recommendations.

Such cases beg the question: Who is to blame when an algorithm does not work as designed? Answers may be easy to reach when an AI's conclusions are objectively wrong, as in certain medical diagnostics. But other situations are much more ambiguous.

For years, Facebook enabled companies to target their advertising based on algorithmically derived information that allowed the platform to infer a user's race, an option now discontinued. "Black people wouldn't be able to see certain job advertisements, or advertisements for housing or financial services, for example," says Wachter. "But those people didn't know about it."

The victims of discrimination might have a claim in the courts after the fact. But the best solution is to pre-empt the introduction of destructive bias in the first place with ethical AI design.

Rules for robots

The idea of imbuing machines with ethics is not new. Author Isaac Asimov penned his Three Laws of Robotics when thinking of androids more than 75 years ago, and all three of his laws raise ethical considerations. In the research labs around the world, science fiction is now edging towards reality as researchers grapple with how to embed ethics into AI.

Current work entails identifying sets of internal guidelines that would be compatible with human laws, norms, and moral expectations, and could serve to keep AIs from making harmful or otherwise inappropriate decisions. Van Wynsberghe pushes back against the idea of calling such AI systems 'ethical machines' per se. "It's like a sophisticated toaster," she says. "This is about embedding ethics into the procedure of making the machines."

In 2018, the Institute of Electrical and Electronics Engineers (IEEE), a non-profit organization headquartered in New York City, US, convened an interdisciplinary group of hundreds of experts from around the world to hash out some of the core principles underlying 'ethically aligned design' for AI systems. Bertram Malle, a cognitive scientist specializing in human-robot interaction at Brown University in Providence, Rhode Island, US, who co-chaired one of the effort's working groups, says, "We can't just build robots that are 'ethical' – you have to ask ethical for whom, where and when." Accordingly, the ethical framework for any given AI, Malle says, should be developed with close input from the communities of people with which they will ultimately be interacting.

A recent law review article from Wachter's team highlighted some of this complexity. After assessing a variety of metrics designed to assess the level of bias in an AI system, her team determined that 13 out of 20 failed to meet the legal guidelines of the European Union's non-discrimination law.

"One of the explanations is because the majority, if not all, of those bias tests were developed in the US… under North American assumptions," she says. This work was conducted in collaboration with Amazon, and the company has subsequently adopted an improved bias-testing system based on the open-source toolkit that resulted from the study.

A trustworthy AI system also requires a measure of transparency, where users can get a clear sense of how an algorithm arrived at a particular decision or outcome. This can be tricky, given the 'black box' complexity and proprietary nature of many AI systems, but is not an insurmountable problem. "Building systems that are completely transparent is both unrealistic and unnecessary," says Malle. "We need to have systems that can answer the kinds of questions that humans have."

That has been another priority for Wachter's team, which uses a strategy called 'counterfactual explanation' to probe AI systems with different inputs in order to determine which factors lead to which outcomes. She cites the example of interrogating diagnostic software with different metabolic parameters to understand how the algorithm determines that a patient has diabetes.

Ethics for all

If embedding ethics and transparency into AI is a difficult problem, the ethical and transparent development of AI, by humans, could be even more challenging. Private companies like Google, Facebook, Baidu and Tesla account for a large portion of overall AI development, while new start-ups seem to emerge on a weekly basis. Ethical oversight in such settings can vary considerably.

> "We see glimmers of hope, where [companies] have hired their own ethicists. The problem is that they're not transparent about what the ethicists are doing – it's all behind non-disclosure agreements."

"We see glimmers of hope, where [companies] have hired their own ethicists," van Wynsberghe says. "The problem is that they're not transparent about what the ethicists are doing, what they're learning – it's all behind non-disclosure agreements." The firing of Gebru and other ethicists highlights the precariousness of allowing companies to police themselves.

But there are potential solutions. To overcome the opacity of private AI development, for example, van Wynsberghe advocates the notion that companies could collectively sponsor an independent ethical review organization to act analogously to the institutional review boards that supervise clinical trials. In this approach, corporations would collectively fund a board of ethicists to take on rotating 'shifts' at the companies to oversee work. "So you'd have this kind of flow of information and shared experiences and whatnot, and the ethicists are not dependent on the company for their paycheck," she says. "Otherwise, they're scared to speak up."

New legal frameworks could help as well, and Wachter believes that many companies are likely to welcome some guidance rather than operating in an environment of uncertainty and risk. "Now examples are being put on the table that concretely tell them what it means to be accountable, what it means to be bias-free, and what it means to protect privacy," she says.

The European Union currently leads the way, with an 'AI Act' that provides a detailed framework for the risk-based assessment of where AI systems can be deployed safely and ethically. China is also implementing strict regulations designed to prevent AI-based exploitation of or discrimination against users – although these same regulations could also provide a vehicle for further censorship of online speech.

Above all, automation should not be seen as a universal solution and the collective good, for all humans not just AI developers, should always be a consideration. Malle favours a focus on systems that complement rather than replace human expertise in areas such as education, healthcare and social services. For example, AI could help overextended teachers to get a better handle on students who need more individual attention or are struggling in particular areas of the curriculum. Or AI could take care of routine tasks in the hospital ward, so that nurses can better focus on the specific needs of their patients.

The goal should be to amplify what can be achieved with available human intellect, expertise and judgement – not to take those out of the equation altogether. "I really see opportunities in the domains where we really don't have enough humans or not enough trained humans," Malle says. "Let's think about domains of need first." ■

PETER CROWTHER

Interested to learn more?

To keep track of the latest and greatest research in robotics and artificial intelligence from across the Nature Portfolio journals, as well as reports from journalists on topics of special interest, sign up for the newsletter here: