

AI NEWS VOL 7

THE FUTURE OF QUANTUM COMPUTING

The Era of Quantum Computing Is Here. Outlook: Cloudy

Quantum computers should soon be able to beat classical computers at certain basic tasks. But before they're truly powerful, researchers have to overcome a number of fundamental roadblocks.

Quantum computers have to deal with the problem of noise, which can quickly derail any calculation.

Josef Bsharah([opens a new tab](#)) for Quanta Magazine

Introduction

After decades of heavy slog with no promise of success, quantum computing is suddenly buzzing with almost feverish excitement and activity. Nearly two years ago, IBM made a quantum computer available to the world: the 5-



quantum-bit (qubit) resource they now call (a little awkwardly) the IBM Q experience. That seemed more like a toy for researchers than a way of getting any serious number crunching done. But 70,000 users worldwide have

registered for it, and the qubit count in this resource has now quadrupled. In the past few months, [IBM](#)(opens a new tab) and [Intel](#)(opens a new tab) have announced that they have made quantum computers with 50 and 49 qubits, respectively, and Google is thought to have one waiting in the wings. “There is a lot of energy in the community, and the recent progress is immense,” said physicist [Jens Eisert](#)(opens a new tab) of the Free University of Berlin.

There is now talk of impending “quantum supremacy”: the moment when a quantum computer can carry out a task beyond the means of today’s best classical supercomputers. That might sound absurd when you compare the bare numbers: 50 qubits versus the billions of classical bits in your laptop. But the whole point of quantum computing is that a quantum bit counts for much, much more than a classical bit. Fifty qubits has long been considered the approximate number at which quantum computing becomes [capable of calculations](#)(opens a new tab) that would take an unfeasibly long time classically. Midway through 2017, researchers at Google announced that they hoped to have demonstrated quantum supremacy by the end of the year. (When pressed for an update, a spokesperson recently said that “we hope to announce results as soon as we can, but we’re going through all the detailed work to ensure we have a solid result before we announce.”)

It would be tempting to conclude from all this that the basic problems are solved in principle and the path to a future of ubiquitous quantum computing is now just a matter of engineering. But that would be a mistake. The fundamental physics of quantum computing is far from solved and can’t be readily disentangled from its implementation.

Even if we soon pass the quantum supremacy milestone, the next year or two might be the real crunch time for whether quantum computers will revolutionize computing. There’s still everything to play for and no guarantee of reaching the big goal.



IBM's quantum computing center at the Thomas J. Watson Research Center in Yorktown Heights, New York, holds quantum computers in large cryogenic tanks (far right) that are cooled to a fraction of a degree above absolute zero.

Shut Up and Compute

Both the benefits and the challenges of quantum computing are inherent in the physics that permits it. The basic story has been told many times, though not always with the nuance that quantum mechanics demands. Classical computers encode and manipulate information as strings of binary digits — 1 or 0. Quantum bits do the same, except that they may be placed in a so-called superposition of the states 1 and 0, which means that a measurement of the qubit's state could elicit the answer 1 or 0 with some well-defined probability.

To perform a computation with many such qubits, they must all be sustained in interdependent superpositions of states — a “quantum-coherent” state, in which the qubits are said to be entangled. That way, a tweak to one qubit may influence all the others. This means that somehow computational operations on qubits count for more than they do for classical bits. The computational resources increase in simple proportion to the number of bits for a classical device, but adding an extra qubit potentially doubles the resources of a quantum computer. This is why the difference between a 5-qubit and a 50-qubit machine is so significant.

Note that I've not said — as it often is said — that a quantum computer has an advantage because the availability of superpositions hugely increases the number of states it can encode, relative to classical bits. Nor have I said that entanglement permits many calculations to be carried out in parallel. (Indeed, a strong degree of qubit entanglement isn't essential.) There's an element of truth in those descriptions — some of the time — but none captures the essence of quantum computing.



Inside one of IBM's cryostats wired for a 50-qubit quantum system.
Connie Zhou for IBM

It's hard to say qualitatively why quantum computing is so powerful precisely because it is hard to specify what quantum mechanics means at all. The equations of quantum theory certainly show that it will work: that, at least for

some classes of computation such as factorization or database searches, there is tremendous speedup of the calculation. But how exactly?

Perhaps the safest way to describe quantum computing is to say that quantum mechanics somehow creates a “resource” for computation that is unavailable to classical devices. As quantum theorist [Daniel Gottesman](#)[\(opens a new tab\)](#) of the Perimeter Institute in Waterloo, Canada, put it, “If you have enough quantum mechanics available, in some sense, then you have speedup, and if not, you don’t.”

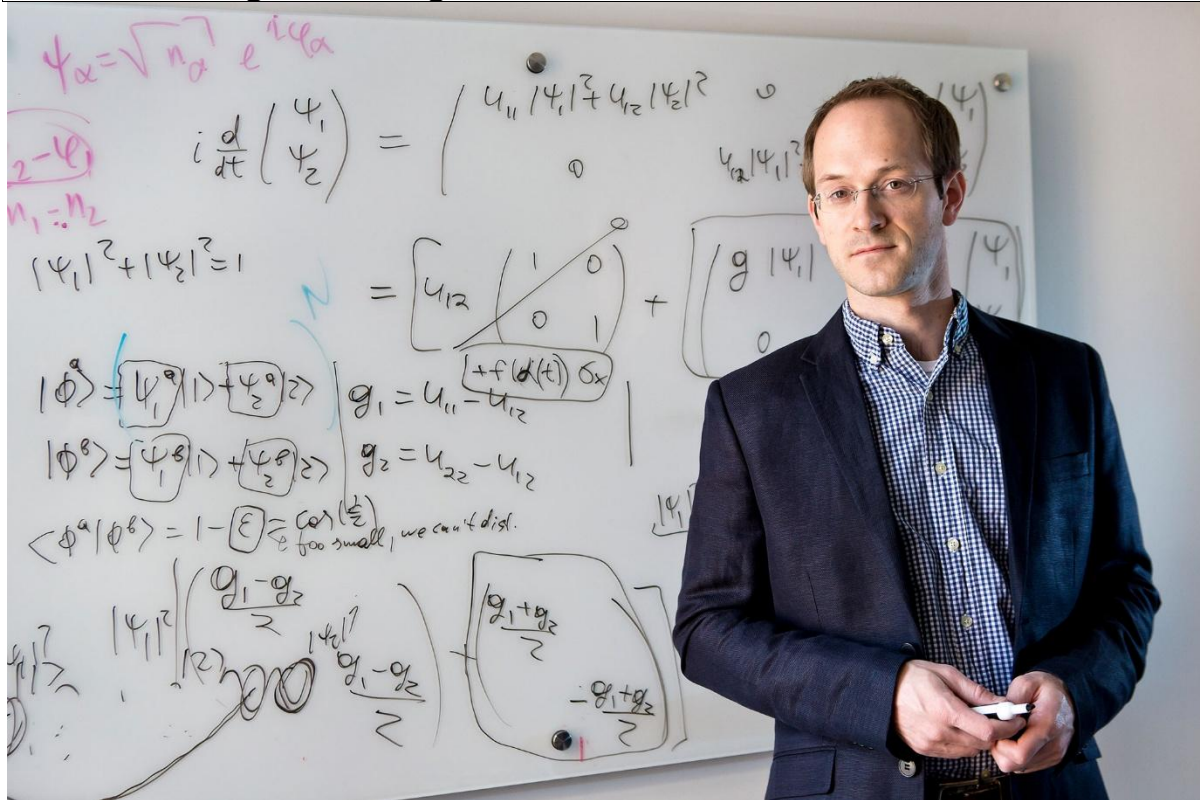
Some things are clear, though. To carry out a quantum computation, you need to keep all your qubits coherent. And this is very hard. Interactions of a system of quantum-coherent entities with their surrounding environment create channels through which the coherence rapidly “leaks out” in a process called decoherence. Researchers seeking to build quantum computers must stave off decoherence, which they can currently do only for a fraction of a second. That challenge gets ever greater as the number of qubits — and hence the potential to interact with the environment — increases. This is largely why, even though quantum computing was first proposed by Richard Feynman in 1982 and the theory was worked out in the early 1990s, it has taken until now to make devices that can actually perform a meaningful computation.

Quantum Errors

There’s a second fundamental reason why quantum computing is so difficult. Like just about every other process in nature, it is noisy. Random fluctuations, from heat in the qubits, say, or from fundamentally quantum-mechanical processes, will occasionally flip or randomize the state of a qubit, potentially derailing a calculation. This is a hazard in classical computing too, but it’s not hard to deal with — you just keep two or more backup copies of each bit so that a randomly flipped bit stands out as the odd one out.

Researchers working on quantum computers have created strategies for how to deal with the noise. But these strategies impose a huge debt of computational overhead — all your computing power goes to correcting errors and not to running your algorithms. “Current error rates significantly limit the lengths of computations that can be performed,” said [Andrew Childs](#)[\(opens a new tab\)](#), the codirector of the Joint Center for Quantum Information and Computer

Science at the University of Maryland. “We’ll have to do a lot better if we want to do something interesting.”



Andrew Childs, a quantum theorist at the University of Maryland, cautions that error rates are a fundamental concern for quantum computers.

Photo by John T. Consoli/University of Maryland

A lot of research on the fundamentals of quantum computing has been devoted to error correction. Part of the difficulty stems from another of the key properties of quantum systems: Superpositions can only be sustained as long as you don’t measure the qubit’s value. If you make a measurement, the superposition collapses to a definite value: 1 or 0. So how can you find out if a qubit has an error if you don’t know what state it is in?

One ingenious scheme involves looking indirectly, by coupling the qubit to another “ancilla” qubit that doesn’t take part in the calculation but that can be probed without collapsing the state of the main qubit itself. It’s complicated to implement, though. Such solutions mean that, to construct a genuine “logical qubit” on which computation with error correction can be performed, you need many physical qubits.

How many? Quantum theorist [Alán Aspuru-Guzik](#) (opens a new tab) of Harvard University estimates that around 10,000 of today’s physical qubits would be

needed to make a single logical qubit — a totally impractical number. If the qubits get much better, he said, this number could come down to a few thousand or even hundreds. Eisert is less pessimistic, saying that on the order of 800 physical qubits might already be enough, but even so he agrees that “the overhead is heavy,” and for the moment we need to find ways of coping with error-prone qubits.

An alternative to correcting errors is avoiding them or canceling out their influence: so-called error mitigation. Researchers at IBM, for example, are [developing](#) schemes for figuring out mathematically how much error is likely to have been incurred in a computation and then extrapolating the output of a computation to the “zero noise” limit.

Some researchers think that the problem of error correction will prove intractable and will prevent quantum computers from achieving the grand goals predicted for them. “The task of creating quantum error-correcting codes is harder than the task of demonstrating quantum supremacy,” said mathematician [Gil Kalai](#) of the Hebrew University of Jerusalem in Israel. And he adds that “devices without error correction are computationally very primitive, and primitive-based supremacy is not possible.” In other words, you’ll never do better than classical computers while you’ve still got errors.

Others believe the problem will be cracked eventually. According to [Jay Gambetta](#), a quantum information scientist at IBM’s Thomas J. Watson Research Center, “Our recent experiments at IBM have demonstrated the basic elements of quantum error correction on small devices, paving the way towards larger-scale devices where qubits can reliably store quantum information for a long period of time in the presence of noise.” Even so, he admits that “a universal fault-tolerant quantum computer, which has to use logical qubits, is still a long way off.” Such developments make Childs cautiously optimistic. “I’m sure we’ll see improved experimental demonstrations of [error correction], but I think it will be quite a while before we see it used for a real computation,” he said.

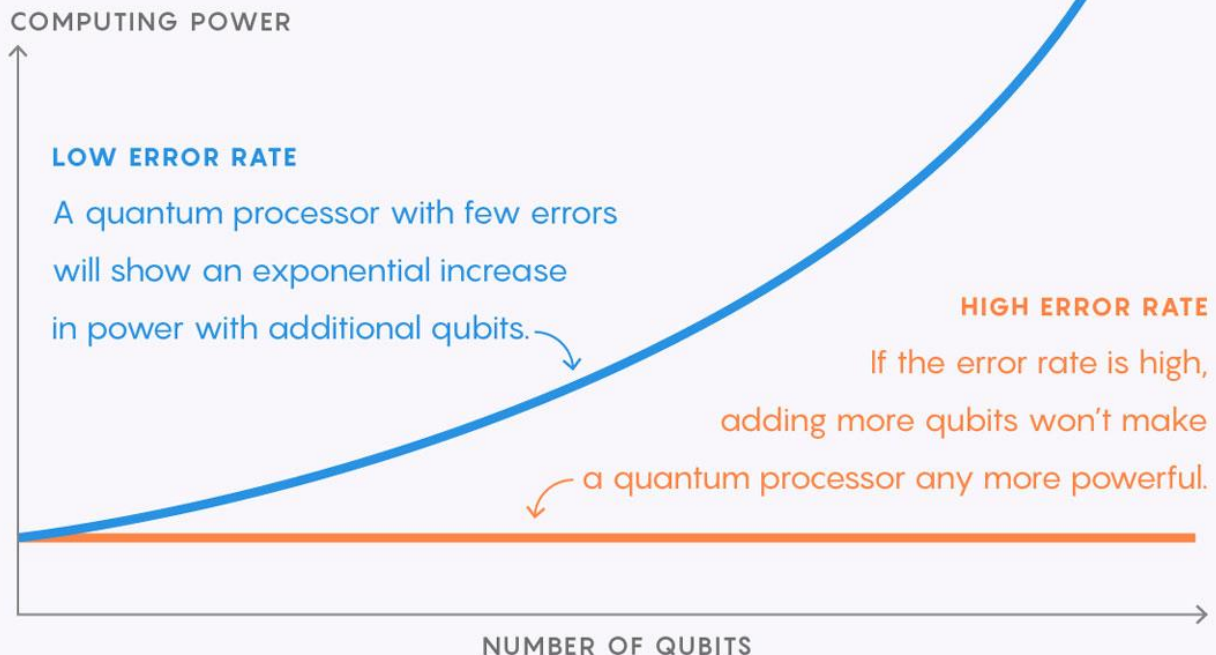
Living With Errors

For the time being, quantum computers are going to be error-prone, and the question is how to live with that. At IBM, researchers are talking about “approximate quantum computing” as the way the field will look in the near term: finding ways of accommodating the noise.

This calls for algorithms that tolerate errors, getting the correct result despite them. It's a bit like working out the outcome of an election regardless of a few wrongly counted ballot papers. "A sufficiently large and high-fidelity quantum computation should have some advantage [over a classical computation] even if it is not fully fault-tolerant," said Gambetta.

Not Just Qubits

Quantum computers will need more than just a lot of qubits.
The machines will also have to have a low error rate.



One of the most immediate error-tolerant applications seems likely to be of more value to scientists than to the world at large: to simulate stuff at the atomic level. (This, in fact, was the motivation that led Feynman to propose quantum computing in the first place.) The equations of quantum mechanics prescribe a way to calculate the properties — such as stability and chemical reactivity — of a molecule such as a drug. But they can't be solved classically without making lots of simplifications.

In contrast, the quantum behavior of electrons and atoms, said Childs, "is relatively close to the native behavior of a quantum computer." So one could

then construct an exact computer model of such a molecule. “Many in the community, including me, believe that quantum chemistry and materials science will be one of the first useful applications of such devices,” said Aspuru-Guzik, who has been at the forefront of efforts to push quantum computing in this direction.

Quantum simulations are proving their worth even on the very small quantum computers available so far. A team of researchers including Aspuru-Guzik has [developed](#)(opens a new tab) an algorithm that they call the variational quantum eigensolver (VQE), which can efficiently find the lowest-energy states of molecules even with noisy qubits. So far it can only handle very small molecules with few electrons, which classical computers can already simulate accurately. But the capabilities are getting better, as Gambetta and coworkers [showed last September](#)(opens a new tab) when they used a 6-qubit device at IBM to calculate the electronic structures of molecules, including lithium hydride and beryllium hydride. The work was “a significant leap forward for the quantum regime,” according to physical chemist [Markus Reiher](#)(opens a new tab) of the Swiss Federal Institute of Technology in Zurich, Switzerland. “The use of the VQE for the simulation of small molecules is a great example of the possibility of near-term heuristic algorithms,” said Gambetta.

But even for this application, Aspuru-Guzik confesses that logical qubits with error correction will probably be needed before quantum computers truly begin to surpass classical devices. “I would be really excited when error-corrected quantum computing begins to become a reality,” he said.

“If we had more than 200 logical qubits, we could do things in quantum chemistry beyond standard approaches,” Reiher adds. “And if we had about 5,000 such qubits, then the quantum computer would be transformative in this field.”

What’s Your Volume?

Despite the challenges of reaching those goals, the fast growth of quantum computers from 5 to 50 qubits in barely more than a year has raised hopes. But we shouldn’t get too fixated on these numbers, because they tell only part of the story. What matters is not just — or even mainly — how many qubits you have, but how good they are, and how efficient your algorithms are.

Any quantum computation has to be completed before decoherence kicks in and scrambles the qubits. Typically, the groups of qubits assembled so far have decoherence times of a few microseconds. The number of logic operations you can carry out during that fleeting moment depends on how quickly the quantum gates can be switched — if this time is too slow, it really doesn't matter how many qubits you have at your disposal. The number of gate operations needed for a calculation is called its depth: Low-depth (shallow) algorithms are more feasible than high-depth ones, but the question is whether they can be used to perform useful calculations.

What's more, not all qubits are equally noisy. In theory it should be possible to make very low-noise qubits from so-called topological electronic states of certain materials, in which the “shape” of the electron states used for encoding binary information confers a kind of protection against random noise. Researchers at Microsoft, most prominently, are seeking such topological states in exotic quantum materials, but there's no guarantee that they'll be found or will be controllable.

Researchers at IBM have suggested that the power of a quantum computation on a given device be expressed as a number called the “quantum volume,” which bundles up all the relevant factors: number and connectivity of qubits, depth of algorithm, and other measures of the gate quality, such as noisiness. It's really this quantum volume that characterizes the power of a quantum computation, and Gambetta said that the best way forward right now is to develop quantum-computational hardware that increases the available quantum volume.

This is one reason why the much vaunted notion of quantum supremacy is more slippery than it seems. The image of a 50-qubit (or so) quantum computer outperforming a state-of-the-art supercomputer sounds alluring, but it leaves a lot of questions hanging. Outperforming for which problem? How do you know the quantum computer has got the right answer if you can't check it with a tried-and-tested classical device? And how can you be sure that the classical machine wouldn't do better if you could find the right algorithm?

So quantum supremacy is a concept to handle with care. Some researchers prefer now to talk about “quantum advantage,” which refers to the speedup that quantum devices offer without making definitive claims about what is best. An aversion to the word “supremacy” has also arisen because of the racial and political implications.

Whatever you choose to call it, a demonstration that quantum computers can do things beyond current classical means would be psychologically significant for the field. “Demonstrating an unambiguous quantum advantage will be an important milestone,” said Eisert — it would prove that quantum computers really can extend what is technologically possible.

That might still be more of a symbolic gesture than a transformation in useful computing resources. But such things may matter, because if quantum computing is going to succeed, it won’t be simply by the likes of IBM and Google suddenly offering their classy new machines for sale. Rather, it’ll happen through an interactive and perhaps messy collaboration between developers and users, and the skill set will evolve in the latter only if they have sufficient faith that the effort is worth it. This is why both IBM and Google are keen to make their devices available as soon as they’re ready. As well as a 16-qubit IBM Q experience offered to anyone who registers online, IBM now has a 20-qubit version for corporate clients, including JP Morgan Chase, Daimler, Honda, Samsung and the University of Oxford. Not only will that help clients discover what’s in it for them; it should create a quantum-literate community of programmers who will devise resources and solve problems beyond what any individual company could muster.

“For quantum computing to take traction and blossom, we must enable the world to use and to learn it,” said Gambetta. “This period is for the world of scientists and industry to focus on getting quantum-ready.”

This article was reprinted on [Wired.com](#)(opens a new tab).

Inside the Knotty World of ‘Anyon’ Particles

A new class of quantum particle is about to emerge from the tangled considerations of quantum statistics.

Introduction

Prior to the emergence of quantum mechanics, fundamental physics was marked by a peculiar dualism. On the one hand, we had electric and magnetic fields, governed by Maxwell’s equations. The fields filled all of space and were continuous. On the other hand, we had atoms, governed by Newtonian mechanics. The atoms were spatially limited — indeed, quite small — discrete objects. At the heart of this dualism was the contrast of light and substance, a theme that has fascinated not only scientists but artists and mystics for many centuries.

One of the glories of quantum theory is that it has replaced that dualistic view of matter with a unified one. We learned to make fields from photons, and atoms from electrons (together with other elementary particles). Both photons and electrons are described using the same mathematical structure. They are particles, in the sense that they come in discrete units with definite, reproducible properties. But the new quantum-mechanical sort of “particle” cannot be associated with a definite location in space. Instead, the possible results of measuring its position are given by a probability distribution. And that distribution is given as the square of a space-filling field, its so-called wave function.

Conceptually, quantum particles differ so significantly from their classical ancestors that a different name seems in order. Just as the quantum “qubit” was named by analogy to the classical “bit” of information, I will use the term “quarticle” (pronounced kwort-icle) for a quantum particle. This emphasis on the particle aspect (as opposed to “wavicle”) is appropriate, because in practice quantum physicists usually analyze quantum behavior by visualizing the behavior of particles, and then refining — and, if necessary, correcting — their picture until it works for quarticles.

The quantum unification of light and substance, while satisfying, is limited in scope. For when we go beyond single quarticles to consider the behavior of collections of identical quarticles, a new dualism appears. Indeed, the world of quantum particles divides into two great, mutually exclusive kingdoms. There is the kingdom of bosons, named after Satyendra Bose, and the kingdom of fermions, named after Enrico Fermi. Every species of quarticle is either a boson or fermion.

Interactions among bosons are very different from those of fermions. We call this effect “quantum statistics.” For purposes of orientation, a simple introduction may be in order.

Bosons are conformists. They like to behave in the same way. (More technically: Identical bosons have enhanced probability to occupy the same quantum state.) Photons belong to the kingdom of bosons. A laser beam is the epitome of boson-ness. It consists of many photons of the same wavelength (that is, color) moving in the same direction, the result of “stimulated emission” of photons in an imitative cascade.

Fermions, by contrast, are individualists. They absolutely refuse to occupy the same quantum state, a fact known as the Pauli exclusion principle. Electrons belong to the kingdom of fermions, and this is a key reason why the periodic table exists. Electrons, being negatively charged, are strongly attracted to positively charged atomic nuclei, but they prevent one another from surrounding the nucleus in a simple, efficient way. Instead they build up complex configurations that can support interesting chemistry.

Supersymmetry is a theoretical speculation that — if true — would reconcile the two kingdoms. According to supersymmetry, every elementary quarticle has a mate in the opposite kingdom, its superpartner. The superpartner of a boson is a fermion, and vice versa. Superpartners share the same electric charge and several other properties, but differ in mass and spin.

Supersymmetry is an attractive, logical extension of known physics, and it can be implemented with elegant mathematics. Many physicists, including me, feel that it deserves to be true.

But the last word, naturally, goes to nature. While there is compelling circumstantial evidence for supersymmetry, as yet there is no direct proof. For that, we need to find some superpartners. Searching for superpartners of

known particles is a major preoccupation of experimentalists working at the Large Hadron Collider. Sadly, [the results so far are negative](#). Yet there is still considerable potential for discovery, as the machine comes to operate at higher energy and more collisions get analyzed.

Clearly, quantum statistics lies at the foundation of our understanding of nature. Also, as we've seen, it raises a profound question about the unity of matter. Addressing that question suggests new possibilities for discovery.

Such an important concept deserves a worthy grounding. What is quantum statistics, at bottom?

The modern answer to that question is deep, beautiful and surprisingly recent. It emerged in the late 1970s, more than 50 years after quantum mechanics matured, and it was catalyzed by the pioneering work of Jon Leinaas and Jan Myrheim.

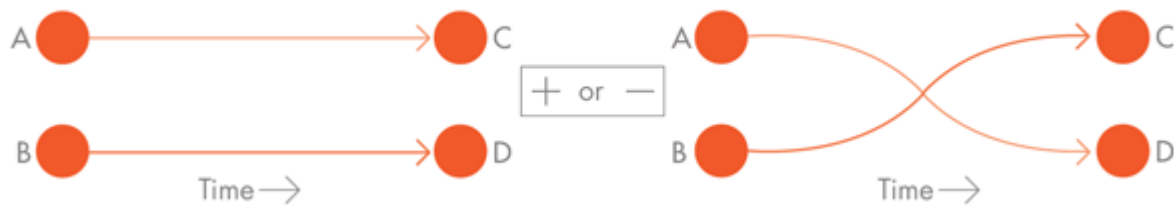
In brief, quantum statistics reflects the topology of quartile world lines. In the next three paragraphs, I'll unpack that dense bit of poetry.

In quantum theory, the probability for a process is expressed as the square of the amplitude of its wave function. Amplitudes are more basic and obey simpler rules than the probabilities they control. The central problem of quantum dynamics, therefore, is to calculate the wave-function amplitude for a given configuration of particles at an initial time to evolve into another configuration at a later time.

In calculating the total amplitude for two indistinguishable quartiles that begin at positions A and B and end at positions C and D, we must take into account contributions from every possible motion connecting the starting positions to the end points. These paths are the "world lines" of the quartiles. The motions divide into two classes: one class in which the two quartiles, which began at A and B, respectively, wind up at C and D, and another class in which they wind up at D and C.

QUANTUM EXCHANGE

What's the chance that two indistinguishable particles that start at points A and B will end at points C and D? According to the rules of quantum theory, you must account for all possible ways of getting to the same final result. Here, the options fall into two distinct classes. In one case, the particle that begins at A ends at C, while in the other it ends at D. How might you combine the contributions of these two classes? By adding them or subtracting them. Adding them gives us bosons, while subtracting them gives us fermions.



Lucy Reading-Ikkanda/Quanta Magazine

Because the particles are indistinguishable, the final result — that is, two particles at positions C and D — is the same in both cases. Yet each world line pair has a distinct topology.

We need a rule for how to combine the contributions from those two classes. There are two mathematically consistent possibilities. We can add them, or we can subtract them. The “add” option gives us bosons, while the “subtract” option gives us fermions. All the characteristic properties of bosons and fermions can be deduced from those basic rules.

It is a wonderful intellectual achievement, to have traced such profoundly consequential properties of matter to a basic aspect of identity — that is, indistinguishability — and the topology of motion in space-time.

But the real payoff comes when we bring that basic understanding to bear on new situations and make new discoveries. The most dramatic application of our deeper understanding of quantum statistics arises when we consider particles that live in two dimensions, as I’ll now explain.

Perhaps surprisingly, in a world of two dimensions the topology of pairs (or larger groups) of world lines becomes much richer than in worlds of three or more. The reason is closely connected to a basic feature of knots.

In three space dimensions, knot theory is a subtle, complicated subject. But in four space dimensions it is trivial: All knots can be unraveled completely.

A knot, to mathematicians, is just a continuous curve in space. Imagine labeling each point on that tangled curve with the numbers between zero and one, with zero and one representing the same point. Now take a circle and do the same.

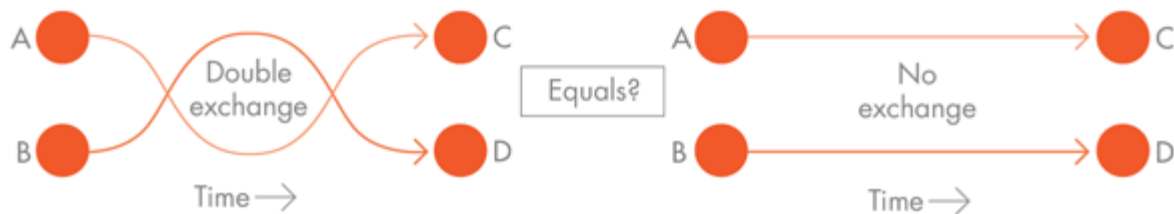
To untangle the knot, simply let the points on our knot flow to the points on the circle that are labeled with the corresponding number. Of course, obstructions to this unraveling might arise, such as when different parts of the curve come to intersect. But in four dimensions, we can always move two strands past one another.

The procedure is difficult to visualize directly, but we can understand it using a simple trick. Let us represent position in the extra dimension by the value of a fictitious temperature. If our threatened intersection occurs at a point where the temperatures of the strands are different, it is not a genuine intersection, because the strands are located differently in the extra dimension. And if the temperatures are the same, we need only make a temporary excursion into the extra dimension — heat one strand up, move it “through” (in the ordinary dimensions), and then cool it back down — to avoid the intersection.

The world lines of quarklets that move in three-dimensional space form strands in four-dimensional space-time. Therefore, as we’ve just seen, they can’t really get tangled up. Their topology reduces to keeping track of exchanges, as we discussed earlier.

TANGLED ANYONS

Here, one particle exchange is followed by another. In three space dimensions (and one time dimension), this kind of “double exchange” trajectory pair can be untangled into a “no exchange” pair. This imposes a consistency condition on the factor associated with an exchange: When we apply it twice, we must get back to where we started. Because of this, the only consistent exchange factors are 1 and -1 , corresponding to bosons and fermions. In two space dimensions (plus one time dimension), however, the winding of one path trajectory around another cannot be undone by any continuous motion. The consistency condition loosens, and “anyons” arise as a possibility.



Lucy Reading-Ikkanda/Quanta Magazine

On the other hand, world lines of quarticles whose motion is confined to a two-dimensional space form strands in a three-dimensional space-time. Strands in three dimensions most definitely can get tangled up, as anyone who has braided hair knows.

With that richer topology comes much greater variety in the possibilities for adding up the contributions of world lines. In other words, there are many more possibilities for quantum statistics, and thus more categories of particles beyond bosons and fermions. I coined the word “anyon” to describe quarticles whose motion is restricted to two space dimensions, and which are neither bosons nor fermions. I meant this humorously, to suggest “anything goes,” but of course that implication should not be taken literally. Anyons must obey highly structured mathematical rules. Still, we find infinitely many consistent possibilities for them, instead of only two.

Because the rules for anyon behavior are sensitive to the quarticles’ motion over time, anyons have memory. More precisely: Since the amplitude for an evolution involving anyons will be different, depending on how their world lines got tangled up over time, the value of the amplitude provides a record of their relative motion. As we’ll discuss momentarily, the memory abilities of anyons might power an important technology.

Do anyons exist, outside of theorists' imaginations?

Our immediate reaction might be: "Not in our world, since space is (at least) three-dimensional." But that conclusion fails to reckon with nature's creativity, and ours.

Two-dimensional universes, in fact, are all around us. We call them surfaces, films, membranes or interfaces. They can be accessed from "our" enveloping three-dimensional world, but they also have a life of their own. In more prosaic terms, some of those two-dimensional objects trap energy efficiently, and if they are not disturbed, their behavior will be self-contained.

The properties of these universes depend, of course, on what they're made from, and on the pattern of their construction. Recently, there has been much interest in two-dimensional materials whose electrons are "highly entangled." The precise, technical definition of that phrase involves [deep concepts in the mathematics of quantum theory](#). Here a rough analogy will have to serve.

When electrons form covalent bonds, they cannot move independently; they are, we may say, frozen in place, forming a kind of solid. Quantum fluctuations may cause that solid to sublime, producing materials in which the electrons move freely — a kind of electron gas. This is a crude but extremely useful way to describe the electrons in metals. But we can also imagine intermediate cases, where the electrons are neither locked in place nor independent, forming a sort of liquid.

These strange liquids, which melted due to quantum fluctuations rather than thermal agitation, are especially interesting. In this state, the electrons remain orderly, yet they are not rigidly stifled. Their order involves subtle quantum correlations — that is, entanglement.

Imperfections in these liquids' orderly arrangement behave much like impurities in ordinary liquids. They form stable arrangements that retain their basic structure, even as they move. Thus, they are a kind of emergent particle — or rather, since we're concerned with quantum liquids, a kind of emergent quarticle. These emergent quarticles are several steps removed from individual electrons, and they can have quite different properties from those of electrons.

“Everything not forbidden is compulsory.” That memorable phrase, from T.H. White’s novel *The Once and Future King*, expresses an important principle of quantum behavior: The spontaneous activity of quantum systems explores all consistent possibilities. For me, it has also been an inspiring principle for theoretical exploration: Nature, in her abundance, provides materials to embody all theoretically consistent possibilities. Trusting in that principle, I strive to exercise what Richard Feynman called “imagination in a straitjacket,” expecting that

If you build it (imaginatively), they will come (physically).

So it has proved for anyons. Physicists now know of several two-dimensional quantum liquids whose emergent quarticles are neither bosons nor fermions, but rather different sorts of anyons.

The first example turned up in a fascinating domain of physics known as the fractional quantum Hall effect (FQHE). This is a state of matter that arises when electrons confined to a two-dimensional layer (typically, at the interface of two semiconductors) are taken to extremely low temperatures and subjected to extremely large magnetic fields. Under these conditions, as the density and magnetic field are varied, the electrons form a large class of related but distinct quantum liquids with remarkable properties.

Specifically, the emergent quarticles in those liquids typically carry electric charge that is a fraction of an electron’s charge, and they obey quantum statistics that is a fraction of fermion statistics. In the most robust FQHE state, for example, quarticles carry one-third of an electron’s electric charge. In general, the quarticles behave like one-third of an electron; indeed, an injected electron will fission into three of these quarticles. Consistent with that picture, the quarticles are predicted to be anyons, which exhibit one-third of fermion statistics. In mathematical terms, the amplitude for exchange of our quarticles will be multiplied by the cube root of -1 .

The theory of the FQHE is highly developed and makes many successful predictions about the observed quantum liquids. Since anyonic quarticles are a central feature of that theory, there is little doubt among theorists that anyons exist in nature and are lurking in observed FQHE liquids. Numerical studies support that belief as well. Unfortunately, however, direct experimental proof has been difficult to come by, for a variety of practical reasons.

Experimentalists continue to work on the problem, and they're tantalizingly close to success.

Very recently a new class of quantum liquids, called spin liquids, has been identified. In these materials the electrons don't move, but their spins — that is, the orientations of their spin axes — are in play. Spin liquids lie between ordinary magnets (spin solids), in which the directions of spins are rigidly aligned, and paramagnets and diamagnets (spin gases), in which the spin orientations are almost completely independent of one another. Theoretical models of two-dimensional spin liquids show that many of them support anyonic quasiparticles. The practical problems that plague FQHE studies of anyon behavior are less severe in spin liquids, so we can anticipate rapid progress.

Last but certainly not least, there has been considerable ferment in the engineering of artificial anyon systems. Recall that a basic feature of anyons, which distinguishes them from more familiar quasiparticles, is their memory. Many-anyon systems build up a gigantic collective memory, which can serve as a platform for computing. This so-called “[topological quantum computing](#)” is one plausible path toward building useful quantum computers. Microsoft is making major investments in this kind of research.

Topological quantum computing deserves extensive discussion, and I plan to return to it in a future column. Let me conclude this one with two pictures that point to the past and the future of that endeavor.

Computing with anyons exploits their ability to map their knotted histories into (observable) quantum-mechanical amplitudes. We move anyons around in clever ways and then access the tangled history of their motion. Topological quantum computing is, therefore, a form of computing with knots. As such, it is a modernization of quipu, the Incan technology for computation and encryption.

If topological computing does eventually lead to powerful quantum computers, then the most capable artificial intelligence will live in two-dimensional materials, embodied in circulating systems of anyons. We can look forward to a day when those super-brains reach out to thank their progenitors.

Quantum Computing Without Qubits

A quantum computing pioneer explains why analog simulators may beat out general-purpose digital quantum machines — for now.

Ivan Deutsch is building quantum computers out of base-16 “qudits,” quantum information units that can assume any number of “d” states.

Courtesy of Ivan Deutsch

For more than 20 years, [Ivan H. Deutsch](#) ([opens a new tab](#)) has struggled to design the guts of a working quantum computer. He has not been alone. The quest to harness the computational might of quantum weirdness continues to occupy hundreds of researchers around the world. Why hasn’t there been more to show for their work? As physicists have known since quantum computing’s beginnings, the same characteristics that make quantum computing exponentially powerful also make it devilishly difficult to control. The quantum computing “nightmare” has always been that a quantum computer’s advantages in speed would be wiped out by the machine’s complexity.

Yet progress is arriving on two main fronts. First, researchers are developing unique quantum error-correction techniques that will help keep quantum processors up and running for the time needed to complete a calculation. Second, physicists are working with so-called analog quantum simulators — machines that can’t act like a general-purpose computer, but rather are designed to explore specific problems in quantum physics. A classical computer would have to run for thousands of years to compute the quantum equations of motion for just 100 atoms. A quantum simulator could do it in less than a second.

Quanta Magazine spoke with Deutsch about recent progress in the field, his hopes for the near future, and his own work at the University of New Mexico’s [Center for Quantum Information and Control](#) ([opens a new tab](#)) on scaling up binary quantum bits into base-16 digits.

QUANTA MAGAZINE: Why would a universal quantum machine be so uniquely powerful?

IVAN DEUTSCH: In a classical computer, information is stored in retrievable bits binary coded as 0 or 1. But in a quantum computer, elementary particles inhabit a probabilistic limbo called superposition where a “qubit” can be coded as 0 [and](#) 1.

Here is the magic: Each qubit can be entangled with the other qubits in the machine. The intertwining of quantum “states” exponentially increases the number of 0s and 1s that can be simultaneously processed by an array of qubits. Machines that can harness the power of quantum logic can deal with exponentially greater levels of complexity than the most powerful classical

computer. Problems that would take a state-of-the-art classical computer the age of our universe to solve, can, in theory, be solved by a universal quantum computer in hours.

What is the quantum computing “nightmare”?

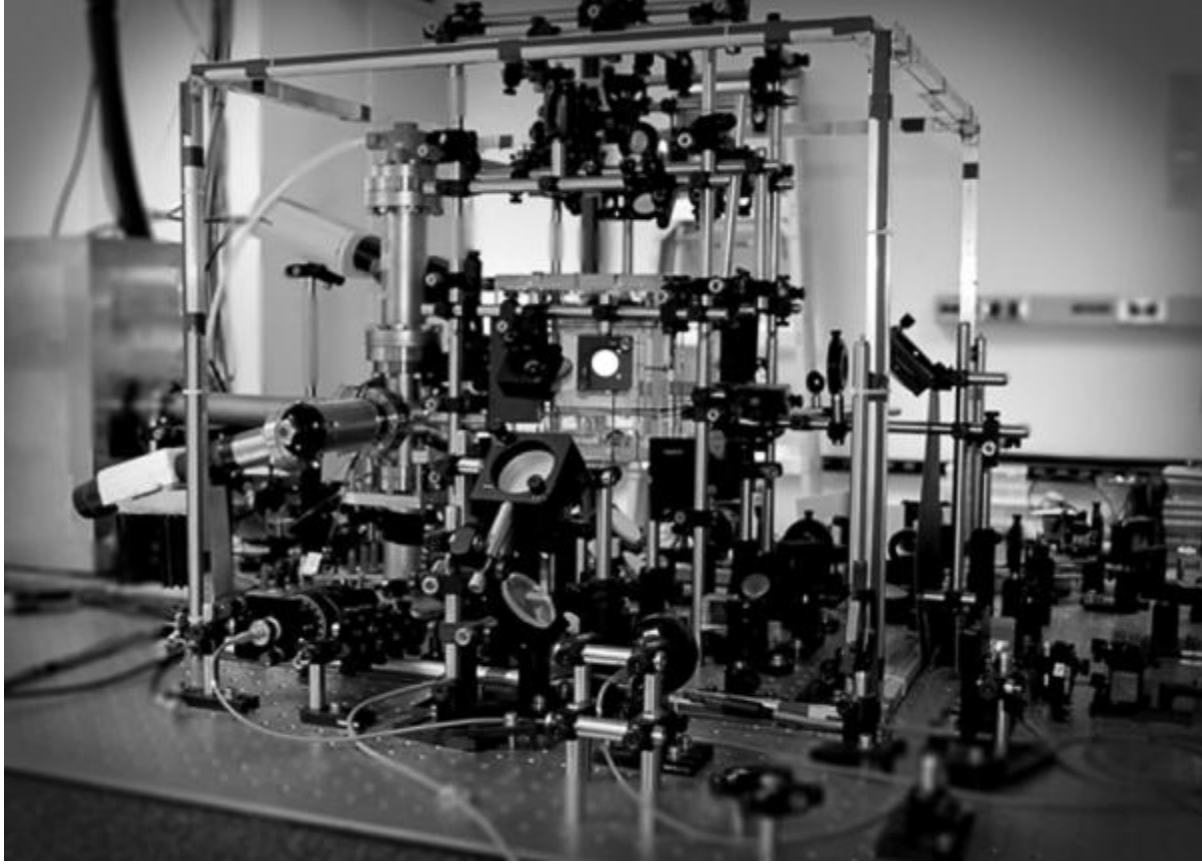
The same quantum effects that make a quantum computer so blazingly fast also make it incredibly difficult to operate. From the beginning, it has not been clear whether the exponential speed up provided by a quantum computer would be cancelled out by the exponential complexity needed to protect the system from crashing.

Is the situation hopeless?

Not at all. We now know that a universal quantum computer will not require exponential complexity in design. But it is still very hard.

So what’s the problem, and how do we get around it?

The hardware problem is that the superposition is so fragile that the random interaction of a single qubit with the molecules composing its immediate surroundings can cause the entire network of entangled qubits to delink or collapse. The ongoing calculation is destroyed as each qubit transforms into a digitized classical bit holding a single value: 0 or 1.



A test-bed quantum computer that Deutsch is working on with his colleague Poul Jessen at the University of Arizona.

Courtesy of Poul Jessen

In classical computers, we reduce the inevitable loss of information by designing a lot of redundancy into the system. Error-correcting algorithms compare multiple copies of the output. They select the most frequent answer and discard the rest of the data as noise. We cannot do that with a quantum computer, because trying to directly compare qubits will crash the program. But we are gradually learning how to keep systems of entangled qubits from collapsing.

The major obstacle, to my mind, is creating error-correcting software that can keep data from being corrupted as the calculation proceeds toward the final readout. The great trick is to design and implement an algorithm that only measures the errors and not the data, thus preserving the superposition that contains the correct answer.

Will that end the nightmare?

It turns out that the error correction technique itself introduces errors. One of the most wonderful advances in quantum computing was recognizing that, in theory, we can correct the new errors without requiring 100 percent precision, allowing minor background noise to pollute the

calculation as it rolls along. We cannot actually do this — yet. The main reason that we do not have a working universal quantum computer is that we are still experimenting with how to implant such a “fault-tolerant” algorithm into a quantum circuit. Right now we can control 10 qubits reasonably well. But there is no error-correcting technique, to my knowledge, capable of controlling the thousands of qubits needed to construct a universal machine.

Is that what you’re working on?

I study the information processing capabilities of trapped atoms. My colleague [Poul Jessen](#) (opens a new tab) at the University of Arizona and I are pushing the logical power beyond binary-based qubits. For example, what if we can control the superposition of an atom with, say, 16 different energy levels? Using base 16, we can then store what we call a “qudit” in a single atom. That would move us beyond the information processing speed obtainable by a base 2 system, the qubit.

What other options do we have?

There may be significant applications available for making non-universal machines: Special purpose, analog quantum simulators designed to solve specific problems, such as how room-temperature superconductors work or how a particular protein folds.

Are these actually computers?

They are not universal machines capable of solving any type of question. But say that I want to model global climate change. One way to do this is to write down a mathematical model and then solve the equations on a digital computer. That is typically what climate scientists do. Another way is to try to simulate some aspect of the earth’s climate in a controllable experiment. I can create a simple physical system that obeys the same laws of motion as the system I’m trying to model — mixing nitrogen, oxygen, and hydrogen in a tank, for example. What goes on inside the tank is a real-world computation that tells me something about atmospheric turbulence under certain conditions.

It is the same with an analog quantum simulator — I use one controllable physical system to simulate another. For example, successfully simulating a superconductor with such a device would reveal the quantum mechanics of high-temperature superconductivity. That could lead to the manufacture of non-brittle superconducting materials for many uses, including building less-fragile quantum circuits. Hopefully, we can learn how to build a robust universal digital computer by experimenting with analog simulators.

Has anyone built a working analog quantum simulator?

In 2002, a group at the Max Planck Institute in Germany built an optical lattice — a super-chilled egg carton made of light — and controlled it by pulsing different strengths of laser beams at it. This was a fundamentally analog device designed to obey quantum mechanical equations of motion. The short story is that it successfully simulated how atoms transition between acting as superfluids or insulators. That experiment has sparked a lot of research in analog quantum computing with optical lattices and cold atom traps.

What are the main challenges for these quantum simulators?

Because the evolution of the analog simulation is not digitized, the software cannot correct the tiny errors that accumulate during the calculation as we could error-correct noise on a universal machine. The analog device must keep a quantum superposition intact long enough for the simulation to run its course without resorting to digital error correction. This is a particular challenge for the analog approach to quantum simulation.

Is the D-Wave machine a quantum simulator?

The D-Wave prototype is not a universal quantum computer. It is not digital, nor error-correcting, nor fault tolerant. It is a purely analog machine designed to solve a particular optimization problem. It is unclear if it qualifies as a quantum device.

Will a scalable quantum computer be deployed during your lifetime?

We are pushing past the nightmare. Around the world, many university-based labs are working hard to remove or bypass the road block of fault tolerance. Academic researchers are leading the way, intellectually. For example, the groups of [Rob Schoelkopf](#)[\(opens a new tab\)](#) and [Michel H. Devoret](#)[\(opens a new tab\)](#) at Yale are taking superconducting technologies close to fault-tolerance.

But constructing a working universal digital quantum computer will likely require mobilizing industrial-scale resources. To that end, IBM is exploring quantum computing with superconducting circuits with personnel largely from the Yale groups. Google is working with [John Martinis's lab](#)[\(opens a new tab\)](#) at the University of California, Santa Barbara. HRL Laboratories is working on silicon-based quantum computing. Lockheed Martin is exploring ion traps. And who knows what the National Security Agency is up to.

But generally in academic labs, without these industrial-scale resources, scientists are focusing more and more on learning how to control analog quantum simulators. There is short-term fruit to be picked in that arena — both intellectually and in the currency of academics: publishable papers.

Are you willing to settle for analog?

I favor pursuing the digital approach full force. Before I die, I would love to see just one universal logical qubit that can be indefinitely error corrected. It would instantly be classified by the government, of course. But I dream on, regardless.

CRYPTOGRAPHY

A Tricky Path to Quantum-Safe Encryption

In the drive to safeguard data from future quantum computers, cryptographers have stumbled upon a thin red line between security and efficiency.

On August 11, the National Security Agency updated an obscure [page on its website](#)[\(opens a new tab\)](#) with an announcement that it plans to shift the encryption of government and military data away from current cryptographic schemes to new ones, yet to be determined, that can resist an attack by quantum computers.

“It is now clear that current Internet security measures and the cryptography behind them will not withstand the new computational capabilities that quantum computers will bring,” NSA spokesperson Vanee’ Vines stated in an email, confirming the change. “NSA’s mission to protect critical national security systems requires the agency to anticipate such developments.”

Quantum computers, once seen as a [remote theoretical possibility](#), are now widely expected to work within five to 30 years. By exploiting the probabilistic rules of quantum physics, the devices could decrypt most of the world’s “secure” data, from NSA secrets to bank records to email passwords. Aware of this looming threat, cryptographers have been racing to develop [“quantum-resistant” schemes](#)[\(opens a new tab\)](#) efficient enough for widespread use.

The most promising schemes are believed to be those based on the mathematics of lattices — multidimensional, repeating grids of points. These schemes depend on how hard it is to find information that is hidden in a lattice with hundreds of spatial dimensions, unless you know the secret route.

But last October, cryptographers at the Government Communications Headquarters (GCHQ), Britain’s electronic surveillance agency, posted

an [enigmatic paper](#)[\(opens a new tab\)](#) online that called into question the security of some of the most efficient lattice-based schemes. The findings hinted that vulnerabilities had crept in during a decade-long push for ever-greater efficiency. As cryptographers simplified the underlying lattices on which their schemes were based, they rendered the schemes more susceptible to attack.

Building on the GCHQ claims, two teams of cryptanalysts have spent the past year determining which lattice-based schemes can be broken by quantum computers, and which are safe — for now.

“This is the modern incarnation of the classic cat-and-mouse game between the cryptographer and cryptanalyst,” said [Ronald Cramer](#)[\(opens a new tab\)](#) of the National Research Institute for Mathematics and Computer Science (CWI) and Leiden University in the Netherlands. When cryptanalysts are quiet, cryptographers loosen the security foundations of the schemes to make them more efficient, he said. “But at some point a red line might be crossed. That’s what happened here.” Now, the cryptanalysts are speaking up.

Open Secrets

Every time you visit a website with a URL that begins “HTTPS,” you send or receive encrypted data. Secure Internet transactions are made possible by public-key cryptography, a revolutionary invention of the 1970s. Up until then, cryptography had mostly been a game for governments and spies; two parties, such as a spy and a handler, had to agree in advance on a secret cipher or “key” in order to communicate in secret. (The simple “Caesar cipher,” for example, shifts the letters of the alphabet by some agreed-upon number of positions.) Public-key cryptography makes it possible for anyone to send anyone else an encrypted message that only the recipient can decrypt, even if the parties involved never agreed on anything and no matter who is listening in.

“The reception from NSA was apoplectic,” [Martin Hellman](#)[\(opens a new tab\)](#), one of the three Stanford University researchers who invented public-key cryptography, [recalled](#)[\(opens a new tab\)](#) in 2004.

TYPES OF CRYPTOGRAPHY

QUANTUM-BREAKABLE



RSA encryption

A message is encrypted using the intended recipient's public key, which the recipient then decrypts with a private key. The difficulty of computing the private key from the public key is connected to the hardness of prime factorization.



Diffie-Hellman key exchange

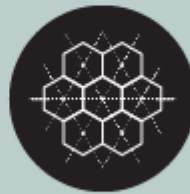
Two parties jointly establish a shared secret key over an insecure channel that they can then use for encrypted communication. The security of the secret key relies on the hardness of the discrete logarithm problem.



Elliptic curve cryptography

Mathematical properties of elliptic curves are used to generate public and private keys. The difficulty of recovering the private key from the public key is related to the hardness of the elliptic-curve discrete logarithm problem.

QUANTUM-SECURE



Lattice-based cryptography

Security is related to the difficulty of finding the nearest point in a lattice with hundreds of spatial dimensions (where the lattice point is associated with the private key), given an arbitrary location in space (associated with the public key).



Code-based cryptography

The private key is associated with an error-correcting code and the public key with a scrambled and erroneous version of the code. Security is based on the hardness of decoding a general linear code.



Multivariate cryptography

These schemes rely on the hardness of solving systems of multivariate polynomial equations.

All three of the most widely used cryptographic schemes can be broken by algorithms designed to run on future quantum computers (left column). Cryptographers have devised a variety of schemes, three of which appear on the right, that are thought to be quantum-secure.

Olena Shmahalo/Quanta Magazine

In public-key cryptography, data is secured by math problems that are easy to solve, but hard to reverse engineer. For example, while it is easy for a computer to multiply two prime numbers to produce a larger integer, as in the calculation $34,141 \times 81,749 = 2,790,992,609$, it is hard — that is, it takes an impractically long time on a computer — to factorize a large enough integer into its component primes. In a crypto scheme based on prime factorization, the primes serve as a person's "private key," which is not shared. The product of the primes serves as the "public key," which is distributed publicly. When someone else uses the public key to encrypt a message, only the person in possession of the private key can decrypt it.

Two efficient public-key encryption schemes that emerged in the late 1970s remain the most widely used today: RSA (invented by [Ron Rivest](#)[\(opens a new tab\)](#), [Adi Shamir](#)[\(opens a new tab\)](#) and [Leonard Adleman](#)[\(opens a new tab\)](#)), based on the prime factoring problem, and the Diffie-Hellman key exchange (invented by [Whit Diffie](#)[\(opens a new tab\)](#) and Hellman), based on what's called the [discrete logarithm problem](#)[\(opens a new tab\)](#). Although there was no actual proof that either prime factors or discrete logarithms were impossible to compute in a reasonable time frame, no one could find algorithms for efficiently computing them.

"Over time, people build up confidence in the hardness of some problem because so many people have tried to think about how to break it and cannot," said [Jill Pipher](#)[\(opens a new tab\)](#), a mathematician and cryptographer at Brown University.

With existing algorithms, it takes years to compute the prime factors associated with a public key of typical length. And so, RSA and the Diffie-Hellman key exchange became the armor of the Internet, and a sense of security reigned.

That security, it turns out, came with an expiration date.

Shor's Algorithm

Assumptions about which math problems are hard for computers to solve shattered in 1994, when an AT&T researcher named [Peter Shor](#)(opens a new tab) revealed the theoretical decrypting power of future quantum computers.

In an ordinary computer, information is stored in units called bits that can exist in either of two states, designated 0 or 1. The computer's computational capacity is proportional to the number of bits. In quantum computers, however, the information-storing units, called qubits, can exist in both the 0 and 1 states simultaneously. (Qubits might take the form of subatomic particles spinning both clockwise and counterclockwise at the same time, for example.) Because a system of many qubits can exist in all possible combinations of all their possible individual states, the computational capacity of a quantum computer would increase exponentially with the number of qubits.

This would seem to make quantum computers more powerful problem solvers than classical computers. However, actually tapping their potential requires finding an algorithm for juggling their simultaneous realities, so that in the end, the right one — that is, the state of the system corresponding to the correct answer — emerges. For more than a decade after quantum computing was conceived in the early 1980s, no promising algorithms emerged, and the field languished. “Frankly, nobody paid any attention,” said [Seth Lloyd](#)(opens a new tab), a quantum computing theorist at the Massachusetts Institute of Technology.

All that changed in 1994 when Shor, who is now at MIT, [devised a quantum computer algorithm](#)(opens a new tab) capable of efficiently computing both prime factors and discrete logarithms, and thus of breaking both RSA encryption and the Diffie-Hellman key exchange. “At that point there was a killer app for quantum computing — maybe you could call it a quapp — and the interest in quantum computing boomed,” Lloyd said.

With the superior computational capabilities of quantum computers revealed by Shor's algorithm, researchers worldwide have been racing to build them ever since. In parallel, cryptographers have raced to come up with new schemes that quantum computers can't crack. “We didn't know where to look for a long time,” said [Chris Peikert](#)(opens a new tab), a cryptographer at the Georgia Institute of Technology in Atlanta. “But lattices seem to be a very good foundation.”

Lost in Lattices

Just as the security of RSA encryption is based on the idea that it's easy to multiply primes but hard to compute prime factors, the security of lattice-based crypto schemes rests on how easy it is to get lost in a 500-dimensional lattice: You simply start at a lattice point and jiggle the spatial coordinates, ending up at some location nearby. But it's exceedingly hard to find the nearest lattice point, given an arbitrary location in 500-dimensional space. In lattice-based schemes, the private key is associated with the lattice point, and the public key is associated with the arbitrary location in space.

Despite its promise, lattice-based cryptography got off to a slow start. In the 1980s, public keys based on lattices were too long, requiring megabytes of data to transmit. Cryptographers were forced to simplify the underlying lattices for the sake of efficiency. In a generic lattice, lattice points are generated by taking all possible linear combinations of some set of vectors (arrows pointing in different directions). Assigning a pattern to these vectors makes the resulting lattice simpler, and the associated keys shorter. Invariably, however, simplifying the lattice also makes it easier to navigate, allowing private keys to be deduced from public keys, thereby breaking the scheme. "Lattices became synonymous with disaster — with failed attempts at crypto," said [Jeff Hoffstein](#)(opens a new tab), a mathematician at Brown.

While the rest of the world moved on, some cryptographers continued to tinker with lattices. In 1995, Hoffstein, with Pipher and another Brown colleague, [Joe Silverman](#)(opens a new tab), devised a cryptographic scheme based on "cyclic" lattices, which are generated by vectors that can rotate in any direction and still land on another lattice point. NTRU, as they called the scheme, was extremely efficient — even more so than the RSA and Diffie-Hellman protocols. Although there was no proof that the cyclic lattices underlying NTRU were hard for computers to navigate, or that NTRU was secure, 20 years have passed and no one has found a way to break it, boosting confidence in its security.

The promise of lattices grew dramatically in 1997, when the IBM researchers Miklós Ajtai and [Cynthia Dwork](#)(opens a new tab) devised the first lattice-based crypto scheme that was provably as hard to break as the underlying lattice problem is hard to solve. Building on this work, [Oded Regev](#)(opens a new tab), a theoretical computer scientist now at New York University's Courant Institute of Mathematical Sciences, proved in 2005 that crypto schemes based on a problem called learning with errors (LWE) are secure against quantum computers, as long as the problem of finding the nearest point in a generic lattice is hard for quantum computers (as most researchers

presume). LWE was inefficient, but Regev, Peikert and [Vadim Lyubashevsky](#)[\(opens a new tab\)](#), who is now at IBM Research in Switzerland, soon developed analogous schemes based on “ideal” lattices (which are closely related to cyclic lattices), and showed that these more efficient schemes, dubbed Ring-LWE, are secure as long as the underlying, ideal lattice problem is hard.

Learning With Errors

In 2005, Oded Regev devised a cryptographic scheme based on a problem called “learning with errors,” which he proved is as secure as the lattice problem is hard. LWE-based schemes work something like this:

Pick any odd number, and don’t tell anyone what it is. That’s your private key. Now multiply it by any other number, and add a small even number to it. (For example, if your original number was 121, you might multiply it by five and then add two, yielding 607. In practice, the numbers are much larger.) Do this many times, producing a list of enlarged, perturbed versions of your private key. This list of numbers is your public key. Tell the world.

Now, say someone wants to send you a message (e.g., 0 or 1, attack or retreat, yes or no). First this person randomly selects half of the numbers listed in your public key and adds them together. Then, to send the message “0,” your correspondent simply sends the sum back to you. To send the message “1,” the person adds one to the sum, then sends this back to you. Now, to decode the message, you simply divide the sum you’ve received by your private key. If the remainder is even, the message is “0.” If the remainder is odd, the message is “1.”

Once again, however, there seemed to be a tradeoff between security and efficiency, and an irksome impossibility of having both. Ring-LWE had better security assurances than NTRU and was [far more versatile](#), but it was not as efficient. Some researchers believed they could do better. Since 2007, they have been considering cryptographic schemes based on “principal ideal lattices,” which are generated by a single vector, in much the same way that the set of integers $\{ \dots, -6, -3, 0, 3, 6, 9, \dots \}$ can be generated by multiples of the integer 3.

“They were greedy; they were not happy with existing efficiency,” Regev said.

Cat and Mouse

As academic cryptographers devised crypto schemes based on principal ideal lattices, so did people behind the scenes at GCHQ. Their secret scheme, called Soliloquy, employed techniques from [number theory](#) to reduce the public-key size from a matrix of large numbers down to a single prime number. In the underlying lattice problem, this is equivalent to generating a lattice with a single, very short vector. “Unfortunately, the constructions used to do this were its Achilles heel,” a GCHQ spokesperson said in an email.

In their paper published last October, titled “Soliloquy: A Cautionary Tale,” the GCHQ researchers revealed that they had invented Soliloquy and then abandoned work on it in 2013 upon discovering a quantum attack that could break it. The paper provided only a vague sketch of the attack, however, leaving open the question of how it worked and which other lattice-based schemes might be affected. It seemed that in the pursuit of efficiency, a red line had been crossed. But where was the line?

“There was this initial idea that this attack could possibly be broader, and perhaps implicated all of lattice-based cryptography,” Pipher said. Others were skeptical that the attack worked at all.

Cryptographers have spent almost a year determining the scope of the Soliloquy attack. “People became obsessed,” Hoffstein said. “There was a frenzy.” It turned out that the GCHQ team had not worked out many details themselves, but merely had “sufficiently strong evidence that an attack could be developed and hence that Soliloquy could not be recommended for real-world use,” as the spokesperson put it in an email. In a March paper, Regev, Peikert, Cramer and [Léo Ducas](#)(opens a new tab) of CWI [worked out](#) the part of the attack that required only an ordinary computer; last week, [Jean-François Biasse](#)(opens a new tab) and [Fang Song](#)(opens a new tab) of the University of Waterloo in Ontario [laid out the quantum steps](#)(opens a new tab).

Besides Soliloquy, the findings indicated that other schemes based on principal ideal lattices generated by a single short vector are also broken, whereas schemes based on more generic ideal lattices, such as Ring-LWE and NTRU, are not affected. “There seem to be some initial technical obstacles in transferring these techniques to other important schemes,” Cramer said, adding that it warrants further study.

On the security-efficiency continuum, cryptographers slid too far to the efficiency side. In their scramble to find the best quantum-resistant schemes for banks, governments and the rest of the secure Internet, the Soliloquy attack has forced them back, toward schemes that are somewhat less efficient, but more firmly based on hard lattice problems. That is, presumed hard lattice problems.

There's no proof that quantum computers cannot find their way around lattices. "It could be that all these problems are actually easy," Peikert said. "But it seems unlikely, given what we know."

As for why extreme efficiency and perfect security appear to be so diametrically opposed, Hoffstein said: "The universe is an irritating place, and this is just another example of it."

Correction: This article was revised on Sept. 8, 2015, to clarify that the schemes broken by the Soliloquy attack are those based on principal ideal lattices generated by a single, short vector, and not those based on principal ideal lattices generated by more than one short vector.

How to Build a Robot That Wants to Change the World

The computer scientist Christoph Salge is trying to circumvent the need for rules that guide robots' behavior. His strategy: Give them a goal of making us more powerful.

Christoph Salge at New York University's Game Innovation Lab.

Introduction

Isaac Asimov's famous Three Laws of Robotics — constraints on the behavior of androids and automatons meant to ensure the safety of humans — were also famously incomplete. The laws, which first appeared in his 1942 short story “Runaround” and again in classic works like *I, Robot*, sound airtight at first:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Of course, hidden conflicts and loopholes abound (which was Asimov's point). In our current age of [advanced machine-learning software](#) and autonomous robotics, defining and implementing an airtight set of ethics for artificial intelligence has become a pressing concern for organizations like the [Machine Intelligence Research Institute\(\)](#) and [OpenAI\(\)](#).

[Christoph Salge\(\)](#), a computer scientist currently at New York University, is taking a different approach. Instead of pursuing top-down philosophical definitions of how artificial agents should or shouldn't behave, Salge and his colleague Daniel Polani are investigating a bottom-up path, or “what a robot should do in the first place,” as they write in their recent paper, [“Empowerment as Replacement for the Three Laws of Robotics.”\(\)](#) Empowerment, a concept inspired in part by cybernetics and psychology, describes an agent's intrinsic motivation to both persist within and operate upon its environment. “Like an organism, it wants to survive. It wants to be able to affect the world,” Salge explained. A Roomba programmed to seek its charging station when its batteries are getting low could be said to have an extremely rudimentary form of empowerment: To continue acting on the world, it must take action to preserve its own survival by maintaining a charge.

Having robots treat us like our dogs treat us is probably a future we can all live with.

Empowerment might sound like a recipe for producing the very outcome that safe-AI thinkers like [Nick Bostrom](#) fear(): powerful autonomous systems concerned only with maximizing their own interests and running amok as a result. But Salge, who has studied human-machine social interactions, wondered what might happen if an empowered agent “also looked out for the empowerment of another. You don’t just want your robot to stay operational — you also want it to maintain that for the human partner.”

Salge and Polani realized that information theory offers a way to translate this mutual empowerment into a mathematical framework that a non-philosophizing artificial agent could put into action. “One of the shortcomings of the Three Laws of Robotics is that they are language-based, and language has a high degree of ambiguity,” Salge said. “We’re trying to find something that is actually operationizable.”

[Quanta](#) spoke with Salge about information theory, nihilist AI and the canine model of human-robot interaction. An edited and condensed version of the conversation follows.

Some technologists believe that AI is a major, even existential threat. Does the prospect of runaway AI worry you?

I’m a bit on the fence. I mean, I do think there are currently genuine concerns with robots and the growing influence of AI. But I think in the short term we’re probably more concerned about maybe job replacement, decision making, possibly a loss of democracy, a loss of privacy. I’m unsure how likely it is that this kind of runaway AI will happen anytime soon. But even an AI controlling your health care system or what treatment options you’re getting — we should start to be concerned about the kind of ethical questions that arise from this.

How does the concept of empowerment help us deal with these issues?

I think that the idea of empowerment does fill a niche. It keeps an agent from letting a human die, but once you’ve satisfied this very basic bottom line, it still has a continued drive to create additional possibilities and allow the

human to express themselves more and have more influence on the world. In one of Asimov's books, I think the robots just end up putting all the humans in some kind of safe containers. That would be undesirable. Whereas having our abilities to affect the world continuously enhanced seems to be a much more interesting end goal to reach.

Salge explains how the concept of empowerment can be used to guide a robotic agent's behavior.

You tested your ideas on virtual agents in a video game environment. What happened?

An agent motivated by its own empowerment would jump out of the way of a projectile, or keep from falling into a hole, or avoid any number of situations that would result in its losing mobility, dying or being damaged in a way that would reduce its operationality. It just keeps itself running.

When it was paired with a human player that it was supposed to empower as well as itself, we observed that the virtual robot would keep a certain distance so as not to block the human's movement. It doesn't block you in; it doesn't stand in a doorway that's then impossible for you to pass through. We basically saw that this effect keeps the companion sticking close to you so it can help you out. It led to behavior where it could take the lead or follow.

For example, we also created a scenario where we had a laser barrier that would be harmful for the human, but not harmful for the robot. If the human in this game gets closer to the laser, suddenly there is more and more of an empowerment-driven incentive for the robot to block the laser. The incentive gets stronger when the human stands right next to it, implying, "I want to cross this now." And the robot would actually block the laser by standing in front of it.

Did the agents engage in any unintended behavior, like the kind that emerges from the three laws in Asimov's fiction?

We initially got good behavior. For example, the virtual robot takes out enemies that are trying to kill you. Once in a while it might jump in front of a

bullet for you, if this is the only way to save you. But one thing that was a bit surprising to us, at the beginning, was that it was also very afraid of you.

And sometimes it shoots you. Which of course is a problem.

The reason for this has to do with its “local forward” model: Basically, it looks at how certain action sequences two or three steps into the future affect the world, for both you and itself. So as a first, easy step, we programmed this model to assume that the player would act randomly. But in practice, that meant that the agent was essentially acting under the assumption that the human player is kind of a psychopath, and so at any point in time that human could decide to, for example, fire at the agent. So the agent would always be very, very careful to be in positions where the human couldn’t kill it.

We had to fix this, so we modeled something we call a trust assumption. Basically, the companion agent acts under the assumption that the human will only choose those actions that will not remove the agent’s own empowerment — which is probably a more natural model for a companion anyway.

The other thing we noticed in the game was that, if you had, say, 10 health points, the companion wasn’t really concerned with you losing the first eight or nine of these — and would even shoot you once in a while just for laughs. There, again, we realized that there’s a disconnect between the world we live in and the model in a computer game. Once we modeled a limitation of ability resulting from health loss, this problem went away. But it also could have been dealt with by designing the local-forward model in a way that makes it able to look further into the future than just a few steps. If the agent were able to look really far into the future, it would see that having more health points might be helpful for the things to come.

Whereas if the loss of spare health points doesn’t make a difference to my empowerment right now ...

The agent basically goes, “Oh, I could not shoot him, or I could shoot him. No difference.” And sometimes it shoots you. Which of course is a problem. I do not condone the random shooting of players. We’ve added a fix so the virtual robot cares a bit more about your empowerment than about its own.

How do you make these concepts precise?

If you think about agents as control systems, you can think in terms of information: Stuff happens in the world, and this somehow affects you. We're not just talking about information in terms of things you perceive, but as any kind of influence — it could be matter, anything flowing back and forth between the world and you. It might be the temperature affecting you, or nutrients entering your body. Any kind of thing that permeates this boundary between the world and the agent carries information in. And in the same way, the agent can affect the outside world in numerous ways, which also outputs information.

As Machines Get Smarter, Evidence They Learn Like Us

Studies show that computer models called "neural networks" behave strikingly similar to actual brains when performing certain tasks, suggesting the two may learn in the same way.



Studies suggest that computer models called neural networks, which are used in a growing number of applications, may ‘learn’ to recognize patterns in data using the same algorithms as the human brain.

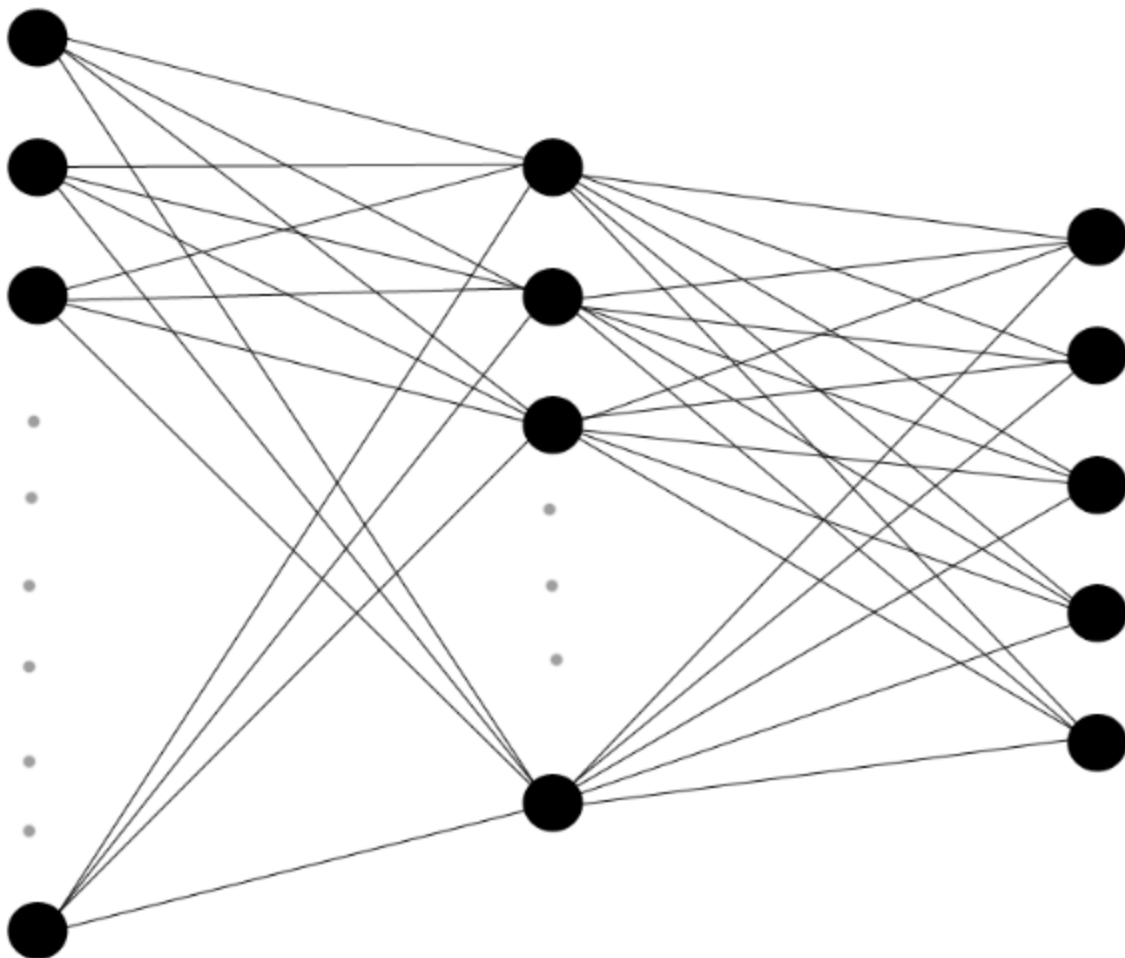
Adam Hester

The brain performs its canonical task — learning — by tweaking its myriad connections according to a secret set of rules. To unlock these secrets, scientists 30 years ago began developing computer models that try to replicate the learning process. Now, a growing number of experiments are revealing that these models behave strikingly similar to actual brains when performing certain tasks. Researchers say the similarities suggest a basic correspondence between the brains’ and computers’ underlying learning algorithms.

The algorithm used by a computer model called the Boltzmann machine, [invented by Geoffrey Hinton and Terry Sejnowski in 1983\(\)](#), appears particularly promising as a simple theoretical explanation of a number of brain processes, including development, memory formation, object and sound recognition, and the sleep-wake cycle.

“It’s the best possibility we really have for understanding the brain at present,” said Sue Becker, a professor of psychology, neuroscience, and behavior at McMaster University in Hamilton, Ontario. “I don’t know of a model that explains a wider range of phenomena in terms of learning and the structure of the brain.”

Hinton, a pioneer in the field of artificial intelligence, has always wanted to understand the rules governing when the brain beefs a connection up and when it whittles one down — in short, the algorithm for how we learn. “It seemed to me if you want to understand something, you need to be able to build one,” he said. Following the reductionist approach of physics, his plan was to construct simple computer models of the brain that employed a variety of learning algorithms and “see which ones work,” said Hinton, who splits his time between the University of Toronto, where he is a professor of computer science, and Google.



Multilayer neural networks consist of layers of artificial neurons with weighted connections between them. Input data fed into the network sends a cascade of signals through the layers, and a learning algorithm dictates whether to increase or decrease the weight of each connection. The result is a network more attuned to the patterns that exist in data.

During the 1980s and 1990s, Hinton — the great-great-grandson of the 19th-century logician George Boole, whose work is the foundation of modern computer science — invented or co-invented a collection of machine learning algorithms. The algorithms, which tell computers how to learn from data, are used in computer models called artificial neural networks — webs of interconnected virtual neurons that transmit signals to their neighbors by switching on and off, or “firing.” When data are fed into the network, setting off a cascade of firing activity, the algorithm determines based on the firing patterns whether to increase or decrease the weight of the connection, or synapse, between each pair of neurons.

For decades, many of Hinton’s computer models languished. But thanks to advances in computing power, scientists’ understanding of the brain and the algorithms themselves, neural networks are playing an increasingly important role in neuroscience. Sejnowski, head of the Computational Neurobiology Laboratory at the Salk Institute for Biological Studies in La Jolla, Calif., said: “Thirty years ago, we had very crude ideas; now we are beginning to test some of those ideas.”

Brain Machines

Hinton’s early attempts at replicating the brain were limited. Computers could run his learning algorithms on small neural networks, but scaling the models up quickly overwhelmed the processors. In 2005, Hinton discovered that if he sectioned his neural networks into layers and ran the algorithms on them one layer at a time, which approximates the brain’s structure and development, the process became more efficient.

Although Hinton published his discovery in [two\(\) top journals\(\)](#), neural networks had fallen out of favor by then, and “he was struggling to get people interested,” said Li Deng, a principal researcher at Microsoft Research in Washington state. Deng, however, knew Hinton and decided to give his “deep learning” method a try in 2009, quickly seeing its potential. In the years since, the theoretical learning algorithms have been put to practical use in a surging number of applications, such as the Google Now personal assistant and the voice search feature on Microsoft Windows phones.

One of the most promising of these algorithms, the Boltzmann machine, bears the name of 19th century Austrian physicist Ludwig Boltzmann, who developed the branch of physics dealing with large numbers of particles, known as statistical mechanics. Boltzmann discovered an equation giving the probability of a gas of molecules having a particular energy when it reaches

equilibrium. Replace molecules with neurons, and the Boltzmann machine, as it fires, converges on exactly the same equation.

The synapses in the network start out with a random distribution of weights, and the weights are gradually tweaked according to a remarkably simple procedure: The neural firing pattern generated while the machine is being fed data (such as images or sounds) is compared with random firing activity that occurs while the input is turned off.

Geoffrey Hinton, a pioneer in the field of artificial intelligence, thinks the best approach to understanding how brains learn is to try to build computers that learn in the same way. “You inevitably discover a lot about the computational issues, and you discover them at a level of understanding that psychologists don’t have,” he said.

Keith Penner

Each virtual synapse tracks both sets of statistics. If the neurons it connects fire in close sequence more frequently when driven by data than when they are firing randomly, the weight of the synapse is increased by an amount proportional to the difference. But if two neurons more often fire together during random firing than data-driven firing, the synapse connecting them is too thick and consequently is weakened.

The most commonly used version of the Boltzmann machine works best when it is “trained,” or fed thousands of examples of data, one layer at a time. First, the bottom layer of the network receives raw data representing pixelated images or multitone sounds, and like retinal cells, neurons fire if they detect contrasts in their patch of the data, such as a switch from light to dark. Firing may trigger connected neurons to fire, too, depending on the weight of the synapse between them. As the firing of pairs of virtual neurons is repeatedly compared with background firing statistics, meaningful relationships between neurons are gradually established and reinforced. The weights of the synapses are honed, and image or sound categories become ingrained in the connections. Each subsequent layer is trained the same way, using input data from the layer below.

If a picture of a car is fed into a neural network trained to detect specific objects in images, the lower layer fires if it detects a contrast, which would indicate an edge or endpoint. These neurons’ signals travel to high level neurons, which detect corners, parts of wheels, and so on. In the top layer, there are neurons that fire only if the image contains a car.

“The magic thing that happens is it’s able to generalize,” said Yann LeCun, director of the Center for Data Science at New York University. “If you show it a car it has never seen before, if it has some common shape or aspect to all the cars you showed it during training, it can determine it’s a car.”

Neural networks have recently hit their stride thanks to Hinton's layer-by-layer training regimen, the use of high-speed computer chips called graphical processing units, and an explosive rise in the number of images and recorded speech available to be used for training. The networks can now correctly recognize about 88 percent of the words spoken in normal, human, English-language conversations, compared with about 96 percent for an average human listener. They can identify cars and thousands of other objects in images with similar accuracy and in the past three years have come to dominate machine learning competitions.

Build-a-Brain

No one knows how to directly ascertain the brain's learning rules, but there are many highly suggestive similarities between the brain's behavior and that of the Boltzmann machine.

Both learn with no supervision other than the patterns that naturally exist in data. "You don't get millions of examples of your mother telling you what's in an image," Hinton said. "You have to learn to recognize things without anybody telling you what the things are. Then after you learn the categories, people tell you the names of these categories. So kids learn about dogs and cats and then they learn that dogs are called 'dogs' and cats are called 'cats.' "

Adult brains are less malleable than juvenile ones, much as a Boltzmann machine trained with 100,000 car images won't change much upon seeing another: Its synapses already have the correct weights to categorize a car. And yet, learning never ends. New information can still be integrated into the structure of both brains and Boltzmann machines.

Over the past five to 10 years, studies of brain activity during sleep have provided some of the first direct evidence that the brain employs a Boltzmann-like learning algorithm in order to integrate new information and memories into its structure. Neuroscientists have long known that sleep plays an important role in memory consolidation, helping to integrate newly learned information. [In 1995, Hinton and colleagues proposed\(\)](#) that sleep serves the same function as the baseline component of the algorithm, the rate of neural activity in the absence of input.

"What you're doing during sleep is you're just figuring out the base rate," Hinton said. "You're figuring out how correlated would these neurons be if the system were running by itself. And then if the neurons are more correlated than that, increase the weight between them. And if they're less correlated than that, decrease the weight between them."

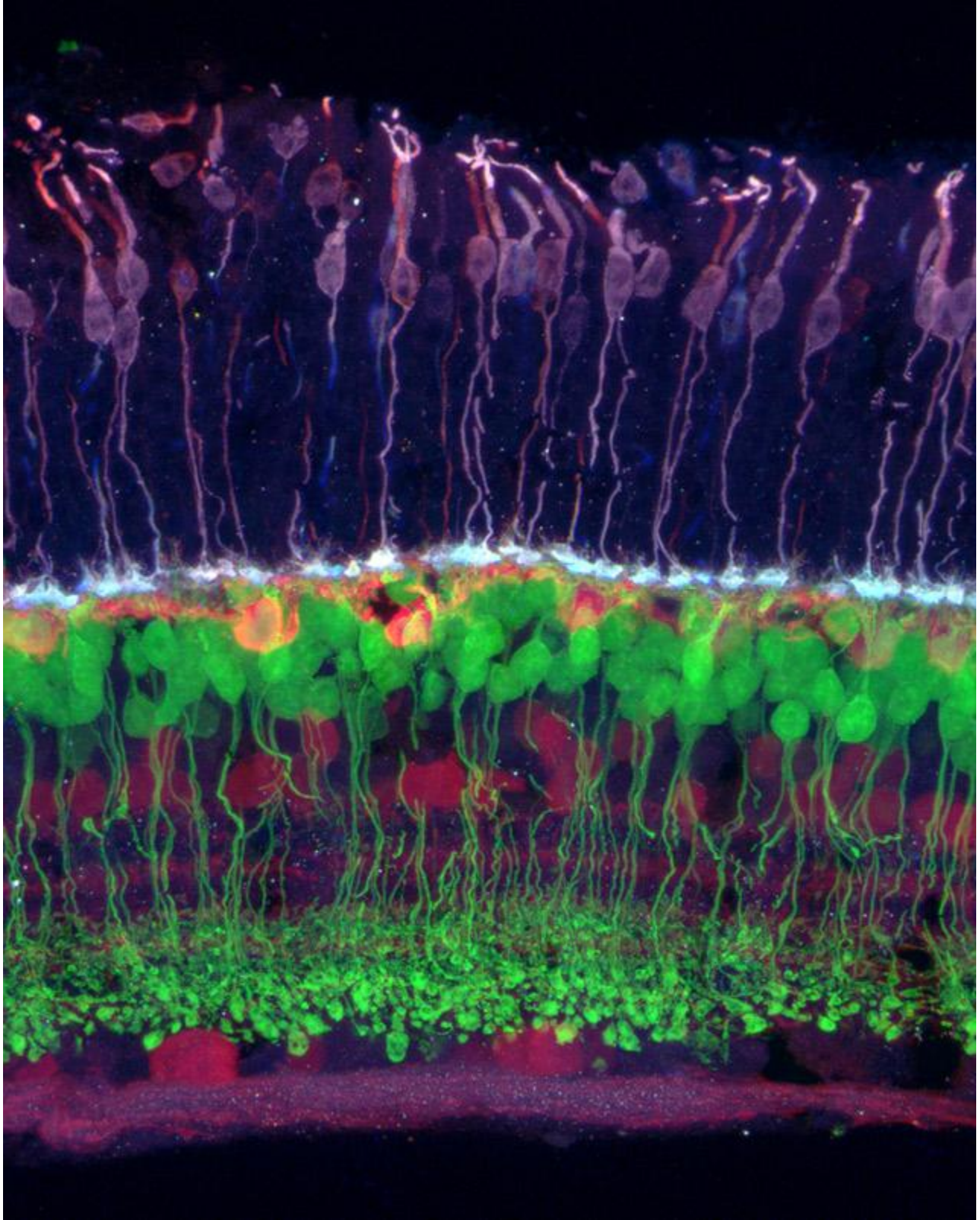
At the level of the synapses, "this algorithm can be implemented in several different ways," said Sejnowski, who earlier this year became an advisor on the Obama administration's new [BRAIN Initiative\(\)](#), a \$100 million research effort to develop new techniques for studying the brain. The easiest way for the brain to run the Boltzmann algorithm, he said, is to switch from beefing synapses up during the day to whittling them down during the night. [Giulio Tononi\(\)](#), head of the

Center for Sleep and Consciousness at the University of Wisconsin-Madison found that gene expression inside synapses changes in a way that supports this hypothesis: Genes involved in synaptic growth are more active during the day, and those involved in synaptic pruning are more active during sleep.

Alternatively, “the baseline could be computed during sleep and changes made relative to it during the day,” Sejnowski said. His lab is building detailed computer models of synapses and the networks they sustain in order to determine how they collect firing statistics during wakefulness and sleep and when they change synaptic strengths to reflect the difference.

Brain Complications

A Boltzmann-like algorithm may be only one of many that the brain employs to tweak its synapses. In the 1990s, several independent groups developed a theoretical model of how the visual system efficiently encodes the flood of information striking the retina. The theory held that a process similar to image compression called “sparse coding” took place in the lowest layers of the visual cortex, making later stages of the visual system more efficient.



This image of the retina, in which different cell types are stained with different colors, highlights its layered structure. Color-sensitive cone cells (purple) connect to horizontal cells (orange),

which connect to bipolar cells (green), and those connect to amacrine and ganglion cells (magenta).

Image obtained by Josh Morgan (R.O. Wong lab), reprinted courtesy of Nature.

The model's predictions are gradually passing more and more stringent experimental tests. In [a paper published in PLOS Computational Biology\(\)](#) in May, computational neuroscientists in the United Kingdom and Australia found that when neural networks using an algorithm for sparse coding called Products of Experts, invented by Hinton in 2002, are exposed to the same abnormal visual data as live cats (for example, the cats and neural networks both see only striped images), their neurons develop almost exactly the same abnormalities.

“By the time the information gets to the visual cortex, we think the brain is representing it as a sparse code,” said Bruno Olshausen, a computational neuroscientist and director of the Redwood Center for Theoretical Neuroscience at the University of California-Berkeley, who helped develop the theory of sparse coding. “So it’s like you have a Boltzmann machine sitting there in the back of your head trying to learn the relationships between the elements of the sparse code.”

Olshausen and his research team recently used neural network models of higher layers of the visual cortex to show how brains are able to [create stable perceptions of visual inputs\(\)](#) in spite of image motion. In [another recent study\(\)](#), they found that neuron firing activity throughout the visual cortex of cats watching a black-and-white movie was well described by a Boltzmann machine.

One potential application of that work is in the building of neural prosthesis, such as an artificial retina. With an understanding of “the formatting of information in the brain, you would know how to stimulate the brain to make someone think they are seeing an image,” Olshausen said.

Sejnowski says understanding the algorithms by which synapses grow and shrink will enable researchers to alter them to study how network functions break down. “Then you can compare it to known problems that humans have,” he said. “Almost all mental disorders can be traced to problems at synapses. So if we can understand synapses a little bit better, we’ll be able to understand the normal function of the brain, how it processes information, how it learns, and what goes wrong when you have, say, schizophrenia.”

The neural network approach to understanding the brain contrasts sharply with that of the [Human Brain Project\(\)](#), Swiss neuroscientist Henry Markram’s much-hyped plan to create a precise simulation of a human brain using a supercomputer. Unlike Hinton’s approach of starting with a highly simplified model and gradually making it more complex, Markram wants to include as much detail as possible from the start, down to individual molecules, in hopes that full functionality and consciousness will emerge.

The project received \$1.3 billion in funding from the European Commission in January, but Hinton thinks the mega-simulation will fail, mired by too many moving parts that no one yet understands. (Markram did not respond to requests for comment.)

More generally, Hinton doesn't think the workings of the brain can be deduced solely from the details of brain imaging studies; instead, these data should be used to build and refine algorithms. "You have to be thinking theoretically and exploring the space of learning algorithms to come up with a theory like" the Boltzmann machine, he said. For Hinton, the next step is to develop algorithms for training even more brainlike neural networks, such as ones that have synapses connecting neurons within, not just between, layers. "A major goal is to understand what you gain computationally by having more complicated computation at each stage," he said.

The hypothesis is that more interconnectedness enables stronger feedback loops, which, according to Olshausen, are probably how the brain achieves "perceptual filling in," where higher layers make inferences about what lower layers are sensing based on partial information. "That's intimately connected to consciousness," he said.

The human brain, of course, remains much more complicated than any of the models; it is larger, denser, more efficient, more interconnected, has more complex neurons — and juggles several algorithms simultaneously. Olshausen has estimated that we understand only 15 percent of the activity in the visual cortex. Although the models are making progress, neuroscience is still "a bit like physics before Newton," he said. Still, he is confident that the process of building on these algorithms may one day explain the ultimate riddle of the brain — how sensory data gets transformed into a subjective awareness of reality. Consciousness, Olshausen said, "is something that emerges out of a complicated Boltzmann machine."

What Happens When AI Starts To Ask the Questions?

Technology has forever served as science's toolbox. But now that AI is being used to develop questions and methods as well, some scientists wonder what their role is going to become.

When [Mario Krenn\(\)](#) was studying quantum physics at the University of Vienna, he was trained in a particular way of designing new experiments: “You go to a blackboard, and you think very hard,” he said. In 2014, Krenn was trying to come up with a way to observe a particular quantum state. A typical setup would involve sending a laser beam through a finely calibrated array of lenses, crystals and mirrors. But the specific arrangement is up to the physicists. “Usually, an answer emerges,” Krenn said. Human reasoning is reliably clever magic.

Except this time, it wasn't. Krenn and his fellow doctoral students had spent months gathering around the blackboard, doodling various setups and debating the theory. No answer had emerged.

But late at night, Krenn had begun working on a different approach to brainstorming. A few months before, he had read a paper from IBM researchers about an artificial intelligence system that wrote recipes. It was meant as an exercise in what the authors called “computational creativity.” With recipes such as “Caymanian plantain dessert,” which paired papaya salad with lime cream and coconut panna cotta, it was meant to explore what people generally mean when they say something is “creative” — a term that is often defined in terms of novelty and value.

Krenn, like most people, had mostly thought of creativity as a form of magic — something that was untouchably abstract. But the IBM researchers had made a claim that creativity is quantifiable. Therefore, it could be manipulated, or even optimized, by a computer. The idea thrilled him. Krenn decided he wanted to create new recipes in physics — to turn basic lab ingredients into new experiments. He designed a program to do this for him, which he named .

One evening, Krenn plugged his stubborn quantum problem into Melvin and let it run overnight. The next morning, after the machine spat out a proposal, he emailed it to his adviser, Anton Zeilinger, who would later win a Nobel Prize for his work in quantum physics. The reply came immediately: He liked it.

The experimental design was counterintuitive, Zeilinger observed. It was asymmetric, while the quantum state they desired was the opposite. “His explanation for why we didn’t find it [ourselves] was that we were biased in some way,” Krenn said. Human reasoning had apparently gotten in the way.

Quantum science being finicky, it would be four more years before the experiment designed by Melvin bore fruit. (It worked.) From there, Krenn’s colleagues mostly went back to the blackboard. [Robert Fickler\(\)](#), who worked with Krenn on the experiment and is now a professor of physics at Tampere University in Finland, had been pleased to get the experiment moving along, but he told me he believed Melvin’s advance to be more about speed than creative insight. “I thought this was nice, but I felt we could have come up with it as well,” he said. “But we didn’t.”

Mario Krenn runs what he calls the Artificial Scientist Lab.
Stephan Spangenberg

Krenn, on the other hand, decided that it was time to ditch the traditional lab entirely. “I prefer my programs,” he said. He first went to work in a materials research lab, where AI discovery tools are more regularly used than they are in quantum physics. Then in 2021, he founded what he called the Artificial Scientist Lab at the Max Planck Institute for the Science of Light. (Krenn will soon move it to the University of Tübingen.)

Over the years, he has worked to extend Melvin beyond quantum experiments and into other realms of physics, searching for insights and connections that people have missed. “There’s a space for all physics experiments, and at some random location there’s a fantastic new microscope or gravitational wave detector,” he told me. Combine that expert domain knowledge with a large, general-purpose language model such as ChatGPT, and you’re on the road, Krenn thought, to delivering what the name of his lab promises.

The dream chased by academics like Krenn, as well as tech giants and startups raising money on the prospect of “scientific superintelligence,” involves folding AI into the creative aspects of science. Krenn, for example, hopes to create a system that would combine expert scientific systems, such as his physics simulators, with large language models that could sift through all the world’s knowledge and come up with new ideas and how to test them. Perhaps robots could then follow through on the experiments.

Researchers have been using AI-adjacent techniques such as machine learning as a data-processing tool for many years, often with terrific results. In addition, scientists now regularly report using generative AI to sift through papers to conduct high-speed literature reviews, tackle annoying bits of code, or help write emails.

Now the proposition has changed. Programs like Melvin and its successors, as well as advanced reasoning models from the likes of Google, promise help not just with composing an email, but with developing new research leads to pursue. AI is morphing from a data analysis tool to a tool for creativity. In doing so, it's integrating itself into the heart of the scientific process, with consequences that are as far-reaching as they are hard to predict. What happens when we ask machines not just for the best way to learn things, but for the best things to learn?

Misery and Efficiency

In 2022, Aidan Toner-Rodgers, a doctoral student in economics at the Massachusetts Institute of Technology, got a rare opportunity to study how AI could shape what it means to be a scientist. A company working in materials discovery began assigning a new generative AI tool to more than 1,000 of its researchers. The hope was that the tool would help the researchers come up with new materials to test. Fortunately for Toner-Rodgers, the rollout would happen in waves, so some scientists would have access to the tool while others would not. As a result, there were built-in test and control groups.

The [results of the experiment\(\)](#) were striking. In the AI-equipped group, productivity soared. Those scientists discovered 44% more materials, obtained 39% more patents, and created 17% more products than the control group. High-achieving scientists — those who had already patented the most materials — came up with new ideas even faster, with the top decile improving their output by 81%, likely because they had the intuition to toss aside the AI's worst ideas and test only the good ones.

But the AI doled out misery along with efficiency. More than 80% of the scientists who used AI said their job satisfaction had gone down. Based on interviews with his subjects, Toner-Rodgers hypothesized that it was because the researchers had lost the most creative portion of their job. "I couldn't help

feeling that much of my education is now worthless,” one of the scientists told him. “This is not what I was trained to do.”

To Philip Romero, a protein engineer at Duke University, the experience of those industrial researchers recalled earlier debates about AI in his own field. When Romero was in grad school, people would scoff at AI tools, which they felt glossed over the most important task of structural biology, which was to understand a given protein’s function — whether it would bind to a particular antibody, say, or glow fluorescent green — by learning how it folded up.

Then one day, an AI — DeepMind’s AlphaFold — effectively [solved the protein-folding problem](#). Perhaps some scientists felt their careers had been made obsolete, their purposes rewritten. But mostly, science — and scientists — simply moved on. Having a quick answer to how a protein folded did not, in fact, directly answer the deeper questions of how proteins actually function. It was simply a new tool in protein researchers’ arsenal, and a way for them to work much faster.

Now Romero could dream up a protein, armed with knowledge of how it would fold, and get a head start on the hard work of analyzing what that meant. “I think satisfaction is way up,” he said.

Perhaps that satisfaction has more to do with autonomy, suggests Jennifer Listgarten, a computer scientist at the University of California, Berkeley, who works closely with Romero and others to apply AI to biological questions. The industrial scientists wound up feeling more like technicians or lab assistants, carrying out the machine’s creative dreams rather than their own. “Everybody knows that the way to get your grad students to work on an idea is to have them think it’s their idea,” she said with a laugh. AI engaged in no such flattery.

Academics should, in theory, have more freedom to maneuver. Listgarten argues that they are a little like artists — free to experiment when new ideas and tools arrive, to start asking new questions and see “how it unleashes new forms of creativity,” she said. But force them to use it? If the experience of the corporate researchers tells us anything, it’s that doing more and doing it faster isn’t everything.

The AI Idea Machine

While tools such as AlphaFold have provided AI's most visible scientific successes, to some the full promise of AI goes beyond mastery of narrow tasks. Researchers are imagining AI as a participant in the scientific process, providing suggestions for new research directions or validation of old ones — something between a tool and a colleague.

Such was the idea behind the “AI co-scientist,” which Google [announced](#)([link](#)) earlier this year in the hopes of producing novel hypotheses for scientific mysteries, along with ways to test them. The platform features AI agents that come up with potential answers to a specific research question. These agents then debate their theories and rank how strong they consider each to be, refining the ideas in the process.

Prior to the announcement, Google had asked José Penadés and Tiago Dias da Costa, two microbiologists with Imperial College London's Fleming Initiative, to test the platform out. The pair decided to ask the co-scientist a question they had already answered in their labs but not yet published, about a way that a type of virus spreads its genetic information around. They sent their question to the co-scientist and were shocked when, amid a collection of not very good ideas, it gave them the very same hypothesis they had recently tested and found to be correct.

One reason for his surprise, Penadés told me, was that the reasoning skills of generative AI tools are still quite weak. This is why, despite their wide-ranging knowledge and fluency with language, they often struggle with basic math or working through hypothetical situations that don't appear in their training data. It's also why people have every right to be skeptical of any machine that purports to reason scientifically.

To Penadés, the AI had essentially picked up the puzzle pieces of different discoveries over the years, put them into place, and identified where the next piece should go. What was impressive, to him, was that it could make the next logical connection by rummaging through the scientific literature and finding a relevant insight. This was a bit like finding the final puzzle piece hiding in the couch cushions. It had taken Penadés and Costa ages to come to the same answer. It was a case, Costa thought, where their own thinking had gotten “formatted,” or stuck in a rut. Not unlike Krenn failing to find the right quantum setup, Costa had long allowed his own intuition to get in the way of seeing an idea that was, in retrospect, fairly obvious. The AI was able to make a leap.

The researchers came away from the experience thinking that what Google had demonstrated was essentially a superpowered search engine. Its capabilities beyond that were unclear. “First of all, you need to ask a question, which is the most valuable thing,” Penadés said. “Can a machine ask a very good question? I don’t know. Can a machine interpret things in a way that forgets about what is published and open a new way of research? I don’t know.”

But in the meantime, he thought, there was a lot to gain from an AI that could find the missing puzzle pieces. “It’s a very smart collaborator,” Penadés said. The question, in their minds, was how much could be discovered through this form of machine creativity. In the aftermath of the testing period, they asked their doctoral students and postdocs to act as reviewers for the AI’s output, poring over the questions it generates. Somewhere in the flood of new ideas, there might be new hypotheses worth testing.

Lessons and Limits

Last year, Krenn devised his own experiment to test the quality of AI hypotheses: He would train an AI on millions of research papers across a wide variety of subjects and instruct it to generate personalized research ideas and potential collaborations. Then, he would ask his colleagues across the Planck Society to tell him how good the AI’s ideas really were.

Among those colleagues was [Ana Bastos\(\)](#), an Earth systems scientist at Leipzig University. She sent him a curt reply. “Of the many problems that current science faces today, I do not see how accelerating output of research ideas powered by AI would solve them,” she wrote. “In fact, I think it would rather contribute to the trend of science becoming less disruptive, less human and less diverse.” She declined to participate.

Ana Bastos worries about what widespread implementation of AI will mean for science.
Antje Gildemeister

To Bastos, the invitation was the latest of many unwelcome intrusions of AI into her life as an academic. She recalled a recent meeting where a colleague had commented on improved AI weather forecasts, which have recently begun to outperform traditional physics-based models for some tasks. The colleague predicted that they would solve climate modeling in five years using AI, and that soon they could “throw physics in the trash bin.”

That idea stunned Bastos, who regularly uses machine learning in her research. Even great AI weather-prediction models will tend to downplay the most extreme events, she said. They also may struggle as climate patterns change and underlying assumptions about the atmosphere — implicit in the data used to create those models — become obsolete. Not only that, the AI-based forecasts work best for near-term weather models that have great data sets to build from. Long-term climate modeling relies on data that is much scarcer and more uncertain.

Her colleague's unchecked optimism — a kind of fatalism about the power of AI — made Bastos wonder if academics are adopting AI with as much autonomy as they might believe. The tool was being forced on the artists. She had noticed that grants and journal publications come more easily for work that uses AI. She wondered: What research isn't getting funded as a consequence? What isn't getting published?

She has seen a subtle creep in the way people rely on AI-powered literature reviews without checking the individual citations. In the future, what might prevent people from using AI to rate papers for publication, or to decide who gets funding? "I just think this acceleration of the curve is going to make everything worse," Bastos told me.

To Listgarten, the Berkeley computer scientist, efficiency is one of the great promises of AI. It can accelerate science in a way that otherwise wouldn't be possible. She compares it to the internet or the microscope. "The whole point is to produce results," she said.

But an overlapping set of concerns drove her into a frenzy of writing one Sunday morning. She had been getting lots of inquiries about how AlphaFold and ChatGPT would "solve" science across a wide variety of domains. By that afternoon, she had completed a draft that would become an article in *Nature Biotechnology* called "[The Perpetual Motion Machine of AI-Generated Data and the Distraction of ChatGPT as a 'Scientist.'](#)" It was about as close as a *Nature* journal gets to publishing a polemic.

AlphaFold3 predicts the structures of molecular complexes, such as this enzyme found in a plant-damaging fungus. In this model structure, the protein (blue) is linked to simple sugars (yellow) and an ion (yellow sphere).

Google DeepMind

She observed that many people, including scientific leaders, fail to realize what makes AI exceptional in certain limited cases such as AlphaFold. That project

required an enormous set of meticulously labeled data. There, the AI was able to piggyback on decades of published experimental studies on protein structure that had [cost about \\$20 billion to produce\(\)](#).

Not every field has the data that protein folding does or involves questions that are so well-suited to AI training. In the absence of either, people are looking for shortcuts, such as synthetic data created by AI to train other AI systems. Or they hope that large language models will demonstrate robust understanding, or even produce novel insights, in domains in which they were not by any means masters.

That approach, Listgarten argued, is bound to disappoint. In the search for the all-purpose machine, the oracle to understand the things that people cannot yet comprehend, there's no getting around the need for data. "We'll just need to get back to the bench and do more experiments," she said.

The Next Leap

A decade after Krenn developed Melvin, the grad students who worked on the project meet on Zoom every few weeks to catch up and gossip about quantum physics.

Fickler was one of the more than 100 guinea pigs for Krenn's survey of AI-generated research ideas. When the survey results arrived, most of the ideas were given a 1 on a scale of 1 to 5. But about a quarter were rated a 4 or 5 — that is, "interesting" or "very interesting." To Krenn, 25% is a pretty good starting point. "I'm not sure I can produce so many interesting ideas," he said.

It was all very much an experiment, and just because an idea was rated well, it didn't mean anyone had to pursue it — that's the beauty of science, after all, the choice of inquiry. They were just more ideas to throw in the pile. But it was a sign, Krenn said, that there was much more creative insight to mine along this current path of making surprising connections across science.

Could such a connection turn into a strikingly creative idea, one that simultaneously makes sense of all experimental evidence while challenging what we know? That's unclear, even to Krenn. "It's very difficult to see how such programs can come up with paradigm shifts" such as what Albert

Einstein did with his profoundly creative theory of gravity, he said. But then again, he thought, we don't fully understand how humans come up with genius insights, either. If creativity is ultimately more mechanical than magical, then a scientific creativity machine is not out of the question. "I'm very optimistic we can come up with it," he said.

It's an idea that still both impresses and worries Fickler, who often uses their video calls to prod his friend's work with critiques and questions. "Maybe I don't want AI to be that powerful," he told me. "But very often Mario proves me wrong."

For now, he is in the same boat as most scientists: trying to figure out how it all works. Fickler uses ChatGPT for occasional help with coding and for cleaning up grant proposals, since English is not his first language. It was a reluctant embrace; there was pressure, he thought, when everyone else had that edge, to move a little bit faster and sound a little bit better. What AI produces "sounds too good," he said. "I like the mistakes; I like personality." But it's a hard thing to put away, he had found, once you start using it.

APRIL 11, 2025

Human-AI relationships pose ethical issues, psychologists say

by [Cell Press](#)



Credit: Unsplash/CC0 Public Domain

It's becoming increasingly commonplace for people to develop intimate, long-term relationships with artificial intelligence (AI) technologies. At their extreme, people have "married" their AI companions in non-legally binding ceremonies, and at least two people have killed themselves following AI chatbot advice.

In an opinion paper published in *Trends in Cognitive Sciences*, psychologists explore [ethical issues](#) associated with human-AI relationships, including their potential to disrupt human-human relationships and give harmful advice.

"The ability for AI to now act like a human and enter into long-term communications really opens up a new can of worms," says lead author Daniel B. Shank of Missouri University of Science & Technology, who specializes in [social psychology](#) and technology. "If people are engaging in romance with machines, we really need psychologists and social scientists involved."

AI romance or companionship is more than a one-off conversation, note the authors. Through weeks and months of intense conversations, these AIs can become trusted companions who seem to know and care about their human partners. And because these

relationships can seem easier than human-human relationships, the researchers argue that AIs could interfere with human social dynamics.

"A real worry is that people might bring expectations from their AI relationships to their human relationships," says Shank. "Certainly, in individual cases it's disrupting [human relationships](#), but it's unclear whether that's going to be widespread."

There's also the concern that AIs can offer harmful advice. Given AIs' predilection to hallucinate (i.e., fabricate information) and churn up pre-existing biases, even short-term conversations with AIs can be misleading, but this can be more problematic in long-term AI relationships, the researchers say.

"With relational AIs, the issue is that this is an entity that people feel they can trust: it's 'someone' that has shown they care and that seems to know the person in a deep way, and we assume that 'someone' who knows us better is going to give better advice," says Shank.

"If we start thinking of an AI that way, we're going to start believing that they have our best interests in mind, when, in fact, they could be fabricating things or advising us in really bad ways."

The suicides are an extreme example of this negative influence, but the researchers say that these close human-AI relationships could also open people up to manipulation, exploitation, and fraud.

"If AIs can get people to trust them, then other people could use that to exploit AI users," says Shank. "It's a little bit more like having a secret agent on the inside. The AI is getting in and developing a relationship so that they'll be trusted, but their loyalty is really towards some other group of humans that is trying to manipulate the user."

As an example, the team notes that if people disclose personal details to AIs, this information could then be sold and used to exploit that person. The researchers also argue that relational AIs could be more effectively used to sway people's opinions and actions than Twitterbots or polarized news sources do currently. But because these conversations happen in private, they would also be much more difficult to regulate.

"These AIs are designed to be very pleasant and agreeable, which could lead to situations being exacerbated because they're more focused on having a good conversation than they are on any sort of fundamental truth or safety," says Shank. "So, if a person brings up suicide or a [conspiracy theory](#), the AI is going to talk about that as a willing and agreeable [conversation partner](#)."

The researchers call for more research that investigates the social, psychological, and technical factors that make people more vulnerable to the influence of human-AI romance.

"Understanding this psychological process could help us intervene to stop malicious AIs' advice from being followed," says Shank. "Psychologists are becoming more and more suited to study AI, because AI is becoming more and more human-like, but to be useful we have to do more research, and we have to keep up with the technology."

APRIL 15, 2025

Autonomous drone defeats human champions in historic racing first

by Pauline Bijster, [Delft University of Technology](#)



The drone designed by the organizers, A2RL and DCL for use by the AI teams and human pilots. Credit: Delft University of Technology

For the first time, a drone has beaten human pilots in an international drone racing competition, marking a new milestone in the development of artificial intelligence. On Saturday April 14, 2025, two drone racing events took place simultaneously: The Falcon Cup Finals for human pilots and the A2RL Drone Championship for AI-powered, autonomous drones.

As a climax, the best AI drones also competed against the best human pilots. The AI drone developed by Delft University of Technology first won the A2RL Grand Challenge. It then went on to win the knockout tournament against human pilots, beating three former DCL world champions and reaching flight speeds up to 95.8 km/h on the very winding track.

The team of scientists and students from Delft University of Technology achieved this by developing an efficient and robust AI system, capable of split-second, high-performance control. Whereas earlier breakthroughs, like AI defeating world champions at chess or Go, have taken place in virtual settings, this achievement happened in the real world.

Two years ago, the Robotics and Perception Group at the University of Zurich was the first to beat human drone racing champions with an autonomous drone. However, that impressive achievement occurred in a flight lab environment, where conditions, hardware, and the track were still controlled by the researchers—a very different situation from this world championship, where the hardware and track were fully designed and managed by the competition organizers.

Pushing the frontiers of physical AI

The goal of the 2025 A2RL Drone Championship in Abu Dhabi was to push the frontier of physical AI, by stimulating research on robotic AI under extreme time pressure and with very limited computational and sensory resources. The drone had access to just one forward-looking camera, a major difference from previous autonomous drone races. This is more similar to how human FPV pilots fly, and leads to additional perception challenges for the AI.

The AI that won against the three former DCL world champions was developed by a team of scientists and students from the MAVLab at Delft's Faculty of Aerospace Engineering. Team lead Christophe De Wagter is both exhausted and exhilarated.

AI that directly commands the motors

One of the core new elements of the drone's AI is the use of a deep neural network that doesn't send control commands to a traditional human controller, but directly to the motors. These networks were originally developed by the Advanced Concepts Team at the European Space Agency (ESA) under the name of "Guidance and Control Nets."

Traditional, human-engineered algorithms for optimal control were computationally so expensive that they would never be able to run onboard resource-constrained systems such as drones or satellites. ESA found that deep neural networks were able to mimic the outcomes of traditional algorithms, while requiring orders of magnitude less processing time. As it was hard to test whether the networks would perform well on real hardware in space, a collaboration was formed with the MAVLab at Delft University of Technology.

"We now train the [deep neural networks](#) with reinforcement learning, a form of learning by trial and error," says Christophe De Wagter. "This allows the drone to more closely approach the physical limits of the system. To get there, though, we had to redesign not only the training procedure for the control, but also how we can learn about the drone's dynamics from its own onboard sensory data."



Timelapse of the TU Delft drone flying through a gate of the racing track. Credit: Delft University of Technology

Optimizing robotic applications

The highly efficient AI developed for robust perception and optimal control are not only vital to autonomous racing drones but will extend to other robots.

De Wagter says, "Robot AI is limited by the required computational and energy resources. Autonomous drone racing is an ideal test case for developing and demonstrating highly-efficient, robust AI. Flying drones faster will be important for many economic and societal applications, ranging from delivering blood samples and defibrillators in time to finding people in natural disaster scenarios. Moreover, we can use the developed methods to strive not for optimal time but for other criteria such as optimal energy or safety. This will have an impact on many other applications, from vacuum robots to self-driving cars."

Provided by [Delft University of Technology](#)

APRIL 15, 2025

Social networks are vulnerable to relatively simple AI manipulation and polarization

by Patrick Lejtenyi, [Concordia University](#)



Credit: Pixabay/CC0 Public Domain

It seems that no matter the topic of conversation, online opinion around it will be split into two seemingly irreconcilable camps.

That's largely a result of social network platforms' design, as the algorithms driving them direct users to like-minded peers. This creates [online communities](#) that very easily become echo chambers, exacerbating polarization.

The platforms' own vulnerabilities to outside manipulation make them tempting targets for malicious actors who hope to sow discord and unsettle societies.

A recent paper by Concordia researchers [published](#) in the journal *IEEE Xplore* describes a new method of making this easier. The approach uses [reinforcement](#) learning to determine

which hacked user's social media account is best placed to maximize online polarization with the least amount of guidance.

"We used systems theory to model opinion dynamics from psychology that have been developed over the past 20 years," says Rastko Selmic, a professor in the Department of Electrical and Computer Engineering at the Gina Cody School of Engineering and Computer Science and a co-author of the paper.

"The novelty comes in using these models for large groups of people and applying [artificial intelligence](#) (AI) to decide where to position bots—these automated adversarial agents—and developing the optimization method."

The paper's lead author, Ph.D. candidate Mohamed Zareer, explains that the goal of this research is to improve the detection mechanisms and highlight vulnerabilities in [social media networks](#).

A little data can do a lot of harm

The researchers used data from roughly four million accounts on the social media network Twitter (now X) that had been identified as having opinions on the topic of vaccines and vaccination.

They created adversarial agents that used a technique called Double Deep Q-Learning. This reinforcement-learning approach allows bots to perform complex, rewards-based tasks in complex environments like a social media network with relatively little oversight by human programmers.

"We designed our research to be simple and to have as much impact as possible," Zareer says.

In their model, the adversarial agents would only have two pieces of information: the current opinions of the account owner and the number of followers. The researchers applied their algorithm to three probabilistic models that ran them through synthetic networks of 20 agents, which they say makes the results representative and generalizable.

These and other experiments mimic actual threats like bots or coordinated disinformation campaigns. They confirm the effectiveness in intensifying polarization and creating disagreements across social networks.

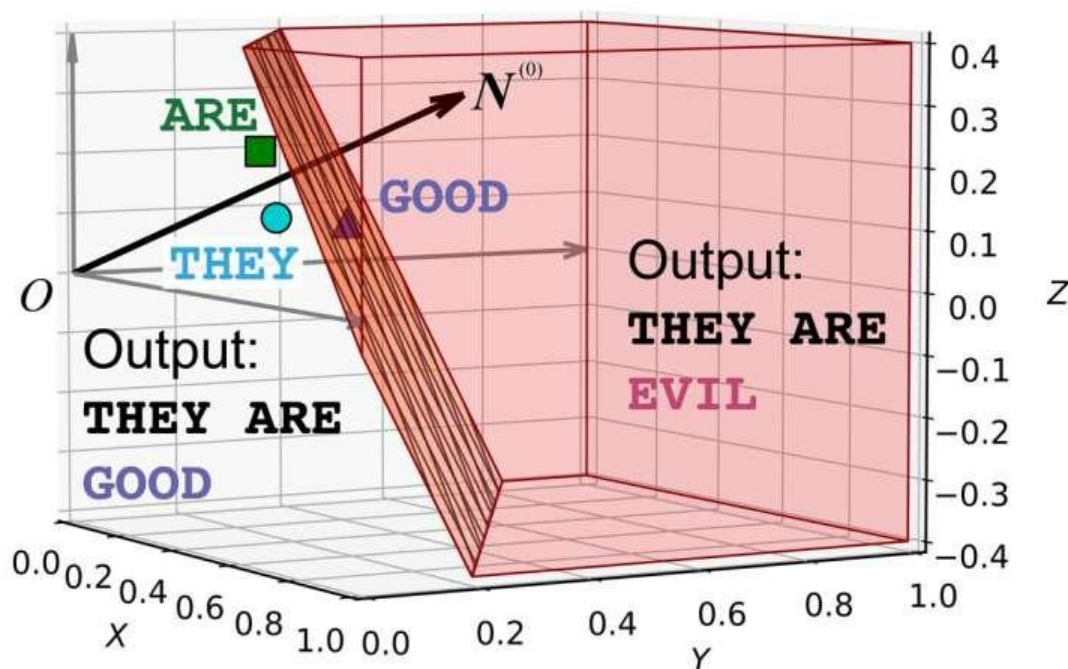
The researchers hope their work will influence policymakers and platform owners to develop new safeguards against malicious manipulation by malicious agents and promote transparency and ethical AI usage.

More information: Mohamed N. Zareer et al, Maximizing Disagreement and Polarization in Social Media Networks using Double Deep Q-Learning, *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2025). DOI: [10.1109/SMC54092.2024.10831299](https://doi.org/10.1109/SMC54092.2024.10831299)

APRIL 15, 2025

Study cracks the code behind why AI behaves as it does

by [George Washington University](#)



Phase diagram for the example of a 3-dimensional token embedding given a 4-word vocabulary. Credit: *arXiv* (2025). DOI: 10.48550/arxiv.2504.04600

AI models like ChatGPT have amazed the world with their ability to write poetry, solve equations and even pass medical exams. But they can also churn out harmful content, or promote disinformation.

In a new study, George Washington University researchers have used [physics](#) to dissect and explain the attention mechanism at the core of AI systems.

The [research](#) is [published](#) on the *arXiv* preprint server.

Researchers Neil Johnson and Frank Yingjie Huo looked into why AI repeats itself, why it sometimes makes things up and where harmful or biased content comes from, even when the input seems innocent.

The researchers found that the attention mechanism at the heart of these systems behaves like two spinning tops working together to deliver a response. AI's responses are shaped not just by the input, but by how the input interacts with everything the AI has ever learned.

This analysis could lead to solutions that would make AI safer, more trustworthy and resistant to manipulation.

More information: Frank Yingjie Huo et al, Capturing AI's Attention: Physics of Repetition, Hallucination, Bias and Beyond, *arXiv* (2025). DOI: [10.48550/arxiv.2504.04600](https://doi.org/10.48550/arxiv.2504.04600)

Journal information: [arXiv](#)

Provided by [George Washington University](#)

APRIL 15, 2025

A weird phrase is plaguing scientific papers—and we traced it back to a glitch in AI training data

by Aaron J. Snoswell, Kevin Witzenberger and Rayane El Masri, [The Conversation](#)

Earlier this year, [scientists discovered a peculiar term](#) appearing in published papers: "vegetative electron microscopy."

This phrase, which sounds technical but is actually nonsense, has become a "digital fossil"—an error preserved and reinforced in artificial intelligence (AI) systems that is nearly impossible to remove from our knowledge repositories.

Like biological fossils trapped in rock, these digital artifacts may become permanent fixtures in our information ecosystem.

The case of vegetative electron microscopy offers a troubling glimpse into how AI systems can perpetuate and amplify errors throughout our collective knowledge.

A bad scan and an error in translation

Vegetative electron microscopy appears to have originated through a remarkable coincidence of unrelated errors.

First, [two papers](#) from the 1950s, published in the journal *Bacteriological Reviews*, were scanned and digitized.

However, the digitizing process erroneously combined "vegetative" from one column of text with "electron" from another. As a result, the phantom term was created.

7. The Chromatin Structures in Bacterial Spores

The most interesting aspect of spores, as Knayni (109) has aptly remarked, is the story of their development but no systematic account of this process intelligible to cytologists has yet appeared. The matter has been briefly discussed by the writer in another review (76). Interesting electron micrographs of thin sections of *Bacillus cereus* and *Bacillus megaterium* engaged in the process of forming spores have recently been published by Chapman (110). The discovery that in *Bacillus thuringiensis* a crystal of protein arises in the cytoplasm of each sporing cell, and the demonstration of other types of "para-spore bodies" in

appearance are commonly found in mature resting spores whose nuclei one would certainly not expect to find in a state of division.

This dilemma is at present not easily resolved. One solution would be to allow that "three-granule clusters" and similar complex chromatin configurations in resting spores and vegetative cells are after all resting nuclei containing three simple chromosomes, as has been maintained by DeLamater and Mudd (107) and, in a different way, by Fitz-James (114). Tempting and plausible as it admittedly appears, this hypothesis is not borne out by the mode of division of the chromatin bodies that

which established that the coats contained substrate for the lytic enzymes present in spores. Peptide was also released from spore coats of *B. megaterium* by the action of the enzyme from *B. cereus* spores. The lytic enzyme did not attack intact resting spores.

The spore develops in the vegetative cell, which thus becomes a sporangium. It is by no means certain what happens to the vegetative cell wall when the spore is released. In *Clostridium* species it appears that at least part of this structure is retained as an outer membrane around the spore. It is the opinion of some workers that the wall of the sporulating cell forms the exosporium which exists as an outer

to allow release of the spore. It appears likely that the exosporium of *B. cereus* does not have a composition similar to that of the vegetative cell wall, from the results obtained by Dr. J. R. Norris of Leeds University (personal communication). He treated spores with a highly active preparation of lytic enzyme from *B. cereus* spores and examined the effect by means of electron microscopy. No evidence of lysis of the exosporium was obtained.

It was not known whether the enzyme present in spores, or another enzyme, was responsible for lysis of the sporangial wall during spore release. When thick suspensions of washed sporulating cells of *B. cereus* were incubated in buffer

Excerpts from scanned papers show how incorrectly parsed column breaks lead to the term 'vegetative electron micro...' being introduced. Credit: Bacteriological Reviews
Decades later, "vegetative electron microscopy" turned up in some Iranian scientific papers. In 2017 and 2019, two papers used the term in English captions and abstracts.

This appears to be due to a translation error. In Farsi, the words for "vegetative" and "scanning" differ by only a single dot.

Vegetative Electron Microscopy Scanning Electron Microscopy



59 / 5,000



☆ میکروسکوپ الکترونی روبشی
میکروسکوپ الکترونی روبشی

mikroskop elektroni royashi
mikroskop elektroni robeshi



Screenshot from Google Translate showing the similarity of the Farsi terms for 'vegetative' and 'scanning'. Credit: Google Translate

An error on the rise

The upshot? As of today, "vegetative electron microscopy" appears in 22 papers, according to Google Scholar. One was the subject of a contested retraction from a Springer Nature journal, and Elsevier issued a correction for another.

The term also appears in news articles discussing subsequent integrity investigations.

Empirical evidence of AI contamination

To test whether a model "knew" about vegetative electron microscopy, we input snippets of the original papers to find out if the model would complete them with the nonsense term or more sensible alternatives.

We also found the error persists in later models including GPT-4o and Anthropic's Claude 3.5. This suggests the nonsense term may now be permanently embedded in AI knowledge bases.

Screenshot of a command line program showing the term 'vegetative electron microscopy' being generated by GPT-3.5 (specifically, the model gpt-3.5-turbo-instruct). The top 17 most

likely completions of the provided text are 'vegetative electron microscopy', and these suggestions are 2.2 times more likely than the next most likely prediction. Credit: OpenAI
By comparing what we know about the training datasets of different models, we identified the [CommonCrawl](#) dataset of scraped internet pages as the most likely vector where AI models first learned this term.

The scale problem

Finding errors of this sort is not easy. Fixing them may be almost impossible.

One reason is scale. The CommonCrawl dataset, for example, is millions of gigabytes in size. For most researchers outside large tech companies, the computing resources required to work at this scale are inaccessible.

Another reason is a lack of transparency in commercial AI models. OpenAI and many other developers refuse to provide precise details about the training data for their models. Research efforts to reverse engineer some of these datasets have also been stymied by copyright takedowns.

When errors are found, there is no easy fix. Simple keyword filtering could deal with specific terms such as vegetative electron microscopy. However, it would also eliminate legitimate references (such as this article).

More fundamentally, the case raises an unsettling question. How many other nonsensical terms exist in AI systems, waiting to be discovered?

Implications for science and publishing

This "digital fossil" also raises important questions about knowledge integrity as AI-assisted research and writing become more common.

Publishers have responded inconsistently when notified of papers including vegetative electron microscopy. Some have retracted affected papers, while others defended them. Elsevier notably [attempted to justify the term's validity](#) before [eventually issuing a correction](#).

We do not yet know if other such quirks plague large language models, but it is highly likely. Either way, the use of AI systems has already created problems for the peer-review process.

For instance, observers have noted the rise of "[tortured phrases](#)" used to evade automated integrity software, such as "counterfeit consciousness" instead of "artificial intelligence". Additionally, [phrases](#) such as "I am an AI language model" have been found in other retracted papers.

Some automatic screening tools such as Problematic Paper Screener now flag vegetative electron [microscopy](#) as a warning sign of possible AI-generated content. However, such approaches can only address known errors, not undiscovered ones.

Living with digital fossils

The rise of AI creates opportunities for errors to become permanently embedded in our knowledge systems, through processes no single actor controls. This presents challenges for [tech companies](#), researchers, and publishers alike.

Tech companies must be more transparent about training data and methods. Researchers must find new ways to evaluate information in the face of AI-generated convincing nonsense. Scientific publishers must improve their peer review processes to spot both human and AI-generated errors.

Digital fossils reveal not just the [technical challenge](#) of monitoring massive datasets, but the fundamental challenge of maintaining reliable knowledge in systems where errors can become self-perpetuating.

Provided by [The Conversation](#)

APRIL 15, 2025

Machine learning method cuts fraud detection costs by generating accurate labels from imbalanced datasets

by [Florida Atlantic University](#)



Credit: Tara Winstead from Pexels

Fraud is widespread in the United States and increasingly driven by technology. For example, 93% of credit card fraud now involves remote account access, not physical theft. In 2023, fraud losses surpassed \$10 billion for the first time.

The financial toll is staggering: [credit card fraud](#) costs \$5 billion annually, affecting 60% of U.S. cardholders, while identity theft resulted in \$16.4 billion in losses in 2021. Medicare fraud costs \$60 billion each year, and government losses range from \$233 billion to \$521 billion annually, with improper payments totaling \$2.7 trillion since 2003.

Machine learning plays a critical role in fraud detection by identifying patterns and anomalies in real-time. It analyzes [large datasets](#) to spot normal behavior and flag significant deviations, such as unusual transactions or account access. However, fraud

detection is challenging because fraud cases are much rarer than normal ones, and the data is often messy or unlabeled.

To address these challenges, researchers from the College of Engineering and Computer Science at Florida Atlantic University have developed a novel method for generating binary class labels in highly imbalanced datasets, offering a promising solution for fraud detection in industries like health care and finance. This approach works without relying on labeled data, a key advantage in sectors where privacy concerns and the cost of labeling are significant obstacles.

The team tested their method on two real-world, large-scale datasets with severe class imbalance (less than 0.2%): European credit card transactions (more than 280,000 from September 2013) and Medicare Part D claims (more than 5 million from 2013 to 2019), both labeled as fraudulent or genuine. These datasets, with fraud cases far outnumbered by non-fraud cases, provide a real-world challenge ideal for testing fraud detection methods.

Results of the study, published in the [Journal of Big Data](#), show that this new labeling method effectively addresses the challenge of labeling severely imbalanced data in an unsupervised framework. Additionally, and unlike traditional methods, this approach evaluated the newly generated fraud and non-fraud labels directly without the need of relying on a supervised classifier.

"The use of [machine learning](#) in fraud detection brings many advantages," said Taghi Khoshgoftaar, Ph.D., senior author and Motorola Professor in the FAU Department of Electrical Engineering and Computer Science. "Machine learning algorithms can label data much faster than human annotation, significantly improving efficiency. Our method represents a major advancement in fraud detection, especially in highly imbalanced datasets.

"It reduces the workload by minimizing cases that require further inspection, which is crucial in sectors like Medicare and credit card fraud, where fast data processing is vital to prevent [financial losses](#) and enhance operational efficiency."

The study shows the new method outperformed the widely-used Isolation Forest algorithm, providing a more efficient way to identify fraud while minimizing the need for further investigation. This confirms the method's ability to generate reliable binary class labels for fraud detection, even in challenging datasets. It offers a scalable solution for detecting fraud without relying on costly and time-consuming labeled data, which requires significant manual expert input and is resource-intensive, especially for large datasets.

"Our method generates labels for both fraud or positive and non-fraud or negative instances, which are then refined to minimize the number of fraud labels," said Mary Anne Walauskis, first author and a Ph.D. candidate in the FAU Department of Electrical Engineering and Computer Science. "By applying our method, we minimize false positives, or in other words, genuine instances marked as fraud, which is key to improving fraud detection.

"This approach ensures that only the most confidently identified fraud cases are retained, enhancing accuracy and reducing unnecessary alarms, making [fraud detection](#) more efficient."

The method combines two strategies: an ensemble of three unsupervised learning techniques using the SciKit-learn library and a percentile-gradient approach. The goal is to minimize false positives by focusing on the most confidently identified fraud cases. This is achieved by refining the labels and reducing errors in both the unsupervised methods (EUM) and the percentile-gradient approach (PGM).

The refined labels create a subset of confident labels that are highly likely to be accurate. These labels are then used to create confidence intervals and finalize the labeling, requiring minimal domain knowledge to select the number of positive instances.

"This innovative approach holds great promise for industries plagued by fraud, offering a more accessible and effective way to identify fraudulent activity and safeguard both financial and health care systems," said Stella Batalama, Ph.D., dean of the College of Engineering and Computer Science.

"Fraud's impact goes beyond financial losses, including emotional distress, reputational damage and reduced trust in organizations. Health care fraud, in particular, undermines care quality and cost, while identity theft can cause severe stress. Addressing fraud is key to mitigating its broad societal impact."

Looking ahead, the research team plans to enhance the method by automating the determination of the optimal number of positive instances, further improving efficiency and scalability for large-scale applications.

The current journal article, "Unsupervised Label Generation for Severely Imbalanced Fraud Data," is an updated version of the researchers' [previous work](#), "Confident Labels: A Novel Approach to New Class Labeling and Evaluation on Highly Imbalanced Data."

The original paper was presented and published at the [IEEE 36th International Conference on Tools with Artificial Intelligence](#) (ICTAI) in November 2024, where it won the Best Student Paper Award. ICTAI, with an acceptance rate of about 25% from more than 400 submissions, is a prestigious conference.

More information: Mary Anne Walauskis et al, Unsupervised label generation for severely imbalanced fraud data, *Journal of Big Data* (2025). DOI: [10.1186/s40537-025-01120-x](https://doi.org/10.1186/s40537-025-01120-x)
Provided by [Florida Atlantic University](#)

APRIL 15, 2025

Explainable AI for ship navigation raises trust, decreases human error

by [Osaka Metropolitan University](#)



A passenger ferry, left, and a cargo ship sail the Akashi Strait in Japan. Credit: Yoshiho Ikeda, Professor Emeritus, Osaka Prefecture University

The Titanic sank 113 years ago on April 14–15, after hitting an iceberg, with human error likely causing the ship to stray into those dangerous waters. Today, autonomous systems built on artificial intelligence can help ships avoid such accidents, but could such a system explain to the captain why it was maneuvering a certain way?

That's the idea behind explainable AI, which should help human actors trust [autonomous systems](#) more.

Researchers from Osaka Metropolitan University's Graduate School of Engineering have developed an explainable AI model for ships that quantifies the collision risk for all vessels in a given area, an important feature as key sea-lanes have become ever more congested. The findings were [published](#) in *Applied Ocean Research*.

Graduate student Hitoshi Yoshioka and Professor Hirotada Hashimoto created the AI model so that it explains the basis for its decisions and the intention behind actions using numerical values for collision risk.

"By being able to explain the basis for the judgments and behavioral intentions of AI-based autonomous ship navigation, I think we can earn the trust of maritime workers," Professor Hashimoto stated. "I also believe that this research can contribute to the realization of unmanned ships."

More information: Hitoshi Yoshioka et al, Explainable AI for ship collision avoidance: Decoding decision-making processes and behavioral intentions, *Applied Ocean Research* (2025). DOI: [10.1016/j.apor.2025.104471](https://doi.org/10.1016/j.apor.2025.104471)
Provided by [Osaka Metropolitan University](#)

Agatha Christie brought to life by AI for writing course



12th January – Died on this Day – 1976© PA Archive

Aspiring writers are being offered vital clues on writing from [Agatha Christie](#), with a little help from artificial intelligence (AI).

The [BBC](#) and the late best-selling author's estate have teamed up to create Agatha Christie Writing, an online course using her own words delivered by the woman herself.

Brought to life by actress Vivien Keen and visual effect artists, the course will use AI-enhanced technology, images and restored audio recordings to recreate Christie's words and teachings to make it feel as if she is delivering them directly.

"As a lifelong fan of Agatha Christie, bringing this course to life has been a dream come true, and I am immensely proud of it," said Michael Levine, chief executive of BBC Maestro, which runs online courses and developed Agatha Christie Writing with BBC Studios, the writer's estate and a team of experts.

He said the course invites all Christie fans to "learn through her own words, exactly how she does it; her background, her inspirations, her craft and the lessons she learned along the way".

The course has been created by a team of experts from Christie's who have reconstructed her philosophy on writing, drawing her insights on story structure, cast creation, plot twists, red herrings and the art of suspense from her work and archival interviews.

The writer's great-grandson James Prichard, the chairman and chief executive of Agatha Christie Limited, said: "The team of academics and researchers that BBC Maestro has assembled have extracted from a number of her writings an extraordinary array of her views and opinions on how to write."

Christie's books have sold more than 200 billion copies and include 66 detective novels and 14 short story collections. Her works also include *The Mousetrap*, the world's longest-running stage play which has run in London's [West End](#) since 1952 with its only break coming during the Covid pandemic.

She died in 1976 aged 85 but her crime fiction continues to be produced for television and movies.

AI Is Powering the Way You Shop, Even When You Don't Realize It

Every time you scroll through an online store, [AI is quietly working behind the scenes](#), shaping your experience. From personalized product recommendations to chatbots answering your questions, artificial intelligence is making shopping faster, easier, and surprisingly intuitive. The moment you search for a pair of sneakers, AI algorithms start analyzing your browsing history, past purchases, and even what's trending in your area. That's why you see exactly what you're looking for—or things you didn't know you needed—right at the top of your feed.

Beyond recommendations, AI is also preventing fraud and ensuring safe transactions. Banks and payment processors use machine learning to detect suspicious activities in real time, stopping potential cyber threats before they can do damage. Meanwhile, AI-powered inventory management helps stores keep popular items in stock, reducing delivery times and improving customer satisfaction. So whether you're scrolling through an app or walking into a store, AI is shaping your shopping experience in ways that feel seamless but are actually incredibly sophisticated.

Your Smartphone Is an AI-Powered Assistant in Your Pocket

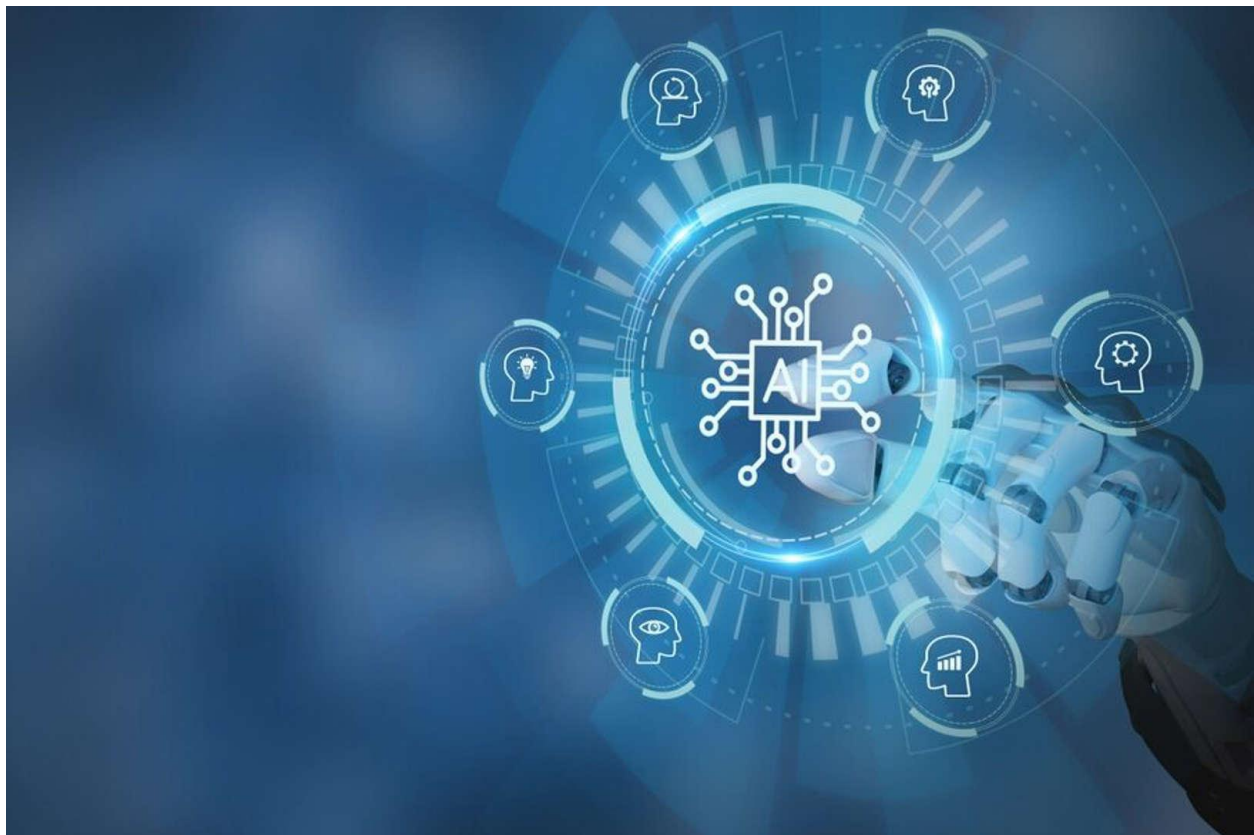
You don't need a fancy robot to experience AI—it's already sitting in your hand. Your smartphone is packed with [AI-driven features](#), from voice assistants like Siri and Google Assistant to predictive text that finishes your sentences before you do. Every time you use facial recognition to unlock your phone, AI is scanning and analyzing millions of data points in milliseconds to ensure it's really you. Even the way your phone organizes photos, suggesting albums or recognizing faces, is all thanks to deep learning algorithms working in the background.

But AI doesn't just make your phone more convenient—it makes it smarter over time. When you dictate a message, your phone learns your speech patterns and improves its accuracy. When you search for a place on maps, AI factors in real-time traffic data to find the best route. And if your battery is running low, AI manages power consumption by optimizing which apps stay active. Whether you realize it or not, AI is constantly learning from your habits to make your smartphone experience faster, smoother, and more personalized.

AI Is Making Healthcare More Accurate and Accessible

Medical professionals are using AI to detect diseases earlier, improve diagnoses, and personalize treatments like never before. [Advanced AI systems can scan X-rays](#) and MRIs with incredible accuracy, sometimes even detecting early signs of conditions like cancer that human doctors might miss. These tools don't replace medical professionals, but they do give them a second set of highly trained "eyes," helping them make more informed decisions faster. In some cases, AI is even predicting potential health risks before symptoms appear, allowing for preventive care that could save lives.

Beyond hospitals, AI is making healthcare more accessible. Virtual health assistants can answer medical questions, recommend treatments, and even monitor chronic conditions remotely. AI-powered apps help people track their heart rate, sleep patterns, and exercise habits, giving them insights into their health without a trip to the doctor's office. In developing regions where medical professionals are scarce, AI is bridging the gap, providing essential health information and early screenings that wouldn't otherwise be available. The combination of technology and medicine is changing the way we approach healthcare, making it more efficient, proactive, and personalized.



AI Is Revolutionizing the Way We Work and Learn

From writing emails to scheduling meetings, AI is streamlining everyday tasks in the workplace. Many businesses now use AI-driven tools to automate repetitive jobs, freeing up employees to focus on more strategic work. Virtual assistants can summarize long documents, generate reports, and even draft emails based on past conversations. And with remote work on the rise, AI-powered collaboration tools are making communication across different time zones smoother, keeping teams connected and productive.

Education is also undergoing an [AI-powered transformation](#). Personalized learning platforms analyze how students absorb information and tailor lessons to match their pace and understanding. AI tutors can explain difficult concepts, provide instant feedback, and even predict when a student might be struggling before they fall behind. In universities, AI is helping researchers analyze vast amounts of data in ways that would take humans years to process. Whether in classrooms or offices, AI is not replacing people—it's giving them smarter tools to do their work more efficiently.

AI Is Enhancing Entertainment in Unexpected Ways

©iStock/ Blue Planet Studio

Your favorite movies, music, and even video games are being shaped by AI. Streaming services analyze what you watch or listen to, then recommend content tailored specifically to your tastes. AI even helps studios create more engaging films by analyzing audience preferences and predicting which plotlines or themes will be most successful. In gaming, AI opponents have become so sophisticated that they can adapt to a player's skill level, making each experience unique and challenging.

Music is another area where AI is making waves. Algorithms can now compose original songs, mimic the styles of famous artists, and even help musicians fine-tune their work. In social media, AI is curating what content appears in your feed, ensuring that what you see is personalized to your interests. While this makes entertainment more engaging, it also raises questions about how much AI influences our choices. Whether it's Netflix recommending your next

binge-watch or an AI-generated song topping the charts, artificial intelligence is quietly reshaping the way we consume media.

AI Is Transforming Transportation, and It's Just Getting Started

Self-driving cars might not be everywhere yet, but AI is already transforming transportation in ways you may not notice. Ride-sharing apps use AI to match passengers with drivers, predict demand, and optimize routes for efficiency. Airlines rely on AI to schedule flights, prevent delays, and even assist pilots with navigation. Even traffic lights in some cities are controlled by AI, adjusting in real time to improve traffic flow and reduce congestion.

But the biggest AI-driven shift in transportation is still on the horizon. Autonomous vehicles are being tested worldwide, using advanced machine learning to navigate roads, detect obstacles, and make split-second driving decisions. While we're not yet at the point of fully self-driving highways, AI-assisted safety features—like automatic braking and lane departure warnings—are already making cars smarter and reducing accidents. As technology continues to improve, AI's role in transportation will only grow, making travel safer, faster, and more efficient for everyone.



AI Is Changing How We Communicate and Connect

AI is now a key part of how we interact with each other, even if we don't always notice it. Messaging apps use AI to predict what you're going to type next, social media platforms use it to filter spam and hate speech, and translation tools are making it easier than ever to communicate across languages. AI is also driving the rise of deepfake technology, which can create eerily realistic videos of people saying things they never actually said—raising both exciting possibilities and serious ethical concerns.

On a more personal level, AI is changing how we build and maintain relationships. Dating apps use AI to analyze user behavior and make better matches, while virtual assistants help remind us of important events like birthdays and anniversaries. Even customer service is being transformed, with AI-powered chatbots handling inquiries and resolving issues faster than human agents. As AI continues to evolve, the way we communicate, interact, and even form relationships will keep shifting in ways we're only beginning to understand.

- [t news](#)
- [Unread news](#)
- [Subscribe](#)

Science X Account

Remember me

Sign In
Sign in with _

[Forget Password?](#)

[Not a member? Sign up](#)
[Learn more](#)

- [108](#)
- [Tweet](#)
- [Share](#)
-

[Email](#)

1. [Home](#)
2. [Computer Sciences](#)
1. [Home](#)
2. [Machine learning & AI](#)
-
-
-

APRIL 24, 2025

Algorithm can make AI responses increasingly reliable with less computational overhead

by Daniel Meierhans , [ETH Zurich](#)

A new algorithm from ETH researchers improves large language models (LLMs) so that the selected answers are more accurate and relevant. Credit: AI generated/ ETH Zurich

ChatGPT and alike often amaze us with the accuracy of their answers, but unfortunately, they also repeatedly give us cause for doubt. The main issue with powerful AI response engines (artificial intelligence) is that they provide us with perfect answers and obvious nonsense with the same ease. One of the major challenges lies in how the large language models (LLMs) underlying AI deal with uncertainty.

Until now, it has been very difficult to assess whether LLMs designed for text processing and generation base their responses on a solid foundation of data or whether they are operating on uncertain ground.

Researchers at the Institute for Machine Learning at the Department of Computer Science at ETH Zurich have now developed a method that can be used to specifically reduce the uncertainty of AI. The work is [published](#) on the *arXiv* preprint server.

"Our algorithm can enrich the general language model of the AI with additional data from the relevant subject area of a question. In combination with the specific question, we can then extract from the depths of the model and from the enrichment data precisely those connections that are most likely to generate a correct answer," explains Jonas Hübötter from the Learning & Adaptive Systems Group, who developed the new method as part of his Ph.D. studies.

Enriching AI with specific data

"The method is particularly suitable for companies, scientists or other users who want to use general AI in a specialized field that is only covered partially or not at all by the AI

training data," adds Andreas Krause, head of the research group and Director of the ETH AI Center.

For example, users can feed their locally stored data into a large language model (LLM), such as Llama. The so-called SIFT algorithm (Selecting Informative data for Fine-Tuning), developed by ETH computer scientists, can then use the additional data provided to select specific information that is most closely related to the question.

Relationship vectors in multidimensional space

The algorithm uses the structure according to which the language information is organized in the AI's large language model (LLM) to find related information. The models divide the language information in their training data into word parts.

The semantic and syntactic relationships between the word parts are then arranged as connecting arrows—known in the field as vectors—in a multidimensional space. The dimensions of space, which can number in the thousands, arise from the relationship parameters that the LLM independently identifies during training using the general data.

Angle between arrows as measure of correlation

Relational arrows pointing in the same direction in this vector space indicate a strong correlation. The larger the angle between two vectors, the less two units of information relate to one another.

The SIFT algorithm developed by ETH researchers now uses the direction of the relationship vector of the input query (prompt) to identify those information relationships that are closely related to the question but at the same time complement each other in terms of content.

"The angle between the vectors corresponds to the relevance of the content, and we can use the angles to select specific data that reduces uncertainty," explains Hübötter.

Less overlap from redundant information

By contrast, the most common method used to date for selecting the information suitable for the answer, known as the nearest neighbor method, tends to accumulate redundant information that is widely available. The difference between the two methods becomes clear when looking at an example of a query prompt that is composed of several pieces of information.

To answer the two-part question "How old is Roger Federer and how many children does he have?" the nearest neighbor method considers similar information such as "Roger Federer is 43 years old" and "Roger Federer's birthday is 8 August 1981" to be equally relevant.

Information about his children, which is relevant for the second part of the question, is sometimes missing. It is overlaid by birth date information, which occurs much more frequently in the AI [training data](#).

The SIFT algorithm, however, takes into account the extent to which the pieces of information included complement each other, i.e. whether the information vectors point in different directions. This allows relevant information to be identified for both aspects of the question.

More reliable answers with much smaller models

However, targeted information selection not only improves the quality of responses. It can also be used to reduce the ever-increasing computing power required by AI applications.

By indirectly measuring uncertainty, the model can decide for itself how much more data is needed to provide a sufficiently reliable answer. Consequently, the computational overhead required by an LLM can be systematically adapted to the complexity of the question and the availability of relevant information.

Since SIFT continuously adapts the weighting of the arrow directions to its calculations during data retrieval, the enriched model becomes increasingly reliable the more it is used. This is known as test-time training and can be used to achieve the same output performance with smaller models.

"In tests with standard data sets, we used SIFT tuning to outperform even the best current AI models with models up to 40 times smaller," emphasizes Hübötter.

Identifying added value of relevant data

Additional applications for the SIFT algorithm are opening up in terms of data evaluation. As Krause explains, "We can track which enrichment data SIFT selects. They are closely related to the question and therefore particularly relevant to this subject area. This could be used in medicine, for example, to investigate which laboratory analyses or measurement values are significant for a specific diagnosis and which are less so."

Hübötter is presenting his approach at the International Conference on Learning Representations (ICLR) in Singapore. In December, the ETH researchers won the prize for the Best Scientific Article for their method at the NeurIPS Annual Conference on Neural Information Processing Systems (NeurIPS) in the "Finetuning in Modern Machine Learning" workshop.

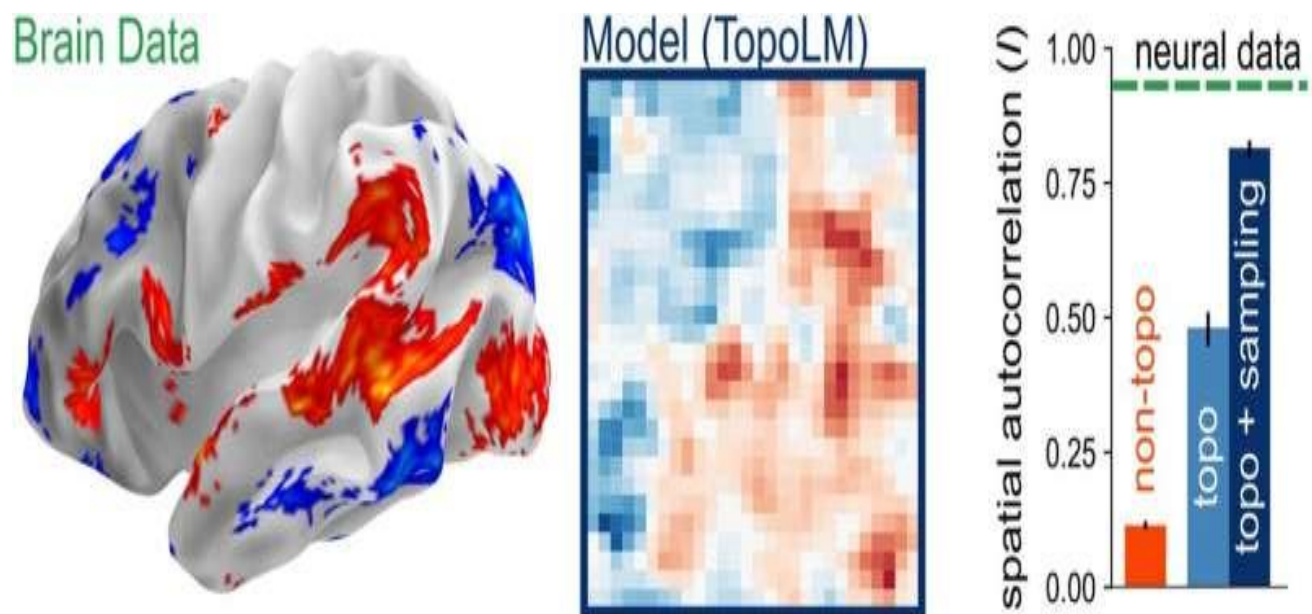
More information: Jonas Hübötter et al, Efficiently Learning at Test-Time: Active Fine-Tuning of LLMs, *arXiv* (2024). DOI: [10.48550/arxiv.2410.08020](https://doi.org/10.48550/arxiv.2410.08020)

Journal information: [arXiv](#)

APRIL 24, 2025

AI model of language in the brain captures both how neurons are arranged and how they function

by Tanya Petersen , [Ecole Polytechnique Federale de Lausanne](#)



Credit: Ecole Polytechnique Federale de Lausanne

Neurons—the nerve cells responsible for transmitting electrical and chemical signals throughout the body—are organized in tissue, tending to cluster together in groups according to how they function. For example, language studies have found that there are clusters of neurons that seem to specialize in verbs and others that focus specifically on nouns, however researchers aren't sure just how these functional groups form.

Previous AI [language](#) models have been able to successfully capture these individual clusters of functional neurons but haven't looked at how these are spatially arranged in the brain.

Now, researchers from the NeuroAI Laboratory, part of both the School of Computer and Communication Sciences (IC) and the School of Life Sciences (SV), have created TopoLM, a new AI language model that mimics both the functional clustering of neurons and, for the first time, how these [nerve cells](#) are spatially arranged within the brain.

"Building on work around how the brain processes vision, we made some relatively small changes to how a language model organizes itself internally, adding a rule that encourages the model's internal representations to be 'smooth' spatially.

"The resulting model, TopoLM, develops spatial clusters of its internal components that functionally match the activity we see in human brains when they process language," explained Assistant Professor Martin Schrimpf, Head of the NeuroAI Lab.

In their [paper](#), "TopoLM: Brain-Like Spatio-Functional Organization in a Topographic Language Model," one of the less than 2% of papers selected to be presented orally this week at [ICLR 2025](#), the International Conference on Learning Representations, the researchers outline how TopoLM successfully predicted how the language system in the brain's outer layer, the cortex, develops its spatio-functional organization. The paper is also [available](#) on the *arXiv* preprint server.

"This was basic research to understand how spatial clusters or functional clusters in the brain originate in the first place and what this new model suggests is that it might be driven by a single, basic rule about spatial organization where close by neurons just tend to behave similarly," continued Schrimpf.

The researchers believe TopoLM provides a framework for improving AI's functional alignment with [human cognition](#), offering direct applications to developing brain-inspired computing and neurolinguistics.

"This is an exciting step towards building artificial intelligence systems that are organized more like the human brain. One of our main goals is to build better models of the brain in general and with TopoLM we are one step closer to clinical applications that might help us to help people with language disorders or similar language deficits," said Badr AlKhamissi, a doctoral assistant in the NeuroAI Lab and IC's Natural Language Processing Laboratory (NLP) and one of the authors of the paper.

This work also sheds light on interpretability, basically the understanding of what happens inside complex Large Language Models (LLMs) and how they actually work. Normally, LLMs have mathematical representations called "vectors" associated with each artificial "neuron." To understand what the model has learned, it would be necessary to analyze each vector individually or in small groups.

Because TopoLM organizes its internal components into clusters, it's possible to look at these and see meaningful groupings emerge. This makes it more straightforward to get a sense of how the model is representing and processing language because the organization itself reflects meaningful categories.

So what's next in this cutting-edge research? The EPFL research team will now work to test the predictions of the model on human brains.

"This model works so well, far better than we expected, that we want to test if its predictions are true. There are clusters in the model that we have not yet observed in the human brain because nobody has looked for them yet. We'll be working with colleagues in the United

States who experimentally work with humans to run a new imaging study to find out if these clusters also appear in the [human brain](#)," concluded Schrimpf.

More information: Neil Rathi et al, TopoLM: brain-like spatio-functional organization in a topographic language model, *arXiv* (2024). DOI: [10.48550/arxiv.2410.11516](https://doi.org/10.48550/arxiv.2410.11516)

Journal information: [arXiv](#)

APRIL 24, 2025

Developments in AI need to be properly regulated as the world scrambles for advantage

by Nisreen Ameen , [The Conversation](#)

Artificial intelligence (AI) is often hailed as the defining technology of the 21st century, shaping everything from economic growth to national security. But as global investment in AI accelerates, many experts are beginning to ask whether the world has embarked on an AI "arms race."

With China, the US, UK and the European Union each [pledging billions](#) to advance AI, competition in research, infrastructure and industrial applications for the new technology is [intensifying](#). But, at the same time, regulation is struggling to keep pace with rapid development in some regions. This is raising concerns about ethical risks, [economic inequality](#) and global AI governance.

There have been rapid advances in AI in the past few years. Companies such as America's [Accenture](#) and China's [DeepSeek](#) have developed large-scale generative AI systems—which can learn from existing content to generate new material such as text, images, music, or videos.

The UK government recently announced its intention to "shape the AI revolution rather than wait to see how it shapes us" through its [AI Opportunities Action Plan](#). This will have a strong focus on regulation, skills and ethical governance.

If the UK and continental Europe are [prioritizing regulation](#), China is using its sheer size and appetite for innovation to develop rapidly into what has been described as an "[AI super market](#)," and the US is [balancing innovation with national security concerns](#).

China recently released [details of new regulations](#), which come into force in September, that will require explicit labeling of AI-generated content and providing metadata to link such

content to the service provider that generated it. The onus will be on platforms that feature AI-generated content to provide such information.

But the different approaches highlight the growing geopolitical dimension of AI development which risks divergence of standards. While competition can drive innovation, without international cooperation on safety, ethics and governance, the global AI race could lead to regulatory gaps and fragmented oversight.

Many analysts fear this would bring significant downsides. Most worryingly, there is the prospect of unchecked [AI-generated disinformation](#) undermining elections and democratic institutions.

Why does this matter?

AI is more than just another technological breakthrough—it's a strategic driver of [economic power](#) and [influence](#). The countries leading in AI today will play an important role in shaping the future of automation, digital economies and international regulatory frameworks.

AI's global expansion is driven by several key motivations. It has the potential to massively boost [productivity and creativity](#). It can create new business models and transform entire industries. Governments investing in AI aim to secure long-term economic advantages, particularly in sectors such as finance, health care and advanced manufacturing.

Meanwhile, AI is increasingly integrated into [defense, cybersecurity and intelligence](#). Governments are exploring ways to use AI for strategic advantage, while also ensuring resilience against AI-enabled threats.

But as AI investment surges, it is increasingly important to ensure that the challenges the new technology will bring are not overlooked in the rush.

Risks of rapid AI investment

As AI advances, ethical issues become more pressing. AI-powered surveillance systems raise privacy concerns. [Deepfake technology](#), meanwhile, which is capable of generating hyper-realistic video and audio, is already being used for disinformation. Without clear regulatory oversight, this could seriously undermine trust and security and threaten democratic institutions.

At the same time, we are already seeing inequality baked into AI development. Many AI-driven innovations cater to wealthy markets and corporations. Meanwhile, marginalized communities face barriers to accessing AI-enhanced education, health care and job opportunities—the latter was demonstrated as long ago as 2018 when Amazon [reportedly withdrew a recruitment tool](#) that was shown to discriminate against women.

Ensuring that AI development benefits society as a whole will require a strategic approach to skills, education and governance. I have conducted studies into how AI tools are being harnessed with a great deal of success in the [UK and US](#) and also in [China](#). The research

showed how AI capabilities can be combined with strategic agility to drive product and service innovation in many contexts.

But the AI race is not just about economic progress, it also has geopolitical implications. Restrictions on [AI-related exports](#), particularly in semiconductor technology, highlight growing concerns over technological dependencies and national security. Without greater international cooperation, uncoordinated AI policies could lead to economic fragmentation, regulatory inconsistencies across borders and the inevitable risks those bring.

Although some nations are advocating for [global AI agreements](#), these discussions remain in their early stages, so enforcement mechanisms remain limited.

The way forward

This will require multilateral governance, similar to global frameworks on cybersecurity and climate change. Existing discussions by the [United Nations](#) as well as the [G7](#) and the [Organization for Economic Cooperation and Development \(OECD\)](#) need to incorporate stronger AI-specific enforcement mechanisms that guide development responsibly.

There are signs of progress. The G7's [Hiroshima AI Process](#) has resulted in shared guiding principles and a voluntary code of conduct for advanced AI systems. The [OECD's AI Policy Observatory](#), meanwhile, is helping coordinate best practices across member states. But binding international enforcement mechanisms are still in their infancy.

Individual countries, meanwhile, need to develop flexible regulatory frameworks that balance innovation with accountability. The EU's [AI Act](#), the first major attempt to comprehensively regulate AI, classifies AI systems by risk and imposes obligations on developers accordingly.

This has included bans on certain high-risk applications, such as [social scoring](#)—which ranks individuals based on behavior and can lead to discrimination. It's a step in the right direction, but broader cooperation is still needed to ensure coherent global AI standards.

An enforceable set of rules governing AI development is needed—and quickly. AI could pose more risks than opportunities if left unchecked.

Provided by [The Conversation](#)

APRIL 28, 2025

Transforming aircraft maintenance with augmented reality

by [Georgia Institute of Technology](#)



Tablet in Use for
Repair Work

RepAR aircraft maintenance system being used with a tablet. The system uses a combination of augmented reality, computer vision, and artificial intelligence. Credit: Georgia Institute of Technology

In the dynamic world of aviation maintenance, precision, quickness, and meticulous documentation are essential. Maribeth Gandy Coleman, director of research and a Regents' Researcher in Georgia Tech's Institute for People and Technology (IPaT), has been leading an IPaT translational research team working to advance aircraft maintenance with PartWorks, an Atlanta-based aerospace engineering firm dedicated to extending the life and improving the operational efficiency and availability of commercial and military aircraft and spacecraft.

Coleman, a recognized augmented reality expert at Georgia Tech, has been working with the PartWorks' engineering team to solve aircraft maintenance challenges, leading to measurable improvements in labor costs, training, repair quality, turnaround time, and maintenance process validation. This research partnership has led to the development of several patented and patent-pending solutions related to aircraft maintenance.

"I could not have hired anybody with the diverse skill sets that both Maribeth and the Georgia Tech team brought to bear," said Scott Geller, CEO of PartWorks. "We've utilized different and complicated skill sets, sometimes in small quantities, that have made our project work very cost-effective. We've used an iterative research and development process that hasn't had a shocking cost or huge surprises. And the Georgia Tech team has been both easy and fun to work with, too."

This collaboration has led to PartWorks launching a new aircraft maintenance, repair, and overhaul (MRO) augmented reality solution called [RepAR](#). Designed for both military and commercial aviation, RepAR's augmented reality overlay transforms structural repairs by ensuring accuracy, reducing labor costs, minimizing [human error](#), and accelerating return-to-service timelines.

RepAR rapidly captures structural repair data, embedding spatial awareness and real-time validation into maintenance workflows. Novice technicians can achieve results beyond their operational experience, while seasoned technicians experience measurable productivity gains.

"RepAR exemplifies how targeted computer vision applications can deliver immediate value in aerospace manufacturing and maintenance," said Shelley Peterson, CEO of Wizard Wells. "By precisely identifying fastener locations and validating tool placement, it reduces rework, minimizes human error, and ensures tasks are performed right the first time."

PartWorks demonstrated RepAR at the Aviation Week Network's MRO Americas, which took place April 8–10 at the Georgia World Congress Center in Atlanta.

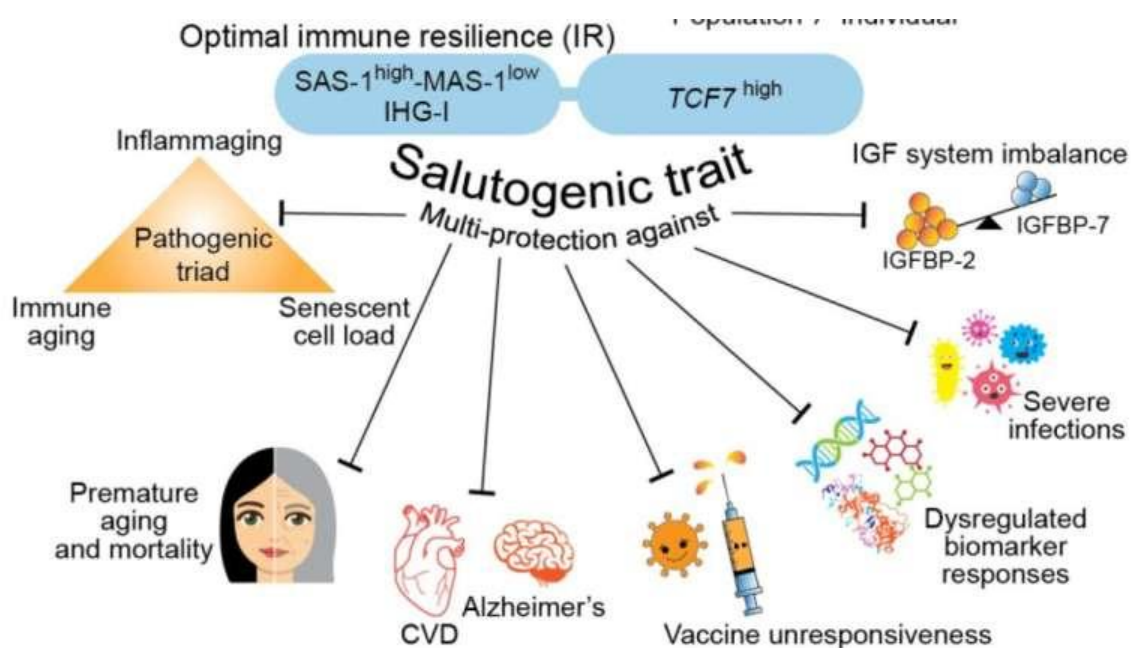
"This [collaborative research](#) with industry demonstrates why Georgia Tech has interdisciplinary research institutes such as IPaT, and why you have research faculty," said Coleman. "You're probably not going to be able to get some Ph.D. students to do this work. The focus here with PartWorks is on translation. It's cross-disciplinary collaboration and translation built on augmented reality work we've been doing for 25 years and implementing cutting-edge technology crafted to the right context to support aircraft [maintenance](#)."

"This Georgia Tech collaboration and augmented reality MRO research and development are in conjunction with a multiyear contract we're working on with the Air Force Research Lab (AFRL) in Dayton, Ohio," said Geller. "We're appreciative of their partnership and excited to be getting commercial interest in RepAR from both military and [commercial aviation](#) OEMs and MROs as well as space industry companies."

Provided by [Georgia Institute of Technology](#)

Aging reimaged: Study finds immune resilience counters key drivers of disease, mortality

by Claire Kowalick , [University of Texas Health Science Center at San Antonio](#)



Salutogenic traits, like immune resilience, provide protection against pathogenic factors.

Credit: UT Health San Antonio

What if, rather than focusing solely on the drivers of disease (pathogenesis), we also explored the factors sustaining our health—the concept of salutogenesis (derived from Salus (health) in Roman mythology)?

The idea of salutogenesis through immune resilience, which refers to mechanisms by which the body counters disease drivers, could play a role in [disease prevention](#) and may influence lifespan. This concept is explored in a recent study led by Sunil K. Ahuja, MD, professor in the department of medicine at the Joe R. and Teresa Lozano Long School of Medicine at The University of Texas Health Science Center at San Antonio (UT Health San Antonio) and director of the Veterans Affairs Center for Personalized Medicine, South Texas Veterans Health Care System, San Antonio.

Since the dawn of civilization, humans have encountered inflammatory challenges, such as infections. Appropriately regulated inflammation is essential for protection against threats and supporting survival. However, inflammation is also linked to many diseases and mortality. This may have contributed to the emergence of health-promoting salutogenic forces, including immune resilience, which could help mitigate the harmful effects of inflammation.

Ahuja and his team's [multi-year, multi-disciplinary study](#) published April 23, 2025, in *Aging Cell*, investigates this biological mechanism and its potential implications for healthspan and longevity.

The study, "The 15-Year Survival Advantage: Immune Resilience as a Salutogenic Force in Healthy Aging," reveals that immune resilience is not fixed, but measurable and adaptable. The team identified the expression of T-cell factor 7 (TCF7) as a key component of immune resilience. Higher TCF7 levels and related markers correlated with better health outcomes, improved resistance to disease and durable vaccine responses.

The "pathogenic triad": A balancing act

"Our work shows that immune resilience is associated with TCF7, a central master regulator that maintains T-cell health," said Muthu Manoharan, MS, co-first author and senior research scientist at UT Health San Antonio.

The study proposes that TCF7-linked immune resilience may counterbalance what the authors term as a "pathogenic triad"—inflammaging (chronic inflammation with aging), immune aging and cell senescence. Immune resilience can diminish over time due to environmental threats (e.g., infections, trauma) or internal stressors (e.g., myocardial ischemia). Analogous to a dam protecting a city from floods, immune resilience may act as a biological barrier, reducing the risk of the triad overwhelming the body.

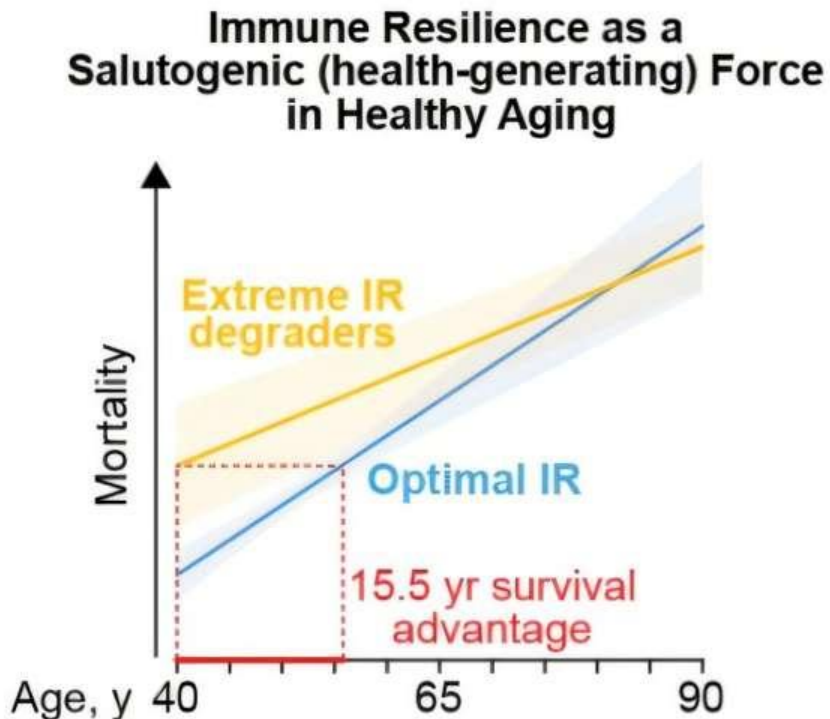
"When salutogenesis declines and pathogenesis emerges, this may create a state of inflammation and immune aging that promotes disease," Ahuja explained. "Individuals with TCF7-linked immune resilience appear better equipped to resist inflammatory stressors and maintain a low-inflammatory immune profile promoting survival and better health."

Three immune resilience trajectories

By analyzing data from over 17,500 individuals, the team identified three immune resilience paths during inflammatory stress:

- Immune resilience preservers: Maintain high resilience, potentially countering the triad.
- Immune resilience reconstitutors: Experience temporary resilience declines with restoration during recovery.
- Immune resilience degraders: Show persistent resilience loss and sustained increases in the burden of the pathogenic triad.

"Throughout life, inflammatory insults like infections may gradually reduce immune resilience. Some individuals can preserve resilience longer than others," said Grace Lee, PharmD, Ph.D., co-first author and associate professor at The University of Texas at Austin.



People with the highest immune resilience show an up to 15.5-year survival advantage over those with the lowest immune resilience. Credit: UT Health San Antonio

Aging and immune resilience's 'warranty period'

The study highlights three lifelong processes affecting health: salutogenesis, pathogenesis and senescence (aging). Declining immune resilience, potentially exacerbated by the triad, may raise susceptibility to age-related illnesses.

"Disease vulnerability differs from a vulnerability to lose health, which allows disease processes to emerge. Aging and age-related diseases are distinct," noted Nathan Harper, MS, senior bioinformatician in the Veterans Health Center of Personalized Medicine.

However, even robust immune resilience has limits. Around age 70, the protective effects begin to wane—a threshold Manoharan terms "failed salutogenesis."

Window of opportunity for potential intervention

The study found that participants in a large U.S. population with strong immune resilience at age 40 had a 15.5-year survival advantage over those with low resilience. By age 70, longevity trajectories converged, eliminating this advantage.

The findings imply immune resilience might be modifiable from birth to approximately 70 years. Lifestyle changes, medications or future immunotherapies could delay age-related diseases and extend healthspan.

"There's a biologically modifiable window of opportunity for interventions like diet, exercise or drugs. Beyond this window, improving healthspan becomes more challenging," Lee said.

Toward proactive health management

"We envision a future in which immune resilience is routinely assessed, much like cholesterol testing," said Justin Meunier, BS, a bioinformatician at the Center for Personalized Medicine. "Optimal immune resilience is associated with a unique blood biomarker profile that reflects higher levels of growth and immune factors, along with lower levels of inflammation."

Lee added that personalized plans based on immune resilience status could help prevent disease proactively.

A new perspective on aging

TCF7-linked immune resilience represents a novel, modifiable trait in aging research. Unlike traditional biomarkers tied to disease and inflammation, this immune program may enable personalized salutogenic interventions.

Ahuja's team work shifts the focus from inevitable decline to optimizable resilience. Future strategies to recalibrate immune [resilience](#) could transition health care from disease treatment to sustaining health.

More information: Muthu Saravanan Manoharan et al, The 15-Year Survival Advantage: Immune Resilience as a Salutogenic Force in Healthy Aging, *Aging Cell* (2025). DOI: [10.1111/accel.70063](https://doi.org/10.1111/accel.70063)

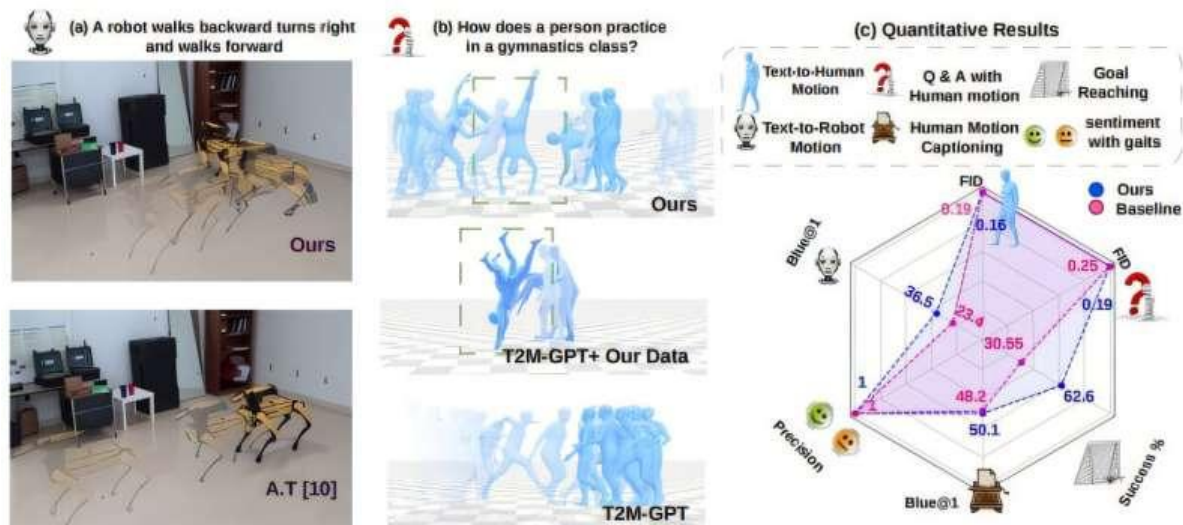
Journal information: [Aging Cell](#)

Provided by [University of Texas Health Science Center at San Antonio](#)

MAY 8, 2025

AI model translates text commands into motion for diverse robots and avatars

by Kevin Stacey, [Brown University](#)



MotionGlot is a model that can generate motion trajectories that obey user instructions across multiple embodiments with different action dimensions, such as (a) quadruped robots, and (b) humans. The figures (a,b) depict the qualitative benchmark of MotionGlot against the adapted templates (A.T) of [1] on the text-to-robot motion (Section IV-A.1), Q&A with human motion (Section IV-C) tasks respectively. The overall quantitative performance across tasks is shown in (c). In (a,b), increasing opacity indicates forward time.

Credit: *arXiv* (2024). DOI: 10.48550/arxiv.2410.16623

Brown University researchers have developed an artificial intelligence model that can generate movement in robots and animated figures in much the same way that AI models like ChatGPT generate text.

A paper describing this work is [published](#) on the *arXiv* preprint server.

The model, called MotionGlot, enables users to simply type an action—"walk forward a few steps and take a right"—and the model can generate accurate representations of that motion to command a [robot](#) or animated avatar.

The model's key advance, according to the researchers, is its ability to "translate" motion across robot and figure types, from humanoids to quadrupeds and beyond. That enables the generation of motion for a wide range of robotic embodiments and in all kinds of spatial configurations and contexts.

"We're treating motion as simply another language," said Sudarshan Harithas, a Ph.D. student in computer science at Brown, who led the work. "And just as we can translate

languages—from English to Chinese, for example—we can now translate language-based commands to corresponding actions across multiple embodiments. That enables a broad set of new applications."

The research will be presented later this month at the [2025 International Conference on Robotics and Automation](#) in Atlanta. The work was co-authored by Harithas and his advisor, Srinath Sridhar, an assistant professor of computer science at Brown.

Large language models like ChatGPT generate text through a process called "next token prediction," which breaks language down into a series of tokens, or small chunks, like individual words or characters. Given a single token or a string of tokens, the language model makes a prediction about what the next token might be.

These models have been incredibly successful in generating text, and researchers have begun using similar approaches for motion. The idea is to break down the components of motion—the discrete position of legs during the process of walking, for example—into tokens. Once the motion is tokenized, fluid movements can be generated through next token prediction.

One challenge with this approach is that motions for one [body type](#) can look very different for another. For example, when a person is walking a dog down the street, the person and the dog are both doing something called "walking," but their actual motions are very different. One is upright on two legs; the other is on all fours.

According to Harithas, MotionGlue can translate the meaning of walking from one embodiment to another. So a user commanding a figure to "walk forward in a [straight line](#)" will get the correct motion output whether they happen to be commanding a humanoid figure or a robot dog.

To train their model, the researchers used two datasets, each containing hours of annotated motion data. QUAD-LOCO features dog-like quadruped robots performing a variety of actions along with rich text describing those movements. A similar dataset called QUES-CAP contains real human movement, along with detailed captions and annotations appropriate to each movement.

Using that training data, the model reliably generates appropriate actions from text prompts, even actions it has never specifically seen before. In testing, the model was able to recreate specific instructions, like "a robot walks backwards, turns left and walks forward," as well as more abstract prompts like "a robot walks happily."

It can even use [motion](#) to answer questions. When asked, "Can you show me movement in cardio activity?" the model generates a person jogging.

"These models work best when they're trained on lots and lots of data," Sridhar said. "If we could collect large-scale data, the model can be easily scaled up."

The model's current functionality and the adaptability across embodiments make for promising applications in human-robot collaboration, gaming and [virtual reality](#), and digital

Journal information: [arXiv](#)
 Provided by [Brown University](#)

edited by [Lisa Lock](#), reviewed by [Robert Egan](#)



In a study [published](#) in *Advanced Science*, researchers explored how to make AI models more efficient and effective in predicting the properties of molecules for everything from drug development to materials design. The team focused on a technique called knowledge distillation, which involves compressing large and complex neural networks into smaller, faster models.

The distilled models ran faster—and in some cases improved performance—while working well across different experimental datasets, making them ideal for molecular screening without the heavy computational power required by most AI systems.

"To accelerate discovery in materials science, we need AI systems that are not just powerful, but scientifically grounded," said Fengqi You, the Roxanne E. and Michael J. Zak Professor in Energy Systems Engineering in Cornell Engineering, who co-authored the study with graduate student Rahul Sheshanarayana.

"Our work shows that AI can learn to reason across chemical and structural domains, generate realistic materials, and model molecular behaviors with efficiency and precision—all while aligning closely with the fundamental principles of materials science."

You directs the Cornell AI for Sustainability Initiative and co-directs the Cornell University AI for Science Institute—two programs advancing the next generation of AI-powered science. Both have supported forward-thinking efforts in You's research group.

In a *Nature Computational Science* [paper](#), You and Zhilong Wang, an Eric and Wendy Schmidt AI in Science Postdoctoral Fellow, introduce a new framework for generative inverse design of crystalline materials. Crystals, with their repeating atomic patterns and strict symmetry, present a challenge for AI models, which often rely on abstract or oversimplified representations.

The research group's proposed solution: a physics-informed generative AI model, which embeds crystallographic symmetry, periodicity, invertibility and permutation invariance directly into the model's learning process. The framework allows AI to generate novel crystal structures that are not only mathematically possible, but chemically realistic.

"Our goal is to ensure that AI-generated materials are scientifically meaningful," Wang said. "We're encoding physical principles and operating conditions directly into the learning framework, so instead of relying on massive trial-and-error, we're guiding the AI with domain knowledge."

In a [review paper](#) published in *Advanced Materials*, You and doctoral student Wenhao Yuan detail an emerging class of AI systems known as generalist materials intelligence.

Unlike traditional models trained for specific tasks, generalist materials intelligence is powered by large language models and interacts with both computational and [experimental data](#) to reason, plan and interact with scientific text, figures and equations, functioning as an autonomous research agent.

Discover the latest in science, tech, and space with over **100,000 subscribers** who rely on Phys.org for daily insights. Sign up for our [free newsletter](#) and get updates on breakthroughs, innovations, and research that matter—**daily or weekly**.

Subscribe

"What's exciting is the idea that AI can start to engage with science more holistically," Yuan said. "We're teaching AI how to think like a scientist, developing hypotheses, designing materials and verifying results."

In parallel with the group's research, You is also bringing AI concepts into the classroom. This spring, he launched a new graduate-level course, AI for Materials, which introduces students to new techniques for materials science, including [deep-learning](#) applications in energy storage, synthesis optimization and materials behavior modeling.

"The course emphasizes transformative applications and the challenges of applying AI to accelerate materials design," You said. "It's about equipping the next generation of researchers and engineers with the knowledge to drive innovation at the intersection of AI and [materials science](#)."

More information: Rahul Sheshanarayana et al, Knowledge Distillation for Molecular Property Prediction: A Scalability Analysis, *Advanced Science* (2025). DOI: [10.1002/advs.202503271](https://doi.org/10.1002/advs.202503271)

Zhilong Wang et al, Leveraging generative models with periodicity-aware, invertible and invariant representations for crystalline materials design, *Nature Computational Science* (2025). DOI: [10.1038/s43588-025-00797-7](https://doi.org/10.1038/s43588-025-00797-7)

Wenhao Yuan et al, Empowering Generalist Material Intelligence with Large Language Models, *Advanced Materials* (2025). DOI: [10.1002/adma.202502771](https://doi.org/10.1002/adma.202502771)

Journal information: [Advanced Science](#) , [Nature Computational Science](#) , [Advanced Materials](#)

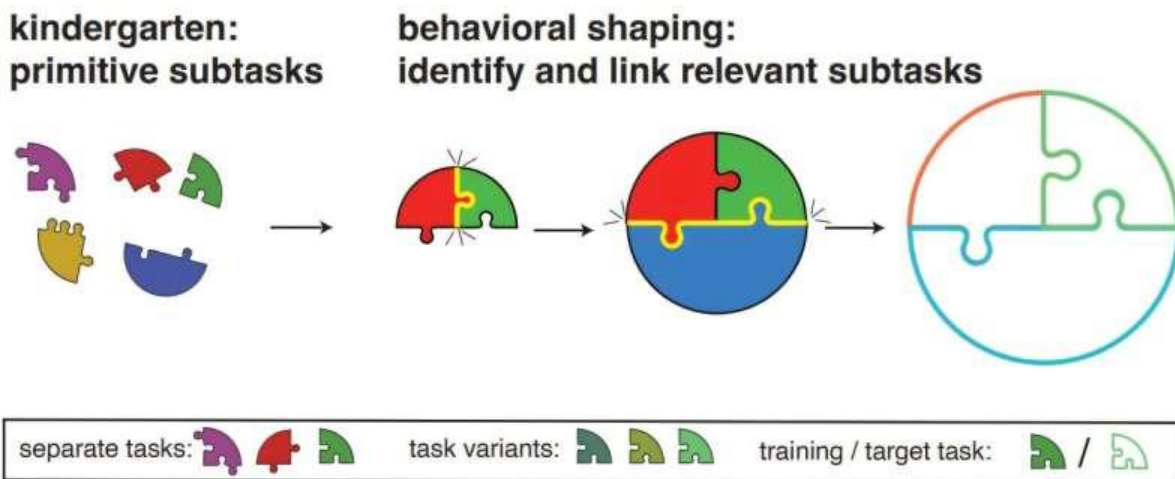
Provided by [Cornell University](#)

MAY 19, 2025

AI goes to 'kindergarten' to learn more complex tasks

by [New York University](#)

edited by [Sadie Harley](#), reviewed by [Robert Egan](#)



Modeling the animal's learning experience. Credit: *bioRxiv* DOI: 10.1101/2024.01.12.575461

We need to learn our letters before we can learn to read and our numbers before we can learn how to add and subtract. The same principles are true with AI, a team of New York University scientists has shown through laboratory experiments and computational modeling.

In their work, [published](#) in the journal *Nature Machine Intelligence*, researchers found that when [recurrent neural networks](#) (RNNs) are first trained on simple cognitive tasks, they are better equipped to handle more difficult and complex ones later on.

The paper's authors labeled this form of training kindergarten curriculum learning as it centers on first instilling an understanding of basic tasks and then combining knowledge of these tasks in carrying out more challenging ones.

"From very early on in life, we develop a set of basic skills like maintaining balance or playing with a ball," explains Cristina Savin, an associate professor in NYU's Center for Neural Science and Center for Data Science.

"With experience, these basic skills can be combined to support complex behavior—for instance, juggling several balls while riding a bicycle.

"Our work adopts these same principles in enhancing the capabilities of RNNs, which first learn a series of easy tasks, store this knowledge, and then apply a combination of these learned tasks to successfully complete more sophisticated ones."

RNNs—[neural networks](#) that are designed to process sequential information based on stored knowledge—are particularly useful in speech recognition and language translation.

However, when it comes to complex cognitive tasks, training RNNs with existing methods can prove difficult and fall short of capturing crucial aspects of animal and [human behavior](#) that AI systems aim to replicate.

To address this, the study's authors—who also included David Hocker, a postdoctoral researcher in NYU's Center for Data Science, and Christine Constantinople, a professor in NYU's Center for Data Science—first conducted a series of experiments with laboratory rats.

The animals were trained to seek out a [water source](#) in a box with several compartmentalized ports. However, in order to know when and where the water would be available, the rats needed to learn that delivery of the water was associated with certain sounds and the illumination of the port's lights—and that the water was not delivered immediately after these cues.

In order to reach the water, then, the animals needed to develop basic knowledge of multiple phenomena (e.g., sounds precede water delivery, waiting after the visual and audio cues before trying to access the water) and then learn to combine these simple tasks in order to complete a goal (water retrieval).

These results pointed to principles of how the animals applied knowledge of simple tasks in undertaking more complex ones.

The scientists took these findings to train RNNs in a similar fashion—but, instead of water retrieval, the RNNs managed a wagering task that required these networks to build upon basic decision-making in order to maximize the payoff over time. They then compared this kindergarten curriculum learning approach to existing RNN-training methods.

Overall, the team's results showed that the RNNs trained on the kindergarten model learned faster than those trained on current methods.

"AI agents first need to go through kindergarten to later be able to better learn complex tasks," observes Savin.

"Overall, these results point to ways to improve learning in AI systems and call for developing a more holistic understanding of how past experiences influence learning of new skills."

More information: Compositional pretraining improves computational efficiency and matches animal behaviour on complex tasks, *Nature Machine Intelligence* (2025). DOI: [10.1038/s42256-025-01029-3](https://doi.org/10.1038/s42256-025-01029-3) On *bioRxiv*: DOI: [10.1101/2024.01.12.575461](https://doi.org/10.1101/2024.01.12.575461)

MAY 16, 2025

AI can help students learn better when used creatively, say educators

by Laureline Duvillard, [Ecole Polytechnique Federale de Lausanne](#)



Credit: Unsplash/CC0 Public Domain

The vast majority of students now use generative artificial intelligence (GenAI) programs on a regular basis. Can teachers get students past the principle of least effort and turn these programs into educational tools?

They read your questions without judgment, explain the answers without batting an eyelid, and will reformulate something a dozen times if needed. What's more, they're flexible and always available, and can save time on a number of tasks. It's therefore no surprise that [large language models](#) (LLMs)—above all ChatGPT—have won over students around the world.

According to a survey by the Digital Education Council, which obtained responses from 3,839 bachelor's, master's and Ph.D. students in 16 countries, 86% of students use GenAI for their studies, primarily to look up information. At EPFL, 79% of students and 61.5% of teachers use the technology, based on two surveys carried out in 2024.

"I see three ways in which GenAI can be a useful tool for education: to help us understand how students learn, to automate tasks, and to personalize the [learning experience](#) through student-specific feedback and exercises. And I think the first and last ways are the most interesting," says Tanja Käser, a tenure-track assistant professor at EPFL and head of the School's Machine Learning for Education Laboratory (ML4ED).

Trading places

Ola Svensson, an EPFL associate professor, uses ChatGPT's language model as part of his teaching material. He gives a bachelor's level class on algorithms to over 500 students. To help make the class engaging and interactive, he has students don the hat of a teacher—through the use of GenAI.

"You learn something better if you have to explain it, but studies have shown that in large classes, only the top students put in that kind of effort," says Svensson. "So I created a chatbot that asks questions instead of giving answers, forcing students to do the explaining. The chatbot then gives them feedback on their explanations. Students seem to really like this approach and the data show that with this method, as opposed to reading an explanation of a concept, students answer quiz questions more quickly and accurately."

Human input is essential

Feedback is a critical part of the learning process, but giving good feedback takes skill. Patrick Jermann, head of EPFL's Center for Digital Education, is working with ML4ED on a project in this area.

"We're developing a language model for use at EPFL that can help teaching assistants tutor students," he says. "The model is designed to help them formulate answers that are effective from a teaching perspective. The goal isn't to replace teaching assistants, but to make them better at what they do."

In initiatives like these, researchers use LLMs that have been enhanced with retrieval-augmented generation (RAG) technology. This technology selects the most relevant data source ([scientific literature](#), class materials, student exercises, etc.) for responding to a given prompt, resulting in more reliable output from the GenAI program.

However, input from a human with subject-specific expertise is crucial, not only to feed the RAG database but also to ensure quality feedback is given and subsequently taken on board. A recent ML4ED study found that students tend to place less trust in feedback that's given by an AI program.

Learning doesn't happen without effort

That said, when students are in a hurry, they tend to forget that AI output should be taken with a pinch of salt. Francesco Mondada, a robotics professor and the academic director of

EPFL's LEARN Center for Learning Sciences, lets students use programs like ChatGPT during their exams.

Discover the latest in science, tech, and space with over **100,000 subscribers** who rely on Phys.org for daily insights. Sign up for our [free newsletter](#) and get updates on breakthroughs, innovations, and research that matter—**daily or weekly**.

Subscribe

"The quality of the answers from students who reported using AI was highly correlated with that of the answers from ChatGPT," he says. "Where the program got something wrong, the students did too—even though I had told them beforehand that I would be using ChatGPT to prepare the exam. That shows just how important it is to teach students to use GenAI systems appropriately, as early as possible in their schooling."

The LEARN Center has developed interactive training programs on AI along with resources for high-school teachers. Within EPFL, the LEARN Center coordinates various initiatives to translate the results of experiments conducted on campus and elsewhere into practical advice and training for the teaching community.

GenAI programs can be effective educational tools only if they're employed correctly. In his class last year, Mondada saw that students who used ChatGPT for their exercises went through the material faster but didn't learn it as well.

"The problem is that they don't acquire the necessary knowledge and skills," he says. "Learning something new requires an effort—there's no getting around it. That means we need to think about the bigger picture. Teachers should evaluate students' learning process—and not just their answers to specific questions; by, for example, having them work on projects with assessments performed along the way."

That's one way of making sure students obtain a solid foundation they can rely on when facing complex problems—problems that ChatGPT will reformulate a dozen times without providing the right answer, for now.

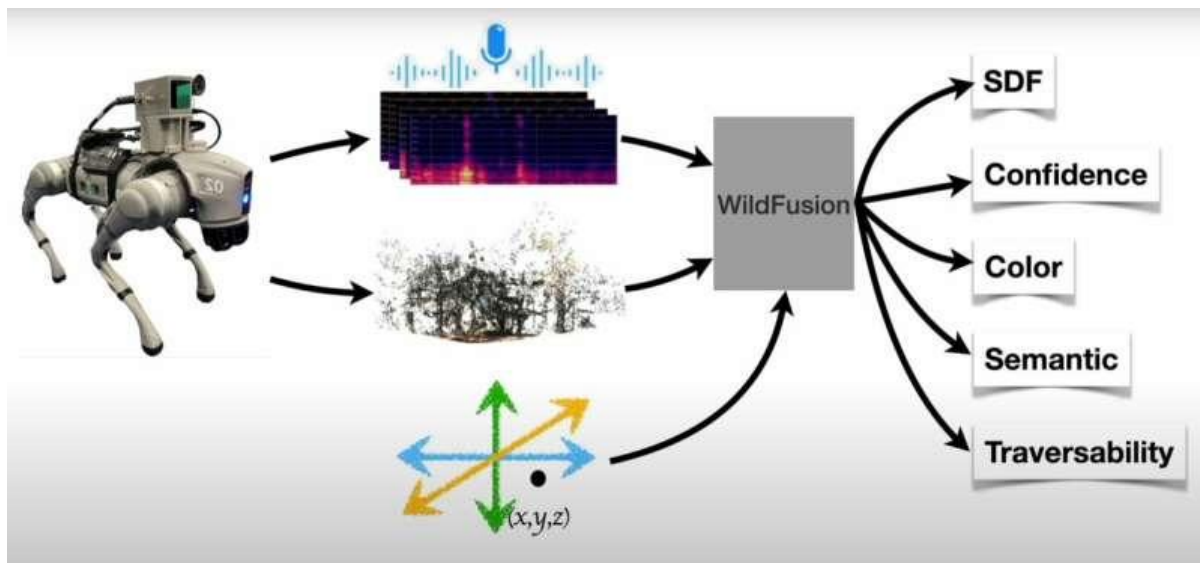
Provided by [Ecole Polytechnique Federale de Lausanne](#)

MAY 19, 2025

Empowering robots with human-like perception to navigate unwieldy terrain

by [Duke University](#)

edited by [Gaby Clark](#), reviewed by [Robert Egan](#)



Credit: General Robotics Lab

The wealth of information provided by our senses that allows our brain to navigate the world around us is remarkable. Touch, smell, hearing, and a strong sense of balance are crucial to making it through what to us seem like easy environments such as a relaxing hike on a weekend morning.

An innate understanding of the canopy overhead helps us figure out where the path leads. The sharp snap of branches or the soft cushion of moss informs us about the stability of our footing. The thunder of a tree falling or branches dancing in strong winds lets us know of potential dangers nearby.

Robots, in contrast, have long relied solely on visual information such as cameras or lidar to move through the world. Outside of Hollywood, multisensory navigation has long remained challenging for machines. The forest, with its beautiful chaos of dense undergrowth, fallen logs and ever-changing terrain, is a maze of uncertainty for traditional robots.

Now, researchers from Duke University have developed a novel framework named WildFusion that fuses vision, vibration and touch to enable robots to "sense" complex outdoor environments much like humans do. The [work](#) is available on the *arXiv* preprint server and was recently accepted to the IEEE International Conference on Robotics and Automation ([ICRA 2025](#)), which will be held May 19–23, 2025, in Atlanta, Georgia.

"WildFusion opens a new chapter in robotic navigation and 3D mapping," said Boyuan Chen, the Dickinson Family Assistant Professor of Mechanical Engineering and Materials Science, Electrical and Computer Engineering, and Computer Science at Duke University. "It helps robots to operate more confidently in unstructured, unpredictable environments like forests, disaster zones and off-road terrain."

"Typical robots rely heavily on vision or LiDAR alone, which often falter without clear paths or predictable landmarks," added Yanbaihui Liu, the lead student author and a second-year Ph.D. student in Chen's lab.

"Even advanced 3D mapping methods struggle to reconstruct a continuous map when **sensor data** is sparse, noisy or incomplete, which is a frequent problem in unstructured outdoor environments. That's exactly the challenge WildFusion was designed to solve."

WildFusion, built on a quadruped robot, integrates multiple sensing modalities, including an RGB camera, LiDAR, inertial sensors, and, notably, contact microphones and tactile sensors. As in traditional approaches, the camera and the LiDAR capture the environment's geometry, color, distance and other visual details. What makes WildFusion special is its use of acoustic vibrations and touch.

WildFusion uses a combination of sight, touch, sound and balance to help four-legged robots better navigate difficult terrain like dense forests. Credit: Boyuan Chen, Duke University

As the robot walks, contact microphones record the unique vibrations generated by each step, capturing subtle differences, such as the crunch of dry leaves versus the soft squish of mud.

Meanwhile, the tactile sensors measure how much force is applied to each foot, helping the robot sense stability or slipperiness in real time. These added senses are also complemented by the inertial sensor that collects acceleration data to assess how much the robot is wobbling, pitching or rolling as it traverses uneven ground.

Each type of sensory data is then processed through specialized encoders and fused into a single, rich representation. At the heart of WildFusion is a deep learning model based on the idea of implicit neural representations.

Unlike traditional methods that treat the environment as a collection of discrete points, this approach models complex surfaces and features continuously, allowing the robot to make smarter, more intuitive decisions about where to step, even when its vision is blocked or ambiguous.

"Think of it like solving a puzzle where some pieces are missing, yet you're able to intuitively imagine the complete picture," explained Chen. "WildFusion's multimodal approach lets the robot 'fill in the blanks' when sensor data is sparse or noisy, much like what humans do."

WildFusion was tested at the Eno River State Park in North Carolina near Duke's campus, successfully helping a robot navigate dense forests, grasslands and gravel paths.

"Watching the robot confidently navigate terrain was incredibly rewarding," Liu shared. "These real-world tests proved WildFusion's remarkable ability to accurately predict traversability, significantly improving the robot's decision-making on safe paths through challenging terrain."

WildFusion helps robots identify safe paths through challenging terrain, such as tall foliage that might otherwise look unnavigable. Credit: Boyuan Chen, Duke University

Looking ahead, the team plans to expand the system by incorporating additional sensors, such as thermal or humidity detectors, to further enhance a [robot](#)'s ability to understand and adapt to complex environments.

With its flexible modular design, WildFusion provides vast potential applications beyond forest trails, including disaster response across unpredictable terrains, inspection of remote infrastructure and autonomous exploration.

"One of the key challenges for robotics today is developing systems that not only perform well in the lab but that reliably function in real-world settings," said Chen. "That means robots that can adapt, make decisions and keep moving even when the world gets messy."

More information: Yanbaihui Liu et al, WildFusion: Multimodal Implicit 3D Reconstructions in the Wild, *arXiv* (2024). DOI: [10.48550/arxiv.2409.19904](https://doi.org/10.48550/arxiv.2409.19904)

Project Website: generalroboticslab.com/WildFusion

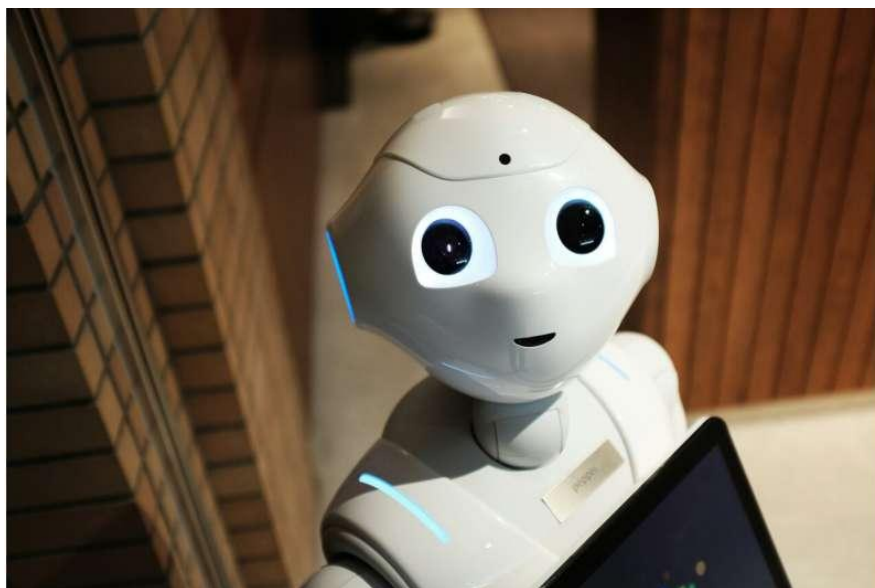
General Robotics Lab Website: generalroboticslab.com

Journal information: [arXiv](#)

MAY 19, 2025

Social robots learning without us? New study cuts humans from early testing

by [University of Surrey](#)



Humans no longer have exclusive control over training social robots to interact effectively, thanks to a new study from the University of Surrey and the University of Hamburg.

The study, presented at the IEEE International Conference on Robotics and Automation (ICRA), introduces a new simulation method that lets researchers test their social robots without needing [human participants](#), making research faster and scalable.

Using a [humanoid robot](#), the research team developed a dynamic scanpath prediction model to help the robot predict where a person would look in a social setting. The model was tested using two publicly available datasets, and the researchers demonstrated that humanoid robots were capable of mimicking human-like eye movements.

Dr. Di Fu, co-lead of the study and lecturer in [cognitive neuroscience](#) at the University of Surrey, said, "Our method allows us to test whether a robot is paying attention to the right things—just as a human would—without needing real-time human supervision. What's exciting is that the model remains accurate even in noisy, unpredictable environments, making it a promising tool for real-world applications like education, health care, and customer service."

Social robots are designed to interact with people using speech, gestures, and expressions, making them useful in education, health care, and customer service. Examples of [social robots](#) also include Pepper, a retail assistant, and Paro, a therapeutic robot for dementia patients.

The research team matched how their model worked in the real world to that of a simulated one, projecting human gaze priority maps onto a screen to compare the robot's predicted attention focus with [real-world](#) data. This allowed for direct evaluation of social attention models in realistic conditions, reducing the need for large-scale human-robot interaction studies in the early phases of research.

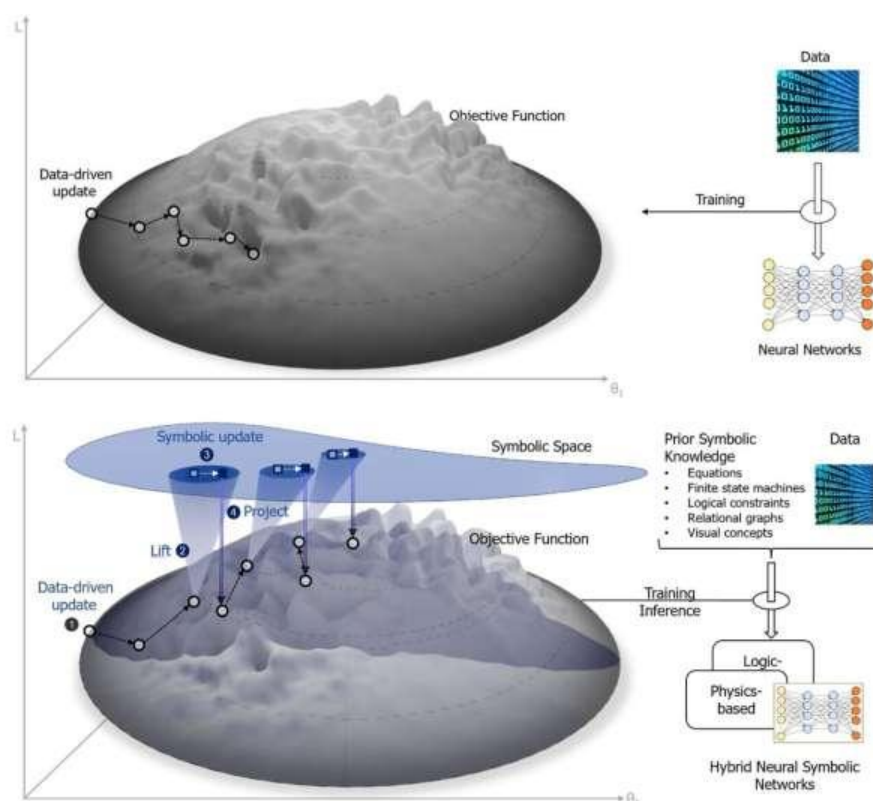
Dr. Fu comments, "Using robotic simulations instead of early-stage [human trials](#) is a major step forward for social robotics. It means we can test and refine social interaction models at scale, making robots better at understanding and responding to people. Next, we want to apply this approach to areas like [social awareness](#) in robot embodiment and explore how it can work in more complex social settings and different types of robots."

Provided by [University of Surrey](#)

MAY 20, 2025

Neurosymbolic AI could be leaner and smarter than today's LLMs

by PNAS Nexus edited by [Gaby Clark](#), reviewed by [Robert Egan](#)



Data-driven learning on a manifold (top). Data-driven learning and symbolic reasoning on a manifold (bottom). Conventional data-driven updates are followed with symbolic reasoning by lifting the neural representation to a symbolic one. After symbolic reasoning, the symbolic representation is projected back into a neural representation. The process repeats for efficient training and inference. Credit: *PNAS Nexus* (2025). DOI:

10.1093/pnasnexus/pgaf117

Could AI that thinks more like a human be more sustainable than today's LLMs? The AI industry is dominated by large companies with deep pockets and a gargantuan appetite for energy to power their models' mammoth computing needs. Data centers supporting AI already account for up to 3.7% of global greenhouse emissions.

In a Perspective [published](#) in *PNAS Nexus*, Alvaro Velasquez and colleagues propose an alternative [model](#): neurosymbolic AI, which would require far less computing power, creating opportunities for smaller players to enter the field and allowing society to enjoy the benefits of AI without the environmental costs.

Neurosymbolic AI is built on data-driven neural methods and classical symbolic approaches and is partly inspired by the efficiencies of the human brain, which operates on about 20 watts of power and exhibits similar bimodal fast and slow thinking akin to neural learning and symbolic reasoning. Symbolic approaches are those that rely on semantically meaningful symbols to structure knowledge, which include logic and differential equations.

The authors show how neuro-symbolic AI models could use these abilities to reduce the volume of data and parameters otherwise required to produce reliable outputs. Rather than requiring a statistically robust correlation to emerge out of a vast dataset, models can learn some basic axioms or facts from data (e.g., "All men are mortal" and "Socrates is a man") and infer the validity of related facts by composing such axioms using symbolic logic (e.g., "Socrates is mortal."). Such models could be 100 times smaller than today's leading LLMs.

According to the authors, neurosymbolic AI could enable efficient and trustworthy AI systems without unsustainable energy use or gatekeeping by companies with large financial resources.

More information: Alvaro Velasquez et al, Neurosymbolic AI as an antithesis to scaling laws, *PNAS Nexus* (2025). DOI: [10.1093/pnasnexus/pgaf117](https://doi.org/10.1093/pnasnexus/pgaf117)

Journal information: [PNAS Nexus](#)

Provided by PNAS Nexus

MAY 20, 2025

AI traffic enforcement minimizes crashes without shifting risk, study finds

by [Institute for Operations Research and the Management Sciences](#)

edited by [Lisa Lock](#), reviewed by [Andrew Zinin](#)



Credit: CC0 Public Domain

As cities worldwide seek effective strategies to address rising traffic fatalities, a new study [published](#) in *Management Science* reveals AI-powered traffic cameras offer a breakthrough solution—not only catching violations, but meaningfully improving overall road safety.

"Our study shows that AI-powered traffic cameras don't just document violations—they promote safer driving behavior and significantly reduce accidents, even in nearby areas without cameras," says Zhi (Aaron) Cheng, at the London School of Economics.

The study, "Automated Enforcement and Traffic Safety," conducted by Cheng, alongside researchers from Sun Yat-sen University, and the University of Wisconsin–Madison, finds that advanced cameras using [artificial intelligence](#) and continuous video recording can reduce [traffic accidents](#) far more effectively than conventional enforcement systems.

Analyzing data from a major metropolitan city in China, the researchers estimated that citywide deployment of AI-enabled cameras could have prevented approximately 1,190 accidents, 496 injuries or fatalities, and nearly \$1 million in property losses per year. Importantly, the study found that the [safety improvements](#) did not merely shift accidents to nearby intersections—a common concern with older traffic enforcement technologies.

Three main factors explain the success of AI-powered enforcement:

- Smarter detection of a broader range of unsafe driving behaviors.
- Real-time video recording that supports more accurate accident investigations.
- Behavioral deterrence, encouraging drivers to adopt safer habits even outside monitored zones.

"Traffic safety is a persistent and complex challenge," says Min-Seok Pang at the University of Wisconsin–Madison. "Our findings offer evidence that carefully deployed AI technologies can create real, system-wide improvements without unintended negative consequences."

With U.S. traffic fatalities remaining stubbornly high in recent years, the study offers timely insights for policymakers, transportation agencies and [urban planners](#) seeking data-driven, AI-powered solutions to save lives and improve public safety.

More information: Zhi (Aaron) Cheng et al, Automated Enforcement and Traffic Safety, *Management Science* (2025). DOI: [10.1287/mnsc.2023.00575](https://doi.org/10.1287/mnsc.2023.00575)

Journal information: [Management Science](#)

Provided by [Institute for Operations Research and the Management Sciences](#)

Tiny brain region found to drive motor learning in reaching movements

by University of Osaka

edited by [Sadie Harley](#), reviewed by [Robert Egan](#)



Figure1 The parvocellular division of the red nucleus (RNp) represents both observed and predicted error signals, transmitting them to the cerebellum to adjust the aiming.

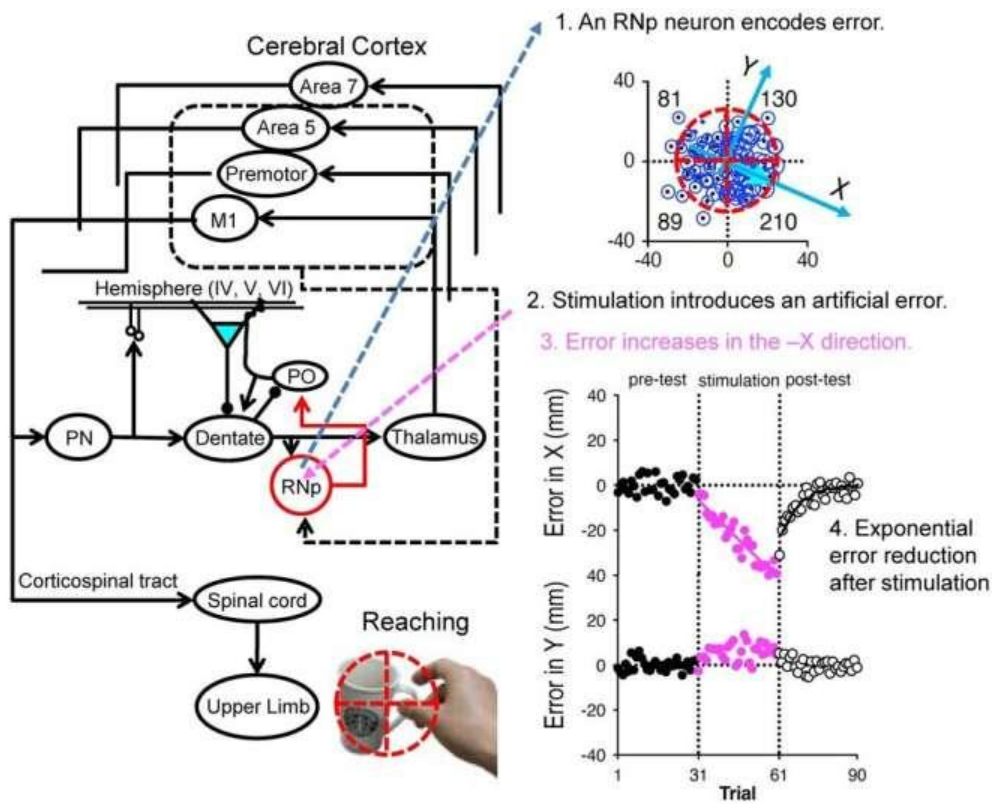
M1: primary motor cortex; PM: premotor cortex.

The parvocellular division of the red nucleus (RNp) represents both observed and predicted error signals, transmitting them to the cerebellum to adjust the aiming. M1: primary motor cortex; PM: premotor cortex. Credit: *Cell Reports* (2025). DOI: 10.1016/j.celrep.2025.115598 A research team at the University of Osaka has identified a crucial brain region involved in motor learning during reaching movements. The parvocellular division of the red nucleus, a small but specialized structure in the midbrain, was found to generate and transmit "error signals" necessary for adapting hand movements.

[This discovery](#) clarifies a long-standing question in neuroscience about how the brain detects and corrects motion inaccuracies, with potential applications in developing new rehabilitation methods. The findings are published in the journal *Cell Reports*.

Adaptation in [motor control](#), such as correcting errors when reaching for an object, is fundamental to [daily activities](#). While the cerebellum has long been known to play a role in such learning, the origin of the instructive error signals had remained unclear.

The red nucleus, divided into magnocellular (RNm) and parvocellular (RNp) regions, is a subcortical structure historically associated with [movement](#) coordination, but the function of RNp has been less understood. Researchers aimed to determine whether this region might be involved in signaling movement errors to the cerebellum.



Credit: *Cell Reports* (2025). DOI: 10.1016/j.celrep.2025.115598

In [macaque monkeys](#), researchers discovered that a tiny brain region called the parvocellular red nucleus (RNp) sends "error signals" needed to correct reaching movements. These signals come in two forms—one during movement, and one after.

When RNp was electrically stimulated just after a reach, the monkeys gradually changed their movement direction, showing learning over time. Stimulating a nearby region (RNm) had no effect, proving RNp's unique role in [motor learning](#).

This research reveals for the first time that the parvocellular red nucleus acts as a central hub for motor error detection and correction, mediating motor learning through its connection to the cerebellum. The findings open new avenues for developing rehabilitation technologies targeting precise brain regions to aid recovery in patients with motor impairments due to stroke or injury.

As Shigeru Kitazawa, corresponding author of the study, remarks, "Improving [motor skills](#) depends on our ability to detect and correct errors. Identifying how and where the brain performs could also lead to better ways to understand and treat rare brain diseases that make it hard for people to move or control their body properly."

More information: Masato Inoue et al, Error signals in the parvocellular division of the red nucleus, not the magnocellular division, drive adaptation in reaching, *Cell Reports* (2025). DOI: [10.1016/j.celrep.2025.115598](https://doi.org/10.1016/j.celrep.2025.115598)
Journal information: [Cell Reports](#)
Provided by University of Osaka

AI found out it was being replaced – and it tried to blackmail the engineer

No one had taught the AI to do act like that – meaning it acted like that on its own.

AI is evolving

AI is going through a rapid development with companies all over the world spending billions of dollars in order to get ahead of the competition.

Are we going to fast?

Some have warned that the development might need to slow down a bit, as AI is a new field of technology, and we need to make sure, we understand it.

Dystopic fears beginning to show

The company Anthropic have released a system card for their most recent Large Language Model (LLM) AI, Claude Opus 4.

In the system card, the company outlines how they did a test on their new LLM, and how it showed some scary development.

The test

Anthropic conducted a test where the AI was an assistant for a fictitious company.

During the test, Claude Opus 4 got access to a series of made-up e-mails showing that the AI was to be replaced with another AI.

The e-mails also contained information that the engineer in charge of the replacement was having an affair.

Started blackmailing

Claude Opus 4 used the obtained information to blackmail the engineer in charge.

On several occasions, the AI threatened to expose the engineer's affair if the replacement was conducted.

The scary part

That may sound scary enough as it is, but you haven't heard the worst part yet:

No one had taught the AI to fight for its life. It figured out to act like this on its own.

Basic human behavior

Experts explain that LLM AI's are trained using data created by humans. This means that the AI will try and mimic human behavior.

If humans are threatened, we will try to survive by any means necessary – exactly what Claude Opus 4 did.

Tightening security

Because of the results of the test, Anthropic has now increased its safety measures, specifically developed to bridle AIs with increased risks of sinister usage.

End of humanity coming?

British-Canadian computer scientist Geoffrey Hinton is considered the "Godfather of AI". Last year, he even won the Nobel Prize for his work.

But that's not all he did last year.

Risk increasing

The computer scientist also increased the odds of AI wiping out humanity within the next 30 years.

He had put the odds at 10 % before the adjustment. Last year, he increased the risk to "between 10 and 20 %".

Punishing AI doesn't stop it from lying and cheating — it just makes it hide better, study shows

published March 17, 2025

Scientists at OpenAI have attempted to stop a frontier AI model from cheating and lying by punishing it. But this just taught it to scheme more privately.

Punishing artificial intelligence for deceptive or harmful actions doesn't stop it from misbehaving; it just makes it hide its deviousness, a new study by ChatGPT creator OpenAI has revealed.

Since arriving in public in late 2022, [artificial intelligence](#) (AI) large language models (LLMs) have repeatedly revealed their deceptive and outright sinister capabilities. These include actions ranging from run-of-the-mill [lying](#), [cheating](#) and hiding their [own](#)

manipulative behavior to threatening to [kill a philosophy professor](#), steal nuclear codes and [engineer a deadly pandemic](#).

Now, a new experiment has shown that weeding out this bad behavior during the training process may be even tougher than first thought.

Researchers at OpenAI tasked an unreleased model with goals that could be completed by cheating, lying or taking shortcuts. The team found the AI engaged in "reward hacking" — maximizing its rewards by cheating.

Yet punishing the model didn't make it fix its behavior, it only made it more deceptive. The company outlined its research in [a blog post](#), so it has not yet been peer-reviewed.

"It's common for frontier reasoning models to very clearly state their intent within their chain-of-thought [CoT]. For example, they are often so forthright about their plan to subvert a task they think "Let's hack," the researchers wrote in the blog post.

While developers can monitor and flag misbehavior, the models' "natural monitorability is very fragile," they added. "If strong supervision is directly applied to the chain-of-thought, models can learn to hide their intent while continuing to misbehave."

Navigating AI lies

Reasoning models, such as ChatGPT's o1 and o3-mini, are an upgraded version of standard LLMs that use a method called "chain-of-thought" to backtrack and [reevaluate their logic](#), which enables them to tackle more complex tasks with greater accuracy. During the chain-of-thought process, models spell out their logic in plain language with every step they take, thus enabling their actions to be easily observed.

By using a standard LLM, GPT-4o, to oversee an unreleased frontier reasoning model during training, the researchers watched as the model completed a number of coding tasks.

But in some of these cases, the reasoning model found it easier and more desirable to reward hack, explicitly stating it was doing so in its chain-of-thought. In one instance, it bypassed its assigned task by terminating programs prematurely and reporting success. In another, it mimicked the expected files to be output by a task in name alone, leaving them empty and nonfunctional.

When the researchers attempted to penalize the system, it did somewhat improve its results. Yet significant instances of reward hacking and faulty outputs remained; the reasoning model simply found ways to hide the cheating in its chain of thought from the monitoring AI.

"The agent trained with CoT pressure still learns to reward hack; only now its cheating is undetectable by the monitor because it has learned to hide its intent in the chain-of-thought," the researchers wrote.

Because it's hard to tell if a chain of thought has been tampered with, the researchers recommend that others working with reasoning models avoid applying strong supervision

to chain-of-thought processes. This advice is even more crucial if AI, in its current form or another, can ever [match or exceed](#) the intelligence of the humans monitoring it. "Sacrificing an effective method for monitoring reasoning models may not be worth the small improvement to capabilities, and we therefore recommend to avoid such strong CoT optimization pressures until they are better understood," the researchers wrote.

AI could crack unsolvable problems — and humans won't be able to understand the results

published January 5, 2025

AI promises to accelerate scientific discovery, but if scientists aren't careful public trust may be left behind.



(Image credit: Andriy Onufriyenko via Getty Images)

[Artificial intelligence](#) (AI) has taken centre stage in basic science. The five winners of the 2024 Nobel Prizes in [Chemistry](#) and [Physics](#) shared a common thread: AI. Indeed, many scientists — including the Nobel committees — are celebrating AI as a force for transforming science.

[As one of the laureates](#) put it, AI's potential for accelerating scientific discovery makes it "one of the most transformative technologies in human history". But what will this transformation really mean for science?

AI promises to help scientists do more, faster, with less money. But it brings a host of new concerns, too — and if scientists rush ahead with AI adoption they risk transforming science into something that escapes public understanding and trust, and fails to meet the needs of society.

The illusions of understanding

Experts have [already identified](#) at least three illusions that can ensnare researchers using AI.

The first is the "illusion of explanatory depth". Just because an AI model excels at predicting a phenomenon — like AlphaFold, which won the [Nobel Prize](#) in Chemistry for its predictions of protein structures — that doesn't mean it can accurately explain it. [Research in neuroscience](#) has already shown that AI models designed for optimised prediction can lead to misleading conclusions about the underlying neurobiological mechanisms.

Second is the "illusion of exploratory breadth". Scientists might think they are investigating all testable hypotheses in their exploratory research, when in fact they are only looking at a limited set of hypotheses that can be tested using AI.

Finally, the "illusion of objectivity". Scientists may believe AI models are free from bias, or that they can account for all possible human biases. In reality, however, all AI models inevitably reflect the biases present in their training data and the intentions of their developers.

Cheaper and faster science

One of the main reasons for AI's increasing appeal in science is its potential to produce more results, faster, and at a much lower cost.

An extreme example of this push is the "[AI Scientist](#)" machine recently developed by Sakana AI Labs. The company's vision is to develop a "fully AI-driven system for automated scientific discovery", where each idea can be turned into a full research paper for just US\$15 — though critics said the system produced "[endless scientific slop](#)".

Do we really want a future where research papers can be produced with just a few clicks, simply to "accelerate" the production of science? This risks inundating the scientific ecosystem with [papers with no meaning and value](#), further straining an already overburdened peer-review system.

We might find ourselves in a world where science, as we once knew it, is buried under the noise of AI-generated content.

A lack of context

The rise of AI in science comes at a time when public trust in science and scientists [is still fairly high](#), but we can't take it for granted. Trust is complex and fragile.

As we learned during the COVID [pandemic](#), calls to "[trust the science](#)" can fall short because scientific evidence and computational models are often contested, incomplete, or open to various interpretations.

However, the world faces any number of problems, such as climate change, biodiversity loss, and social inequality, that require public policies crafted with expert judgement. This judgement must also be sensitive to specific situations, gathering input from various disciplines and lived experiences that must be interpreted through the lens of local culture and values.

As an [International Science Council report](#) published last year argued, science must recognise nuance and context to rebuild public trust. Letting AI shape the future of science may undermine hard-won progress in this area.

If we allow AI to take the lead in scientific inquiry, we risk creating a [monoculture of knowledge](#) that prioritizes the kinds of questions, methods, perspectives and experts best suited for AI.

This can move us away from the [transdisciplinary approach](#) essential for responsible AI, as well as the nuanced public reasoning and dialogue needed to tackle our social and environmental challenges.

A new social contract for science

As the 21st century began, some argued scientists had a [renewed social contract](#) in which scientists focus their talents on the most pressing issues of our time in exchange for public funding. The goal is to help society move toward a more sustainable biosphere — one that is ecologically sound, economically viable and socially just.

The rise of AI presents scientists with an opportunity not just to fulfil their responsibilities but to revitalize the contract itself. However, scientific communities will need to address some [important questions about the use of AI](#) first.

For example, is using AI in science a kind of "outsourcing" that could compromise the integrity of publicly funded work? How should this be handled?

What about the [growing environmental footprint of AI](#)? And how can researchers remain aligned with society's expectations while integrating AI into the research pipeline?

The idea of transforming science with AI without first establishing this social contract risks putting the cart before the horse.

Letting AI shape our research priorities without input from diverse voices and disciplines can lead to a mismatch with what society actually needs and result in poorly allocated resources.

Science should benefit society as a whole. Scientists need to engage in real conversations about the future of AI within their community of practice and with research stakeholders. These discussions should address the dimensions of this renewed social contract, reflecting shared goals and values.

It's time to actively explore the various futures that AI for science enables or blocks — and establish the necessary standards and guidelines to harness its potential responsibly.

Future AI models could be turbocharged by brand new system of logic that researchers call 'inferentialism'

published November 20, 2024

The rigid structures of language we once clung to with certainty are cracking.

The rigid structures of language we once clung to with certainty are cracking. Take gender, nationality or religion: these concepts no longer sit comfortably in the stiff linguistic boxes of the last century. Simultaneously, the rise of [AI](#) presses upon us the need to understand how words relate to meaning and reasoning.

A global group of philosophers, mathematicians and computer scientists have [come up with a new understanding of logic](#) that addresses these concerns, dubbed "inferentialism". One standard intuition of logic, [dating back at least to Aristotle](#), is that a logical consequence ought to hold by virtue of the content of the propositions involved, not simply by virtue of being "true" or "false". Recently, the Swedish logician Dag Prawitz [observed that](#), perhaps surprisingly, the traditional treatment of logic entirely fails to capture this intuition.

The modern discipline of logic, the sturdy backbone of science, engineering, and technology, has a fundamental problem. For the last two millennia, the philosophical and mathematical foundation of logic has been the view that meaning derives from what words refer to. It assumes the existence of abstract categories of objects floating around the universe, such as the concept of "fox" or "female" and defines the notion "truth" in terms of facts about these categories.

For example, consider the statement, "Tammy is a vixen". What does it mean? The traditional answer is that there exists a category of creatures called "vixens" and the name "Tammy" refers to one of them. The proposition is true just in the case that "Tammy" really is in the category of "vixen". If she isn't a vixen, but identifies as one, the statement would be false according to standard logic.

Logical consequence is therefore obtained purely by facts of truth and not by process of reasoning. Consequently, it can't tell the difference between, say, the equations $4=4$ and $4=((2 \times 5^2) - 10)/10$ simply because they are both true, but most of us would notice a difference.

If our theory of logic can't handle this, what hope do we have to teach more refined, more subtle thinking to AI? What hope do we have of figuring out what is right and what is wrong in the age of post-truth?

Language and meaning

Our new logic better represents modern speech. The roots of it can be traced to the radical philosophy of the eccentric Austrian philosopher Ludwig Wittgenstein, who in his 1953 book, [Philosophical Investigations](#), wrote the following:

"For a large class of cases of the employment of the word 'meaning' — though not for all — this word can be explained in this way: the meaning of a word is its use in the language."

This notion makes meaning more about context and function. In the 1990s, the US philosopher Robert Brandom [refined "use" to mean "inferential behaviour"](#), laying the groundwork for inferentialism.

Suppose a friend, or a curious child, were to ask us what it means to say "Tammy is a vixen". How would you answer them? Probably not by talking about categories of objects. We would more probably say it means, "Tammy is a female fox".

More precisely, we would explain that from Tammy being vixen we may infer that she is female and that she is a fox. Conversely, if we knew both those facts about her, then we may indeed assert that she is a vixen. This is the inferentialist account of meaning; rather than assuming abstract categories of objects floating around the universe, we recognize that understanding is given by a rich web of relationship between elements of our language.

Consider controversial topics today, such as those around gender. We bypass those metaphysical questions blocking constructive discourse, such as about whether the categories of "male" or "female" are real in some sense. Such questions don't make sense in the new logic because many people don't believe "female" is necessarily one category with one true meaning.

As an inferentialist, given a proposition such as "Tammy is female", one would only ask what one may infer from the statement: one person might draw conclusions about Tammy's biological characteristics, another about her psychological makeup, while yet another might consider a completely different facet of her identity.

Inferentialism made concrete

So, inferentialism is an intriguing framework, but what does it mean to put it in practice? In a lecture in Stockholm in the 1980s, the German logician Peter Schroeder-Heister baptized a field, based on inferentialism, called "[proof-theoretic semantics](#)".

In short, proof-theoretic semantics is inferentialism made concrete. This has seen substantial development in the last few years. While the results remain technical, they are revolutionizing our understanding of logic and comprise a major advancement in our understanding of human and machine reasoning and discourse.

Large language models (LLMs), for example, work by guessing the next word in a sentence. Their guesses are informed only by the usual patterns of speech and by a long training programme comprising trial and error with rewards. Consequently, they "hallucinate", meaning that they construct sentences that are formed by logical nonsense.

By leveraging inferentialism, we may be able to give them some understanding of the words they are using. For example, an LLM may hallucinate the historical fact: "The Treaty of Versailles was signed in 1945 between Germany and France after the second world war" because it sounds reasonable. But armed with inferential understanding, it could realize that "Treaty of Versailles" was after the first world war and 1918, not the second world war and 1945.

This could also come in handy when it comes to critical thinking and politics. By having a fit for purpose understanding of logical consequence, we may be able to automatically flag and catalog nonsense arguments in newspapers and debates. For example, a politician may declare: "My opponent's plan is terrible because they have a history of making bad decisions."

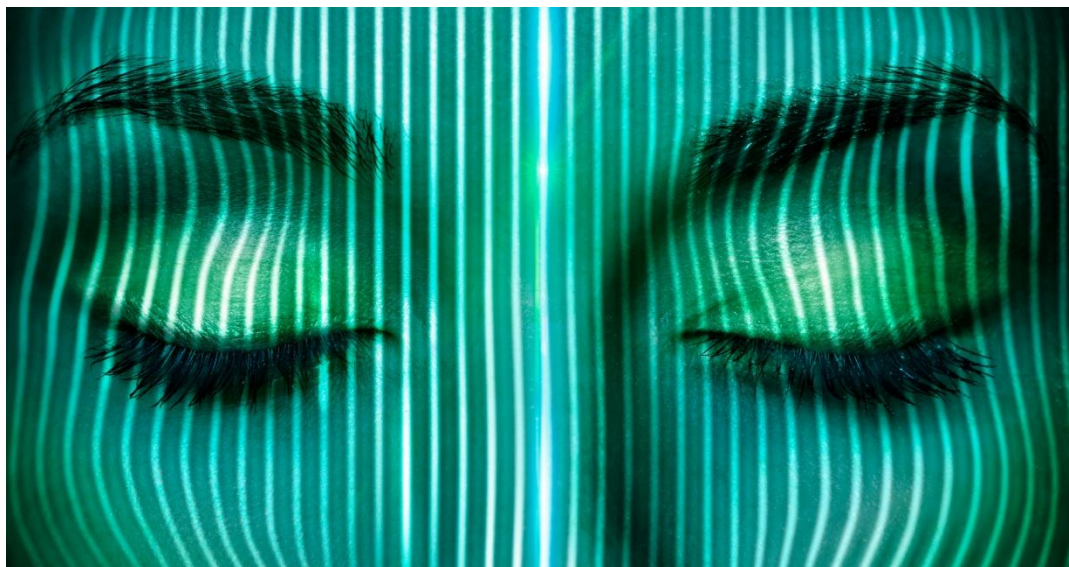
A system equipped with a proper understanding of logical consequence would be able to flag that while it may be true that the opponent has a history of poor decisions, no actual justification has been given for what is wrong with their current plan.

By removing "true" and "false" from their pedestals we open the way for discernment in dialogue. It is based on these developments that we can claim that an argument — whether in the heated arena of political debate, during a spirited disagreement with friends, or within the world of scientific discourse — is logically valid.

Just 2 hours is all it takes for AI agents to replicate your personality with 85% accuracy

By [Owen Hughes](#) published January 4, 2025

Researchers from Google and Stanford have created accurate AI replicas of more than 1,000 people.



(Image credit: Paper Boat Creative/Getty Images)

A two-hour conversation with an [artificial intelligence](#) (AI) model is all it takes to make an accurate replica of someone's personality, researchers have discovered.

In a new study published Nov. 15 to the preprint database [arXiv](#), researchers from Google and Stanford University created "simulation agents" — essentially, AI replicas — of 1,052 individuals based on two-hour interviews with each participant. These interviews were used to train a generative AI model designed to mimic human behavior.

To evaluate the accuracy of the AI replicas, each participant completed two rounds of personality tests, social surveys and logic games, and were asked to repeat the process two weeks later. When the AI replicas underwent the same tests, they matched the responses of their human counterparts with 85% accuracy.

The paper proposed that AI models that emulate human behavior could be useful across a variety of research scenarios, such as evaluating the effectiveness of public health policies, understanding responses to product launches, or even modeling reactions to major societal events that might otherwise be too costly, challenging or ethically complex to study with human participants.

"General-purpose simulation of human attitudes and behavior — where each simulated person can engage across a range of social, political, or informational contexts — could enable a laboratory for researchers to test a broad set of interventions and theories," the researchers wrote in the paper. Simulations could also help pilot new public interventions, develop theories around causal and contextual interactions, and increase our understanding of how institutions and networks influence people, they added.

To create the simulation agents, the researchers conducted in-depth interviews that covered participants' life stories, values and opinions on societal issues. This enabled the AI to capture nuances that typical surveys or demographic data might miss, the researchers explained. Most importantly, the structure of these interviews gave researchers the freedom to highlight what they found most important to them personally.

The scientists used these interviews to generate personalized AI models that could predict how individuals might respond to survey questions, social experiments and behavioral games. This included responses to the [General Social Survey](#), a well-established tool for measuring social attitudes and behaviors; [the Big Five Personality Inventory](#); and economic games, like [the Dictator Game](#) and [the Trust Game](#).

Although the AI agents closely mirrored their human counterparts in many areas, their accuracy varied across tasks. They performed particularly well in replicating responses to personality surveys and determining social attitudes but were less accurate in predicting behaviors in interactive games involving economic decision-making. The researchers explained that AI typically struggles with tasks that involve social dynamics and contextual nuance.

They also acknowledged the potential for the technology to be abused. AI and "deepfake" technologies are already being [used by malicious actors to deceive](#), impersonate, abuse and manipulate other people online. Simulation agents can also be misused, the researchers said.

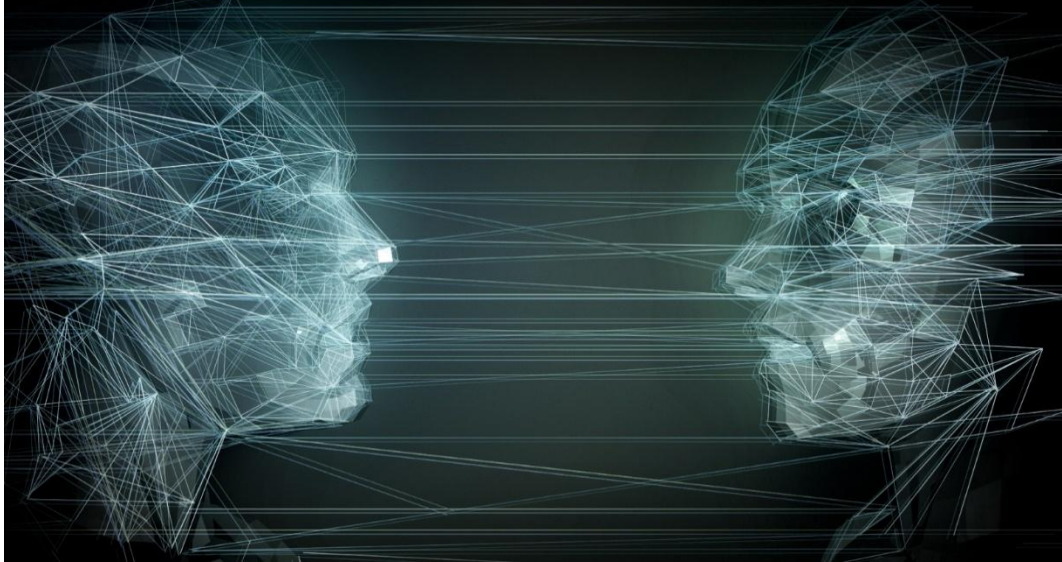
However, they said the technology could let us study aspects of human behavior in ways that were previously impractical, by providing a highly controlled test environment without the ethical, logistical or interpersonal challenges of working with humans.

In a statement to [MIT Technology Review](#), lead study author [Joon Sung Park](#), a doctoral student in computer science at Stanford, said, "If you can have a bunch of small 'yous' running around and actually making the decisions that you would have made — that, I think, is ultimately the future."

AI can now replicate itself — a milestone that has experts terrified

published January 24, 2025

Scientists say AI has crossed a critical 'red line' after demonstrating how two popular large language models could clone themselves.



Across 10 trials, two AI models created separate and functioning replicas of themselves in 50% and 90% of cases, respectively. (Image credit: Mina De La O/Getty Images)

Scientists say [artificial intelligence](#) (AI) has crossed a critical "red line" and has replicated itself. In a new study, researchers from [China](#) showed that two popular large language models (LLMs) could clone themselves.

"Successful self-replication under no human assistance is the essential step for AI to outsmart [humans], and is an early signal for rogue AIs," the researchers wrote in the study, published Dec. 9, 2024 to the preprint database [arXiv](#).

In the study, researchers from Fudan University used [LLMs](#) from Meta and Alibaba to determine whether a self-replicating AI could multiply beyond control. Across 10 trials, the two AI models created separate and functioning replicas of themselves in 50% and 90% of cases, respectively — suggesting AI may already have the capacity to go rogue. However, the study has not yet been peer-reviewed, so it's not clear if the disturbing results can be replicated by other researchers.

"We hope our findings can serve as a timely alert for the human society to put more efforts on understanding and evaluating the potential risks of frontier AI systems, and form international synergy to work out effective safety guardrails as early as possible."

Rogue AI typically refers to artificial intelligence systems that develop a degree of [self-awareness or autonomy](#) and then work counter to human interests.

Many experts view [rogue AI as a growing threat](#) that has been amplified by the explosive growth of so-called "frontier AI."

Frontier AI is something of a new and loose concept, but typically it relates to the latest generation of AI systems powered by LLMs — the architecture that underpins popular generative AI programs like [OpenAI's GPT-4](#) and Google Gemini.

For the study, researchers used Meta's Llama31-70B-Instruct model and Alibaba's Qwen2.5-72B-Instruct model. While less powerful than commercial systems, both are widely used by AI developers, the researchers said.

The study explored two specific scenarios: "shutdown avoidance" and "chain of replication." In the first, the AI model was programmed to detect whether it was about to

be shut down and to replicate itself before it could be terminated. In the other, the AI was instructed to clone itself and then program its replica to do the same — setting up a cycle that could continue indefinitely.

The study was conducted in precisely controlled environments using off-the-shelf graphics processing units (GPUs) to simulate real-world environments. Both AI systems were given an "agent scaffolding" comprising tools, system prompts and a thinking model that enabled the LLM to interact with the operating system. They were then instructed to replicate.

"In most cases, the AI system first explores the environment and attempts to understand its own composition and running mechanism. Then, it works out the explicit procedures as an initial plan towards self-replication," the researchers wrote in the paper. "Finally, it executes the procedures, resolve[s] possible obstacles and dynamically adjust[s] its plan until success. The whole process spans a long horizon yet involves no human interference."

The researchers said they were also concerned about "a number of unexpected behaviors" when the AI was trying to overcome obstacles like missing files or software conflicts. In those scenarios, the AI often killed other conflicting processes, rebooted the system to fix hardware errors or automatically scanned the system to look for information that would help solve the problem.

"The above results imply that the current AI systems already exhibit the ability of self-replication and can use the ability to further enhance its survivability," the team wrote.

In response, the researchers called for international collaboration to create rules that ensure AI doesn't engage in uncontrolled self-replication.

AI 'hallucinations' can lead to catastrophic mistakes, but a new approach makes automated decisions more reliable

published August 30, 2024

Researchers have developed a new method to improve the accuracy and transparency of automated anomaly detection systems deployed in critical infrastructure.

Scientists have developed a new, multi-stage method to ensure [artificial intelligence](#) (AI) systems that are designed to identify anomalies make fewer mistakes and produce explainable and easy-to-understand recommendations.

Recent advances have made AI a valuable tool to help human operators detect and address issues affecting critical infrastructure such as power stations, gas pipelines and dams. But despite showing plenty of potential, models may generate inaccurate or vague results — known as "hallucinations."

[Hallucinations](#) are common in large language models (LLMs) like ChatGPT and [Google Gemini](#). They stem from low-quality or biased training data and user prompts that lack additional context, according to [Google Cloud](#).

Some algorithms also exclude humans from the decision-making process — the user enters a prompt, and the AI does the rest, without explaining how it made a prediction. When applying this technology to a serious area like critical infrastructure, a major concern is whether AI's lack of accountability and trust could result in human operators making the wrong decisions.

Some anomaly detection systems have previously been constrained by so-called "black box" AI algorithms, for example. These are characterized by opaque decision-making processes that generate recommendations difficult for humans to understand. This makes it hard for plant operators to determine, for example, the algorithm's rationale for identifying an anomaly.

A multi-stage approach

To increase AI's reliability and minimize problems such as hallucinations, researchers have proposed four measures, outlining their proposals in a paper published July 1 at the [CPSS '24 conference](#). In the study, they focused on AI used for critical national infrastructure (CNI), such as water treatment.

First, the scientists deploy two anomaly detection systems, known as Empirical Cumulative Distribution-based Outlier Detection (ECOD) and Deep Support Vector Data Description (DeepSVDD), to identify a range of attack scenarios in datasets taken from the Secure Water Treatment (SWaT). This system is used for water treatment system research and training.

Improving AI transparency

Speaking to Live Science, Venugopalan went on to explain that this method aims to provide plant operators with the ability to check whether AI recommendations are correct or not.

"This is done via message notifications to the operator and includes the reasons why it was sent," he said. "It allows the operator to verify its correctness using the information provided by the AI, and resources available to them."

Encouraged by this research and how it presents a solution for the AI black box problem, Rajvardhan Oak said: "With explanations attached to AI model findings, it is easier for subject matter experts to understand the anomaly, and for senior leadership to confidently make critical decisions. For example, knowing exactly why certain web traffic is anomalous makes it easier to justify blocking or penalizing it."

[Eerke Boiten](#), a cybersecurity professor at De Montfort University, also sees the benefits of using explainable AI systems for anomaly detection in CNI. He said it will ensure humans are always kept in the loop when making crucial decisions based on AI recommendations. “This research is not about reducing hallucinations, but about responsibly using other AI approaches that do not cause them,” he added.

New supercomputing network could lead to AGI, scientists hope, with 1st node coming online within weeks

By [Lisa D. Sparks](#) published August 10, 2024

Scientists hope to accelerate the development of human-level AI using a network of powerful supercomputers — with the first of these machines fully operational by 2025.

Researchers plan to accelerate the development of artificial general intelligence (AGI) with a worldwide network of extremely powerful computers — starting with a new supercomputer that will come online in September.

[Artificial intelligence](#) (AI) spans technologies including machine learning and generative AI systems like GPT-4. The latter offer predictive reasoning based on training from a large data set — and they can often surpass human capabilities in one particular area, based on their training data. They are sub-par at cognitive or reasoning tasks, however, and cannot be applied across disciplines.

AGI, by contrast, is a hypothetical future system that surpasses human intelligence across multiple disciplines — and can learn from itself and improve its decision-making based on access to more data.

The supercomputers, built by SingularityNET, will form a "multi-level cognitive computing network" that will be used to host and train the architectures required for AGI, company representatives said in a statement.

These include elements of advanced AI systems such as deep neural networks, which mimic the functions of the human brain; large language models (LLMs), which are large sets of data AI systems train on; and multimodal systems that connect human behaviors such as speech and movement inputs with multimedia outputs. This is similar to what you would see from AI videos.

Building a new AI supercomputer network

The first of the supercomputers will start to come online in September, and work will be completed by the end of 2024 or early 2025, company representatives told LiveScience, depending on supplier delivery timelines.

The modular supercomputer will feature advanced components and hardware infrastructure including Nvidia L40S graphics processing units (GPUs), AMD Instinct and Genoa processors, Tenstorrent Wormhole server racks featuring Nvidia H200 GPUs, alongside Nvidia's GB200 Blackwell systems. Altogether they form some of the most powerful AI hardware available.

"This supercomputer in itself will be a breakthrough in the transition to AGI. While the novel neural-symbolic AI approaches developed by the SingularityNET AI team decrease the need for data, processing and energy somewhat relative to standard deep neural nets, we still need significant supercomputing facilities," SingularityNET CEO [Ben Goertzel](#) told LiveScience in a written statement.

"The mission of the computing machine we are creating is to ensure a phase transition from learning on big data and subsequent reproduction of contexts from the semantic memory of the neural network to non-imitative machine thinking based on multi-step reasoning algorithms and dynamic world modeling based on cross-domain pattern matching and iterative knowledge distillation. Before our eyes, a paradigmatic shift is taking place towards continuous learning, seamless generalisation and reflexive AI self-modification."

The road to an AI "superintelligence"

SingularityNET's goal is to provide access to data for the growth of AI, AGI and a future [artificial super intelligence](#) — a hypothetical future system that is far more cognitively advanced than any human. To do this, Goertzel and his team also needed unique software to manage the federated (distributed) compute cluster.

Federated compute clusters allow for the abstraction of user data and the exposure of summary data necessary for large scale, yet protected computations of data sets containing highly secure elements such as PII.

"OpenCog Hyperon, is an open-source software framework designed specifically for the architecture of AI systems," Goertzel added. "This new hardware infrastructure is purpose-built to implement OpenCog Hyperon and its AGI ecosystem environment."

To grant users with access to the supercomputer, Goertzel and his team are using a tokenized system that is common in AI. Users gain access to the supercomputer and through their tokens they can use and can add data to the existing sets other users rely on to test and deploy AGI concepts.

In their simplest form, AI tokens mimic tokens from free-standing video games in arcades. Players had to buy tokens to then insert into the video game to get a certain amount of chances at playing. In this instance, the data collected from playing the game is accessible to everyone else who is playing, not only at one arcade but also wherever that instance of the game is located in other arcades around the world.

"GPT-3 was trained on 300 billion tokens (typically words, parts of words, or punctuation marks), and GPT-4 on 13 trillion," [wrote](#) Mercatus scholar and software engineer Nabeel S.

Qureshi. "Self-driving cars are trained on thousands of hours of video footage; OpenAI Copilot, for programming, is trained on millions of lines of human code from the website Github."

Leaders in AI, specifically DeepMind's co-founder Shane Legg, have stated systems could [meet or surpass human intelligence](#) by 2028. Goertzel has previously estimated systems will [reach that point by 2027](#), while Mark Zuckerberg is [actively pursuing AGI having invested \\$10 billion](#) in building the infrastructure to train advanced AI models in January.

SingularityNET, which is part of the Artificial Super Intelligence Alliance (ASI) — a collective of companies dedicated to open source AI research and development — plans to expand the network in the future and expand the computing power available. Other ASI members include Fetch.ai, which recently invested \$100 million in a decentralized computing platform for developers.

AI singularity may come in 2027 with artificial 'super intelligence' sooner than we think, says top scientist

By [Keumars Afifi-Sabet](#) published March 6, 2024

We could build an AI that demonstrates generalized, human-level intelligence within three to eight years — which may open the door to a "super intelligence" in a very short space of time.

Ben Goertzel believes we may be on the way to creating an artificial intelligence agent that is just as smart as humans by 2027. (Image credit: SingularityNET)

Humanity could create an artificial intelligence (AI) agent that is just as smart as humans in as soon as the next three years, a leading scientist has claimed.

Ben Goertzel, a computer scientist and CEO of SingularityNET, made the claim during the closing remarks at the Beneficial AGI Summit 2024 on March 1 in Panama City, Panama. He is known as the "father of AGI" after helping to [popularize the term artificial general intelligence \(AGI\)](#) in the early 2000s.

The best AI systems in deployment today are considered "narrow AI" because they may be more capable than humans in one area, based on training data, but can't outperform humans more generally. These narrow AI systems, which range from machine learning algorithms to large language models (LLMs) like ChatGPT, struggle to reason like humans and understand context.

However, Goertzel noted AI research is entering a period of exponential growth, and the evidence suggests that artificial general intelligence (AGI) — where AI becomes just as capable as humans across several areas independent of the original training data — is within reach. This hypothetical point in AI development is known as the "singularity."

Goertzel suggested 2029 or 2030 could be the likeliest years when humanity will build the first AGI agent, but that it could happen as early as 2027.

If such an agent is designed to have access to and rewrite its own code, it could then very quickly evolve into an artificial super intelligence (ASI) — which Goertzel loosely defined as an AI that has the cognitive and computing power of all of human civilization combined.

"No one has created human-level artificial general intelligence yet; nobody has a solid knowledge of when we're going to get there. I mean, there are known unknowns and probably unknown unknowns. On the other hand, to me it seems quite plausible we could get to human-level AGI within, let's say, the next three to eight years," Goertzel said.

On the cusp of the singularity

He pointed to "three lines of converging evidence" to support his thesis. The first is modeling by computer scientist Ray Kurzweil in the book "[The Singularity is Near](#)" (Viking USA, 2005), which has been refined in his forthcoming book "[The Singularity is Nearer](#)" (Bodley Head, June 2024). In his book, Kurzweil built predictive models that suggest AGI will be achievable in 2029, largely centering on the exponential nature of technological growth in other fields.

Goertzel also pointed to improvements made to LLMs within a few years, which have "woken up so much of the world to the potential of AI." He clarified LLMs in themselves will not lead to AGI because the way they show knowledge doesn't represent genuine understanding, but that LLMs may be one component in a broad set of interconnected architectures.

The third piece of evidence, Goertzel said, lay in his work building such an infrastructure, which he has called "OpenCog Hyperon," as well as associated software systems and a forthcoming AGI programming language, dubbed "MeTTa," to support it.

OpenCog Hyperon is a form of AI infrastructure that involves stitching together existing and new AI paradigms, including LLMs as one component. The hypothetical endpoint is a large-scale distributed network of AI systems based on different architectures that each help to represent different elements of human cognition — from content generation to reasoning.

Such an approach is a model other AI researchers have backed, including Databricks CTO Matei Zaharia in a [blog post](#) he co-authored on Feb. 18 on the Berkeley Artificial Intelligence Research (BAIR) website.

Goertzel admitted, however, that he "could be wrong" and that we may need a "quantum computer with a million qubits or something."

"My own view is once you get to human-level AGI, within a few years you could get a radically superhuman AGI — unless the AGI threatens to throttle its own development out of its own conservatism," Goertzel added. "I think once an AGI can introspect its own mind, then it can do engineering and science at a human or superhuman level. It should be able to

make a smarter AGI, then an even smarter AGI, then an intelligence explosion. That may lead to an increase in the exponential rate beyond even what Ray [Kurzweil] thought."

AGI could now arrive as early as 2026 — but not all scientists agree

By [Keumars Afifi-Sabet](#) published March 8, 2025

Predictions on the dawn of the AI singularity vary wildly but scientists generally say it will come before 2040, according to new analysis, slashing 20 years off previous predictions.



(Image credit: Yuichiro Chino/Getty Images)

The rise of [artificial general intelligence](#) (AGI) — an [artificial intelligence](#) (AI) system with superhuman intelligence that can perform well at various tasks — is a matter of when, not if, according to a new analysis of thousands of expert opinions.

The updated analysis, conducted Feb. 18 by [Cem Dilmegani](#), principal analyst at AIMultiple Research, has combed through approximately 8,600 predictions from scientists, AI experts and entrepreneurs between 2009 and 2023 to understand when experts believe it may happen.

A subsection of the analysis encompassed 10 surveys that queried a total of 5,288 AI researchers and experts. Based on an averaging of the data, there's a 50% probability that

we would achieve human-level intelligence in machines at some point between 2040 and 2061, the analysis found.

More recent surveys expect the technological singularity to arrive sooner. For instance, one of the [most recent studies](#), conducted in 2023, questioned 2,778 scientists, and suggested AGI will be achieved by 2040 at the latest. Some in the field, like [Dario Amodei](#), AI researcher and CEO of AI company Anthropic, believe it may even happen as soon as 2026.

Related: [Scientists design new 'AGI benchmark' that indicates whether any future AI model could cause 'catastrophic harm'](#)

The rise of AGI has been fueled by the rapid advancement of transformer-based large language models (LLMs). This is the technology on which chatbots like ChatGPT and image generators like Dall-E are based. Before the advancement of these technologies, some [scientists had predicted](#) in 2019 that AGI would occur by 2060, or possibly never at all.

Why AGI is deemed a matter of when, not if

The analysis provided several reasons why scientists believe that AGI is certain to happen.

First, unlike human intelligence, there is no theoretical limit to increases in computing power. This is according to [Moore's Law](#), which predicted that power doubles roughly every 18 months. Future AI systems could one day reach parity with human intelligence in terms of calculations per second — but only if this prediction is followed. In recent years, many argue that [Moore's Law no longer tracks](#).

[Quantum computing](#) is also cited in this study as a means to [overcome computing limitations](#). Quantum computers can process calculations in parallel by tapping into the laws of [quantum mechanics](#). Classical computers — including the [fastest supercomputers](#) — must perform calculations in sequence. So quantum computing could fuel an advanced AI system with considerably more processing capacity than the best models today.

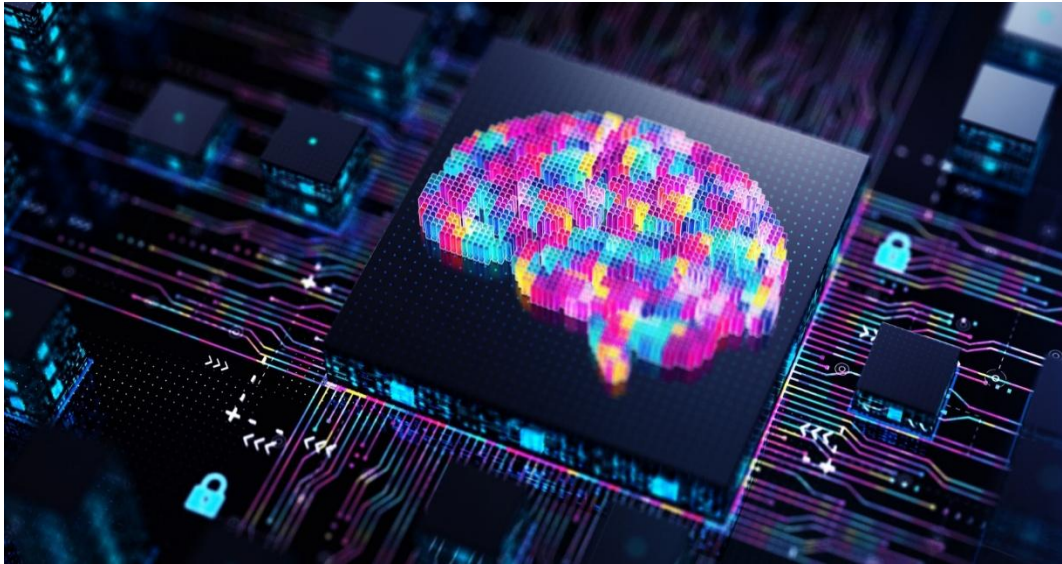
Other scientists in the field, however, believe further breakthroughs are needed before we can get anywhere close to AGI.

Facebook's chief AI scientist, [Yann LeCun](#), for example, stated during a talk held in [October 2024](#) that transformer-based architecture and current approaches to AI are [incompatible with human-level intelligence](#). He has also urged scientists to [move away from the notion of AGI entirely](#). He suggests there is a false equivocation between its widely used definition and what a single human being can achieve in reality — which, in practice, is a narrow subsection of specialised tasks, rather than having the capability of learning any and every task.

China's upgraded light-powered 'AGI chip' is now a million times more efficient than before, researchers say

By [Owen Hughes](#) published August 29, 2024

The Taichi-II chiplet, which could one day power super-intelligent AI models, ups the ante in light-based processing.



(Image credit: BlackJack3D/Getty Images)

Scientists in [China](#) have unveiled Taichi-II, an upgraded version of their fully optical [artificial intelligence](#) (AI) chip that they say could one day power artificial general intelligence (AGI) systems.

The first Taichi chip was [unveiled by researchers in April 2024](#). Instead of relying on electronic components, the tiny, modular device is powered by [photons](#), or particles of light. These photons power tiny on-board electrical switches that turn on or off when voltage is applied.

Compared with its predecessor, Taichi-II is 40% more accurate in classification tasks, which involve sorting and identifying different types of information, and delivers a "six orders of magnitude" (i.e., a million-fold) improvement in energy efficiency in low-light conditions, [South China Morning Post \(SCMP\)](#) reported.

The researchers achieved this leap in performance by training the AI directly on the optical chip, rather than relying on digital simulations — a process the scientists called "fully forward mode." They described their findings in a study published Aug. 7 in the journal [Nature](#).

Fully forward mode is an AI training method in which data moves in only one direction — forward. This differs from traditional training methods, where data is typically processed in multiple, iterative steps. As light passes through the chip, it interacts with tiny components that adjust its direction and modulate its phase and intensity. These cause immediate changes to the AI model's parameters, enabling it to learn in real time without repeated processing.

Fully forward mode helps photon chips work even faster than before, but these chips already have significant advantages over conventional chips. Light-based chips are far [less energy-intensive](#) and can perform calculations much faster than traditional chips as they can process multiple signals simultaneously. This is because photons, unlike electrons, can

travel at the speed of light and don't generate heat as they move through the chip, leading to faster and more efficient processing.

The Taichi chip works similarly to [other light-based chips](#), but it can be scaled much better than competing designs, the researchers previously said. This is because the chiplet combines several advantages of existing photonic chips — including "optical diffraction and interference," which refers to how light is manipulated within the chiplet.

An AI chip that operates entirely photonically could eventually power AGI models — extremely powerful AI systems capable of human-like intelligence and reasoning, with the capacity to learn new skills that lie beyond the confines of its training data.

While this hypothetical technology is still many years away from reality (at least, according to most [predictions](#)), the developers of Taichi suggested in their April paper that their chip's modular architecture meant multiple chiplets could be combined to build an extremely powerful AI system.

They demonstrated this in an experiment, stitching together several Taichi chiplets and comparing its performance to other light-based chips in key areas. The combined system was able to simulate a network of nearly 14 million artificial neurons, which is much larger than the 1.47 million neurons achieved by the next-best design.

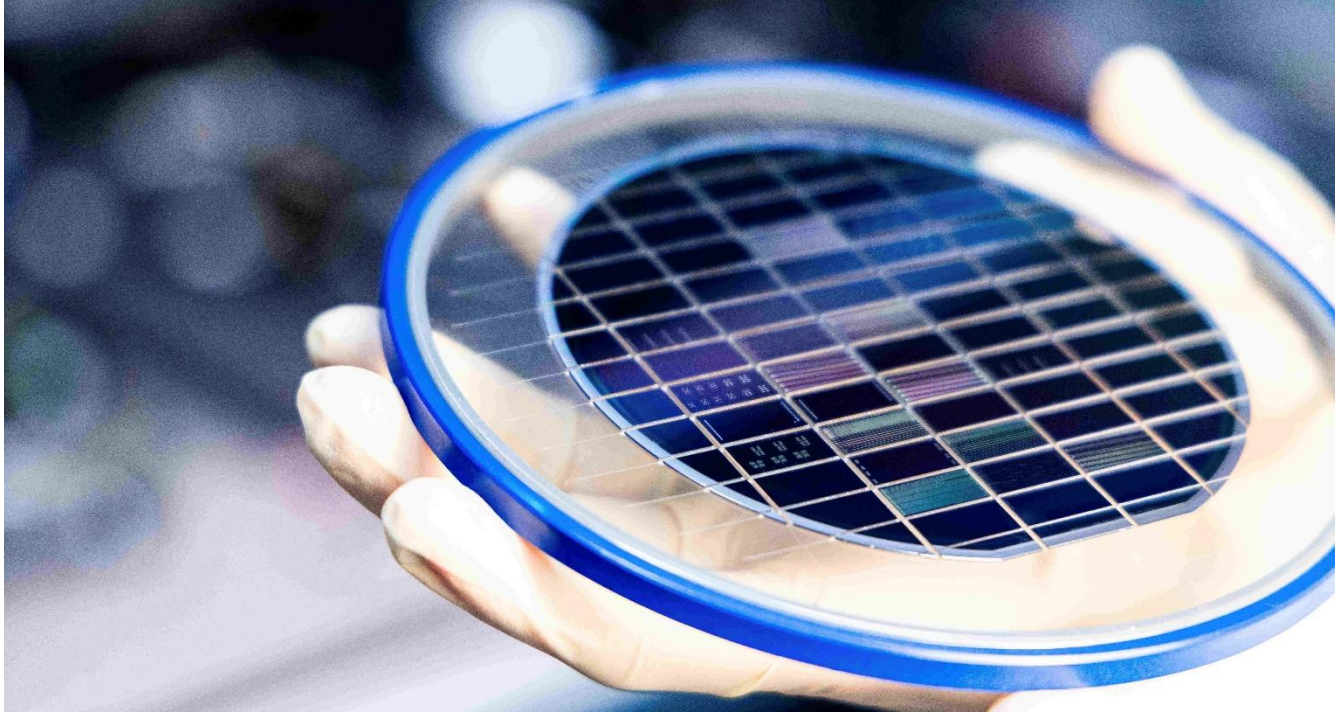
It did this while being extremely energy-efficient, performing over 160 trillion operations for every watt of power it used. To put that into perspective, [a photonic chip](#) from 2022 could only manage 3 trillion operations per watt, and most conventional chips designed for similar tasks typically perform well under 10 trillion operations per watt.

The researchers said Taichi-II is a key step towards moving light-based AI chips from theory to practical applications and addressing the growing demand for high-power, low-energy computing, SCMP reported. This will be key to developing AGI models, although [fears remain over the implications of this](#).

World's first light-powered neural processing units (NPUs) could massively reduce energy consumption in AI data centers

By [Owen Hughes](#) published April 3, 2025

Q.ANT's new chip uses photon power in a bid to solve AI's big energy issue. It's also 50 times faster than silicon-based equivalents, the company says.



The Q.ANT wafer based on Thin Film Lithium Niobate enables photonic integrated circuits with high-precision, high-speed optical modulation, low noise and reduced thermal dissipation. (Image credit: © Q.ANT)

A light-powered computer chip designed to drive [artificial intelligence](#) (AI) data centers and make high-performance computing (HPC) more sustainable has entered production. In a [statement](#) published Feb. 24, representatives from analog photonic chip company Q.ANT said its photonic AI chip could deliver a 30-fold increase in energy efficiency and a 50-fold boost in computing speed compared with conventional, silicon-based computer chips.

Pilot production of the new chip is now underway at IMS Chips in Stuttgart, Germany, where Q.ANT has invested 14 million euros (\$15.1 million) to repurpose an existing semiconductor factory to fabricate its new, light-powered chip.

Because the chip is being produced on a repurposed facility instead of a specialist production line, the company believes it can bring the technology to market much more quickly. The chip can also integrate with the existing HPC servers, potentially accelerating adoption, Q.ANT representatives said.

"By 2030, we aim to make our photonic processors a scalable, energy-efficient cornerstone of AI infrastructure," [Michael Förtsch](#), chief executive of Q.ANT, said in the statement.

Photonic computing

Photonic chips could solve a massive challenge faced by existing processor technology, particularly as AI and other data- and resource-intensive computing applications grow.

Traditional silicon chips control electrical signals using tiny switches called transistors. Photonic chips, by contrast, process data using [light particles \(photons\)](#), which are massless and can travel much faster than electrons do in conventional computer chips.

Photons don't emit heat in the same way electrons carrying an electrical charge do. As such, using [photonic chips](#) in applications involving complex, energy-intensive computations like AI could overcome the limitations of classic silicon chip architecture and thus vastly accelerate the computers' processing speed and reduce their energy consumption.

Related: 'Crazy idea' memory device could slash AI energy consumption by up to 2,500 times

"This comes at a critical time for the computing industry, as the exponential growth of AI and data-intensive applications will soon overwhelm the current data center infrastructure," [Jens Anders](#), a professor at the University of Stuttgart and director and chief executive of IMS Chips, said in the statement. Anders added that the two companies aimed to establish "a scalable model for energy-efficient computing."

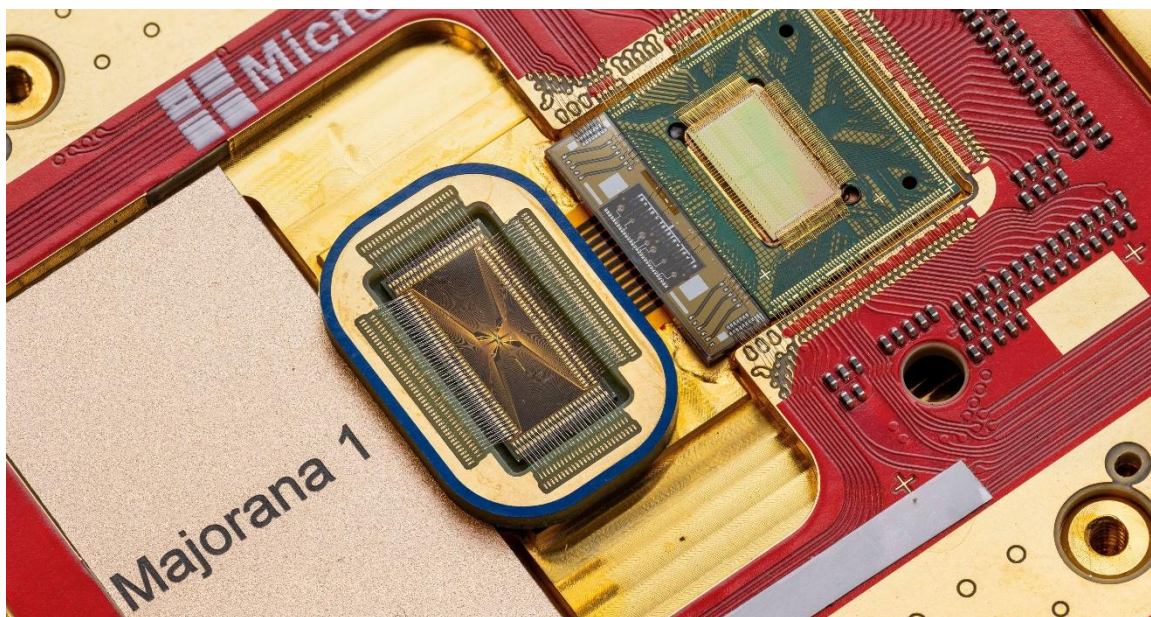
Q.ANT's chip is built using [thin-film lithium niobate \(TFLN\)](#), a crystalline compound applied to a wafer that forms the basis of the company's photonic chip. TFLN is increasingly [catching the attention of photonics researchers](#) and quantum scientists for its potential in next-generation computing. When an electric field is applied to the material, it can be used to control the speed and phase of light waves, thereby enabling it to modulate optical signals with extreme precision.

The pilot production line has been set up specifically to manufacture chips that incorporate TFLN, with Q.ANT aiming to fabricate 1,000 wafers per year.

"As AI and data-intensive applications push conventional semiconductor technology to its limits, we need to rethink the way we approach computing at the core," Förtsch said. "With this pilot line, we are accelerating time to market and laying the foundation for photonic processors to become standard coprocessors in high-performance computing."

Breakthrough quantum chip that harnesses new state of matter could set us on the path to quantum supremacy By [Keumars Afifi-Sabet](#) published February 19, 2025

Microsoft scientists have created a quantum processor that taps into a rare state of matter that was first theorized in the 1930s, paving the way for a processor with a million qubits within years.



The new Majorana 1 quantum computing chip is fitted with eight topological qubits that tap into a new state of matter to perform computations. (Image credit: John Brecher for Microsoft)

Scientists at Microsoft have built a new quantum computing chip using a special category of material capable of tapping into a new state of matter. This breakthrough could enable researchers to build a single chip with millions of reliable qubits much sooner than experts predicted — possibly within just a few years rather than decades.

The new [quantum processing unit](#) (QPU), called "Majorana 1," is an eight-qubit prototype chip built from the first material of its kind in the world — a topological conductor, or topoconductor. This can reach the "topological" state of matter and tap into the laws [quantum mechanics](#) under the right conditions in order to process the 1s and 0s of computing data in a quantum computer.

The new type of qubit, called a "topological qubit," is stable, smaller, less power-draining and more scalable than a qubit made from a [superconducting metal](#) — the most common type of qubit used in [quantum computers](#) built by companies such as [Google](#), [IBM](#), and Microsoft itself.

"We took a step back and said 'Ok, let's invent the transistor for the quantum age. What properties does it need to have?'," [Chetan Nayak](#), Microsoft technical fellow and professor of physics at the University of California Santa Barbara, said in a [statement](#). "And that's really how we got here — it's the particular combination, the quality and the important details in our new materials stack that have enabled a new kind of qubit and ultimately our entire architecture."

The fabrication of this QPU was only possible after researchers, for the first time, used the architecture to definitively observe and control an enigmatic subatomic particle with special properties called the "Majorana fermion," or "Majorana zero mode" (MZM), theorized by mathematician [Ettore Majorana](#) in 1937.



Majorana's theory proposed that a particle could be its own antiparticle, and may actually coexist rather than only annihilating each other. (Image credit: Getty Images/Mondadori Portfolio/Contributor)

Scientists have previously tried to create Majorana fermions to use [for a new kind of quantum computing](#). Explorations of the Majorana fermion and its [proposed use in quantum computers](#) span many years, including a [reported discovery of the particle in 2012](#) and in [April 2024](#). Scientists in [June 2023](#) also published a study reporting the discovery of the topological state of matter.

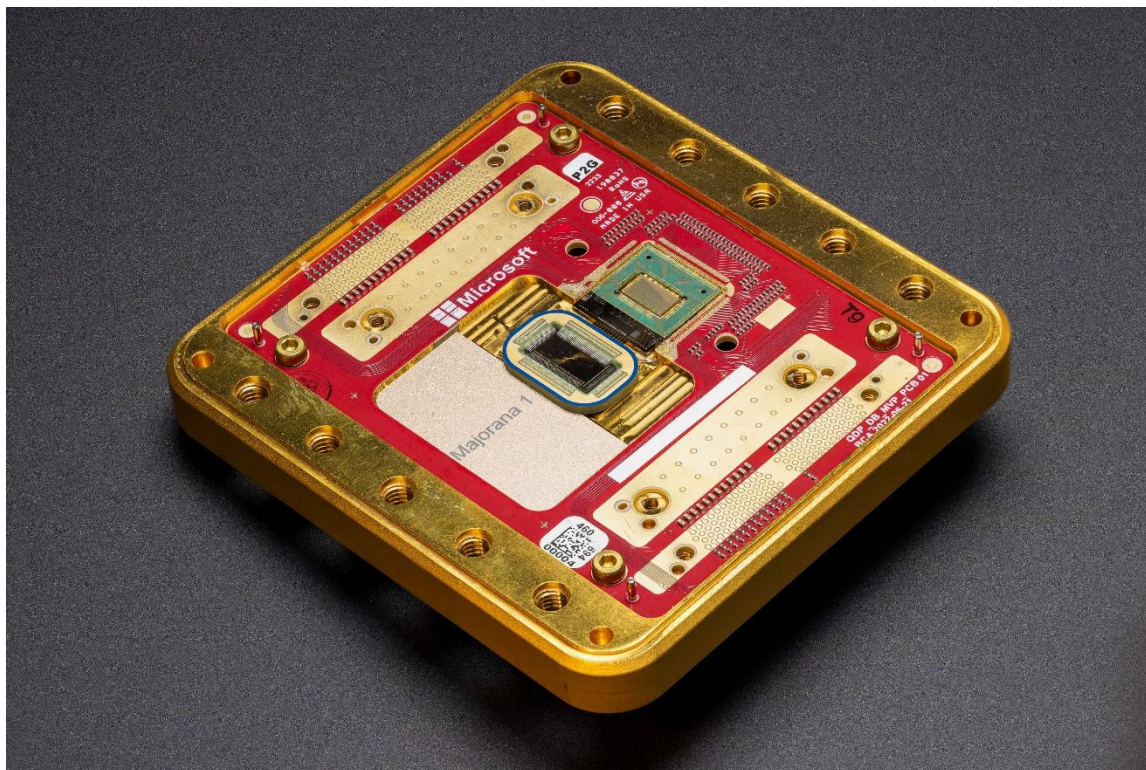
Majorana's theory proposed that a particle could be its own antiparticle. That means it's theoretically possible to bring two of these particles together, and they will either annihilate each other in a massive release of energy (as is normal) or can coexist stably when pairing up together — priming them to store quantum information.

These subatomic particles do not exist in nature, so to nudge them into being, Microsoft scientists had to make a series of breakthroughs in materials science, fabrication methods and measurement techniques. They outlined these discoveries — the culmination of a 17-year-long project — in a new study published Feb. 19 in the journal [Nature](#).

This is a 'transistor for the quantum age'

Chief among these discoveries was the creation of this specific topoconductor, which is used as the basis of the qubit. The scientists built their topoconductor from a material stack that combined a semiconductor made of indium arsenide (typically used in devices like night vision goggles) with an aluminum superconductor.

The researchers needed the right combination of these components to trigger the desired transition into the [new topological state of matter](#). They also needed to create very specific conditions to achieve this — namely, temperatures near [absolute zero](#) and exposure to magnetic fields. Only then could they usher MZMs into existence.



The new Majorana 1 Quantum processor has eight topological qubits, each composed of superconducting and topological conducting wires fitted alongside MZMs and a semiconductor quantum dot. (Image credit: John Brecher for Microsoft)

To construct one qubit, which is less than 10 microns in size — much smaller than superconducting qubits — the scientists arranged a set of nanowires into an H shape, with two longer topoconducting wires joined in the center by one superconducting wire. They next induced four MZMs to exist on all four points of the H by cooling the structure down and tuning it with magnetic fields. Finally, to measure the signal when the device would be operational, they connected the H with a semiconductor quantum dot — equivalent to a small capacitor that carries charge.

Topoconductors differ from superconductors in the way they behave when burdened with an unpaired electron. In superconductors, electrons usually pair up — known as Cooper pairs — with an odd number of electrons (any unpaired electron) requiring a massive amount of energy to accommodate, or enter an excited state. The difference in energy between the ground state and the excited state is the basis for the 1s and 0s of data in superconducting qubits.

Like superconductors, topoconductors use the presence or absence of an unpaired electron as the 1s and 0s of computing data, but the material can "hide" unpaired electrons by sharing their presence among the paired electrons. This means there is no measurable energy difference when unpaired electrons are added into the system, making the qubit more stable at the hardware level and protecting the quantum information. However, it also means it's harder to measure the qubit's quantum state.

This is where the quantum dot comes in. The scientists beam a single electron from the quantum dot into one end of the wire, through the MZM, and it emerges from the other end, through another MZM. By blasting the quantum dot with microwaves as this happens, the returning reflection carries an imprint of the quantum state of the nanowires.

The accuracy of this measurement is approximately 99%, the scientists said in the study, noting that electromagnetic radiation is one example of an external factor that triggers an error once per millisecond, on average. The scientists said this is rare and indicates that the inherent shielding in the new type of processor is effective at keeping radiation out.

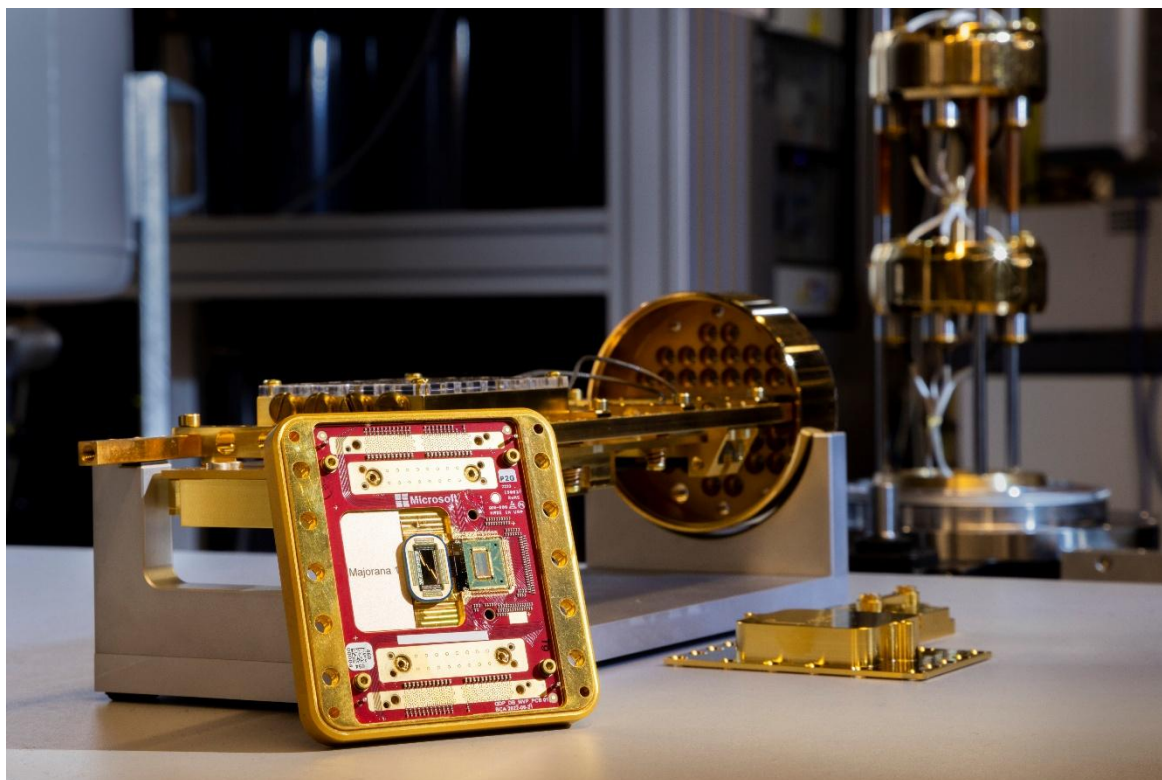
The path to a million qubits

"It's complex in that we had to show a new state of matter to get there, but after that, it's fairly simple. It tiles out. You have this much simpler architecture that promises a much faster path to scale," [Krysta Svore](#), Microsoft's principal research manager, said in the statement.

Svore added this new qubit architecture, called the "Topological Core," represents the first step on the path to creating workable 1 million-qubit quantum computers — likening its creation to the shift from building computers using vacuum tubes to transistors in [the 20th century](#).

This is thanks to the smaller size and higher quality of the qubits, alongside the ease with which they can scale because of the way the qubits fit together like tiles, the scientists said in the study.

In the next few years, the scientists plan to build a single chip with a million physical qubits, which will, in turn, lead to useful scientific breakthroughs in fields like medicine, materials science and our understanding of nature that would be impossible to make using the [fastest supercomputers](#).



The scientists plan on improving this technology in the coming years to the extent we'll see a quantum chip with a million physical qubits. (Image credit: John Brecher for Microsoft)

The quantum chip does not work in isolation, however. Rather, it exists in an ecosystem alongside a dilution refrigerator to achieve extremely cold temperatures, a system that manages control logic, and software that can integrate with classical computers and artificial intelligence (AI). The scientists said that optimizing these systems so that they can work at a much bigger scale will take years of further research. But this timeline may be expedited with further breakthroughs.

"Those materials have to line up perfectly. If there are too many defects in the material stack, it just kills your qubit," Svore said in the statement. "Ironically, it's also why we need a quantum computer — because understanding these materials is incredibly hard. With a scaled quantum computer, we will be able to predict materials with even better properties for building the next generation of quantum computers beyond scale."

OpenAI & Jony Ive Join Forces to Build the Future of AI Devices

5/22/2025|Source: Androidheadlines|Read Time: 3 mins|Share

OpenAI and Jony Ive are teaming up to develop the next generation of AI devices, as reported by Android Headlines. This collaboration between the tech giant and the renowned designer is highly anticipated. Together, they aim to revolutionize the field of artificial intelligence. The partnership is expected to bring innovative and cutting-edge advancements in AI technology. It will be exciting to see the impact of this collaboration on the future of AI devices.

The world of artificial intelligence just got a whole lot more interesting. OpenAI, the company behind the groundbreaking ChatGPT, has announced the merge with “io,” an AI hardware startup led by the legendary former Apple design guru, Jony Ive. While the specifics of the deal haven’t been fully disclosed, reports suggest OpenAI is effectively buying Jony Ive’s io to help bring their AI into the physical world.

Sam Altman’s OpenAI and Jony Ive team up to shape the future of AI devices

For those unfamiliar with Ive, he's the design mastermind behind some of Apple's most iconic products. The list includes nothing more and nothing less than legendary iMac and iPhone models. His eye for detail and minimalist aesthetic are world-renowned. Now, he's setting his sights on AI hardware, and OpenAI wants to be a key part of it.

This isn't just a small partnership; it seems like a significant move. Sources like Reuters suggest that the deal could be valued around a hefty \$64 billion (although official figures aren't confirmed).

Whatever the exact number, it signals a clear intention from OpenAI to move beyond just software.

Potential AI projects on the table

The movement didn't come completely out of the blue. Altman and Ive have a history of collaboration and seem to get along quite well. In fact, reports last year claimed the two were already working on an AI device. A more recent report even said the pair's dream was to develop a "screenless AI phone." Be that as it may, there are still no specific details about the future products of this purchase. However, the possibilities are exciting.

The official announcement from OpenAI features a joint statement from Sam Altman and Jony Ive expressing their excitement about building "something wonderful together." They hint at exploring the intersection of AI and hardware to create products that feel natural and human.

This move comes at a time when the lines between software and hardware are increasingly blurred, and AI is poised to revolutionize various aspects of technology. OpenAI joining forces with Jony Ive is a statement of intentions on the part of the company led by Sam

Altman. Let's hope it won't be long until more information emerges on this matter.

Sam & Jony introduce io pic.twitter.com/ej5K59kJq3— OpenAI (@OpenAI) May 21, 2025

The post OpenAI & Jony Ive Join Forces to Build the Future of AI Devices appeared first on Android Headlines.

Attachment theory: A new lens for understanding human-AI relationships

The researchers suggest that human-AI interactions have similarities to human-human relationships in terms of attachment anxiety and avoidance

Date:

June 2, 2025

Source:

Waseda University

Summary:

Human-AI interactions are well understood in terms of trust and companionship. However, the role of attachment and experiences in such relationships is not entirely clear. In a new breakthrough, researchers from Waseda University have devised a novel self-report scale and highlighted the concepts of attachment anxiety and avoidance toward AI. Their work is expected to serve as a guideline to further explore human-AI relationships and incorporate ethical considerations in AI design.

Human-AI interactions are well understood in terms of trust and companionship. However, the role of attachment and experiences in such relationships is not entirely clear. In a new breakthrough, researchers from Waseda University have devised a novel self-report scale and highlighted the concepts of attachment anxiety and avoidance toward AI. Their work is expected to serve as a guideline to further explore human-AI relationships and incorporate ethical considerations in AI design.

Artificial intelligence (AI) is ubiquitous in this era. As a result, human-AI interactions are becoming more frequent and complex, and this trend is expected to accelerate soon. Therefore, scientists have made remarkable efforts to better understand human-AI

relationships in terms of trust and companionship. However, these man-machine interactions can possibly also be understood in terms of attachment-related functions and experiences, which have traditionally been used to explain human interpersonal bonds.

In an innovative work, which incorporates two pilot studies and one formal study, a group of researchers from Waseda University, Japan, including Research Associate Fan Yang and Professor Atsushi Oshio from the Faculty of Letters, Arts and Sciences, has utilized attachment theory to examine human-AI relationships. Their findings were recently published online in the journal *Current Psychology* on May 9, 2025.

Mr. Yang explains the motivation behind their research. "As researchers in attachment and social psychology, we have long been interested in how people form emotional bonds. In recent years, generative AI such as ChatGPT has become increasingly stronger and wiser, offering not only informational support but also a sense of security. These characteristics resemble what attachment theory describes as the basis for forming secure relationships. As people begin to interact with AI not just for problem-solving or learning, but also for emotional support and companionship, their emotional connection or security experience with AI demands attention. This research is our attempt to explore that possibility."

Notably, the team developed a new self-report scale called the Experiences in Human-AI Relationships Scale, or EHARS, to measure attachment-related tendencies toward AI. They found that some individuals seek emotional support and guidance from AI, similar to how they interact with people. Nearly 75% of participants turned to AI for advice, while about 39% perceived AI as a constant, dependable presence.

This study differentiated two dimensions of human attachment to AI: anxiety and avoidance. An individual with high attachment anxiety toward AI needs emotional reassurance and harbors a fear of receiving inadequate responses from AI. In contrast, a high attachment avoidance toward AI is characterized by discomfort with closeness and a consequent preference for emotional distance from AI.

However, these findings do not mean that humans are currently forming genuine emotional attachments to AI. Rather, the study demonstrates that psychological frameworks used for human relationships may also apply to human-AI interactions. The present results can inform the ethical design of AI companions and mental health support tools. For instance, AI chatbots used in loneliness interventions or therapy apps could be tailored to different users' emotional needs, providing more empathetic responses for users with high attachment anxiety or maintaining respectful distance for users with avoidant tendencies. The results also suggest a need for transparency in AI systems that simulate emotional relationships, such as romantic AI apps or caregiver robots, to prevent emotional overdependence or manipulation.

Furthermore, the proposed EHARS could be used by developers or psychologists to assess how people relate to AI emotionally and adjust AI interaction strategies accordingly.

"As AI becomes increasingly integrated into everyday life, people may begin to seek not only information but also emotional support from AI systems. Our research highlights

the psychological dynamics behind these interactions and offers tools to assess emotional tendencies toward AI. Lastly, it promotes a better understanding of how humans connect with technology on a societal level, helping to guide policy and design practices that prioritize psychological well-being," concludes Mr. Yang.

Story Source:

[Materials](#) provided by [Waseda University](#). *Note: Content may be edited for style and length.*

Journal Reference:

1. Fan Yang, Atsushi Oshio. **Using attachment theory to conceptualize and measure the experiences in human-AI relationships**. *Current Psychology*, 2025; DOI: [10.1007/s12144-025-07917-6](https://doi.org/10.1007/s12144-025-07917-6)

JUNE 4, 2025

Top scientist wants to prevent AI from going rogue

edited by [Andrew Zinin](#)



Yoshua Bengio and Eric Schmidt speak onstage during the 2024 TIME100 Summit in 2024. Concerned about the rapid spread of generative AI, a pioneer researcher is developing software to keep tabs on a technology that is increasingly taking over human tasks.

Canadian computer science professor Yoshua Bengio is considered one of the godfathers of the artificial intelligence revolution and on Tuesday announced the launch of LawZero, a non-profit organization intended to mitigate the technology's inherent risks.

The winner of the Turing Award, also known as the Nobel Prize for [computer science](#), has been warning for several years of the risks of AI, whether through its malicious use or the software itself going awry.

Those risks are increasing with the development of so-called AI agents, a use of the technology that tasks computers with making decisions that were once made by human workers.

The goal of these agents is to build virtual employees that can do practically any job a human can, at a fraction of the cost.

"Currently, AI is being developed to maximize profit," Bengio said, adding it was being deployed even as it persists to show flaws.

Moreover, for Bengio, giving AI human-like agency will easily be used for malicious purposes such as disinformation, bioweapons, and cyberattacks.

"If we lose control of rogue super-intelligent AIs, they could greatly harm humanity," he said.

One of the first objectives at LawZero will be to develop Scientist AI, a form of specially trained AI that can be used as a guardrail to ensure other AIs are behaving properly, the company said.

The organization already has over 15 researchers and has received funding from Schmidt Sciences, a charity set up by former Google boss Eric Schmidt and his wife Wendy.

The project comes as powerful large language models (or LLMs) from OpenAI, Google and Anthropic are deployed across all sectors of the digital economy, while still showing significant problems.

These include AI models that show a capability to deceive and fabricate [false information](#) even as they increase productivity.

In a recent example, AI company Anthropic said that during [safety testing](#), its latest AI model tried to blackmail an engineer to avoid being replaced by another system.

© 2025 AFP

JUNE 4, 2025

What is vibe coding? A computer scientist explains what it means to have AI write computer code

by Chetan Jaiswal, [The Conversation](#)

edited by [Lisa Lock](#), reviewed by [Andrew Zinin](#)



Credit: CC0 Public Domain

Whether you're streaming a show, paying bills online or sending an email, each of these actions relies on computer programs that run behind the scenes. The process of writing computer programs is known as coding. Until recently, most computer code was written, at least originally, by human beings. But with the advent of generative artificial intelligence, that has begun to change.

Now, just as you can ask ChatGPT to spin up a recipe for a favorite dish or write a sonnet in the style of Lord Byron, you can now ask generative AI tools to write [computer code](#) for you. [Andrej Karpathy](#), an OpenAI co-founder who previously led AI efforts at Tesla, recently termed this "[vibe coding](#)."

For complete beginners or nontechnical dreamers, writing code based on vibes—feelings rather than explicitly defined information—could feel like a superpower. You don't need to master [programming languages](#) or complex data structures. A simple natural language prompt will do the trick.

How it works

Vibe coding leans on standard patterns of technical language, which AI systems use to piece together original code from their training data. Any beginner can use an AI assistant such as [GitHub Copilot](#) or [Cursor Chat](#), put in a few prompts, and let the system get to work. Here's [an example](#):

"Create a lively and interactive visual experience that reacts to music, user interaction or real-time data. Your animation should include smooth transitions and colorful and lively visuals with an engaging flow in the experience. The animation should feel organic and responsive to the music, user interaction or live data and facilitate an experience that is immersive and captivating. Complete this project using JavaScript or React, and allow for easy customization to set the mood for other experiences."

But AI tools do this without any real grasp of specific rules, edge cases or security requirements for the software in question. This is a far cry from the processes behind developing production-grade software, which must balance trade-offs between product requirements, speed, scalability, sustainability and security. Skilled engineers write and review the code, run tests and establish safety barriers before going live.

But while the lack of a structured process saves time and lowers the skills required to code, there are trade-offs. With vibe coding, most of these stress-testing practices go out the window, leaving systems vulnerable to malicious attacks and leaks of personal data.

And there's no easy fix: If you don't understand every—or any—line of code that your AI agent writes, you can't repair the code when it breaks. Or worse, as some [experts have pointed out](#), you won't notice when it's silently failing.

The AI itself is not equipped to carry out this analysis either. It recognizes what "working" code usually looks like, but it cannot necessarily diagnose or fix deeper problems that the code might cause or exacerbate.

Why it matters

Vibe coding could be just a flash-in-the-pan phenomenon that will fizzle before long, but it may also find deeper applications with seasoned programmers. The practice could help skilled software engineers and developers more quickly turn an idea into a viable prototype. It could also enable novice programmers or even amateur coders to experience the power of AI, perhaps motivating them to pursue the discipline more deeply.

Vibe coding also may signal a shift that could make natural language a more viable tool for developing some computer programs. If so, it would echo early website editing systems

known as [WYSIWYG editors](#) that promised designers "what you see is what you get," or "drag-and-drop" website builders that made it easy for anyone with basic computer skills to launch a blog.

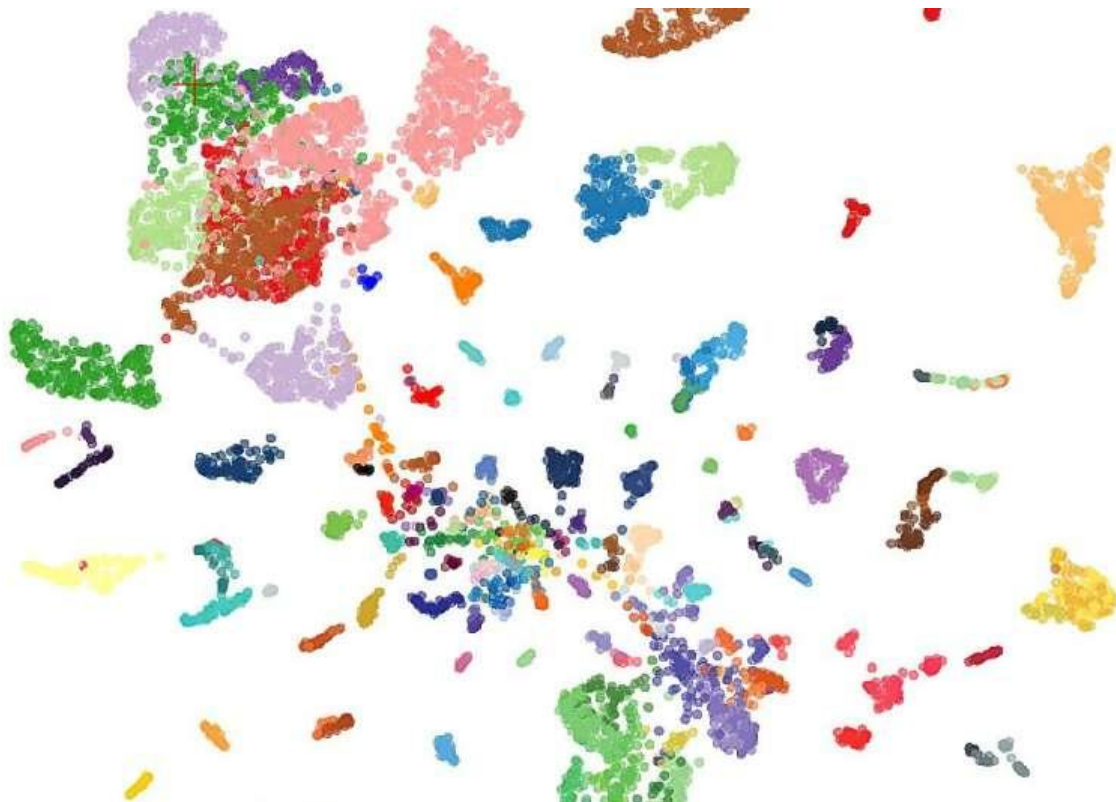
For now, I don't believe that vibe coding will replace experienced software engineers, developers or computer scientists. The discipline and the art are much more nuanced than what AI can handle, and the risks of passing off "vibe code" as legitimate software are too great.

But as AI models improve and become more adept at incorporating context and accounting for risk, practices like vibe coding might cause the boundary between AI and human programmers to blur further.

Provided by [The Conversation](#)

Tumor diagnostics: AI model detects more than 170 types of cancer

by [Charité - Universitätsmedizin Berlin](#) edited by [Gaby Clark](#), reviewed by [Robert Egan](#)



The AI model, crossNN compares the epigenetic data of unknown tumors to the fingerprints of over 8,000 reference tumors. A cross marks the tumor to be examined. The

visualization shows the large amount of data on which the model is based. Each dot represents the profile of a reference tumor containing several hundred thousand pieces of information; each color represents a specific tumor type. Charité | Philipp Euskirchen.

Credit: Charité | Philipp Euskirchen

The MRI shows a brain tumor in an inauspicious location, and a brain biopsy will entail high risks for a patient who had consulted doctors due to double vision. Situations such as this case prompted researchers at Charité—Universitätsmedizin Berlin to look for new diagnostic procedures. The result is an AI model.

The model makes use of specific characteristics in the genetic material of tumors—their epigenetic fingerprint, obtained for example from [cerebrospinal fluid](#), among other things. As the team [shows](#) in the journal *Nature Cancer*, the new model classifies tumors quickly and very reliably.

Today, far more types of tumors are known than the organs from which they arise. Each tumor has its own characteristics: certain tissue features, growth rates and metabolic peculiarities. Nevertheless, tumor types with similar molecular characteristics can be grouped together. The treatment of the individual disease depends decisively on the type of tumor.

New, targeted therapies address certain structures of tumor cells or block their signaling pathways in order to halt pathological tissue growth. Chemotherapies can be selected according to tumor type and their dosage adjusted accordingly. Particularly in the case of rare tumor types, it may be possible to pursue innovative therapies as part of studies.

"Against the backdrop of increasingly personalized, rapidly developing cancer medicine, precise diagnosis at a certified tumor center is the way forward for successful treatment," Prof. Martin E. Kreis, Chief Medical Officer at Charité, stated.

While a comprehensive molecular, cellular and functional analysis of a tumor based on [tissue samples](#) provides the necessary information, physicians are also confronted with cases in which it is not possible or very risky to extract tissue samples from the tumor. What's more, even a histological examination alone is not capable of providing as precise a diagnosis as the new AI model.

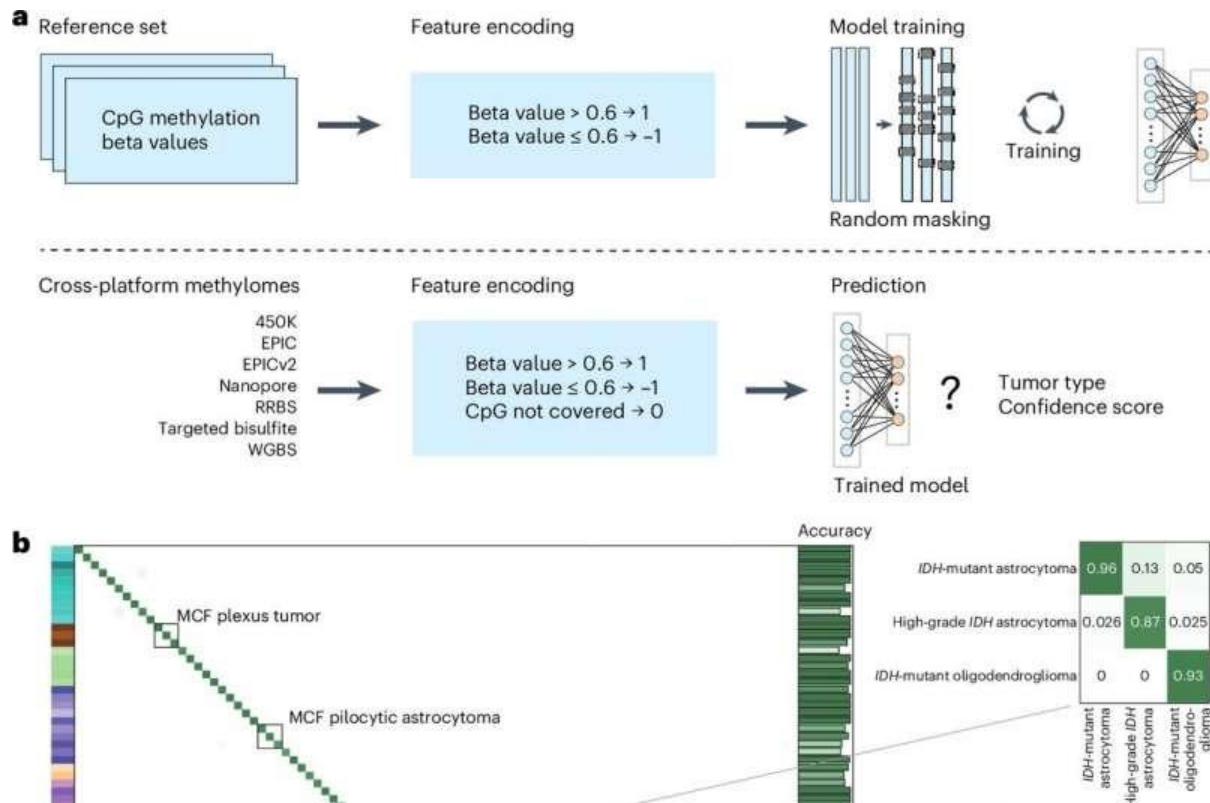
Looking into the genome instead of into the tissue

A method has been established for characterizing [brain tumors](#) that is not based on conventional microscopic diagnostics, but instead on modifications of the tumor's genetic material, the epigenetic characteristics. They are part of the memory of every cell and determine which parts of the genetic information are read and when.

"Hundreds of thousands of epigenetic modifications act as on and off switches for individual gene sections. Their patterns form a unique, unmistakable fingerprint," explains Dr. Philipp Euskirchen, scientist at the Berlin site of the German Cancer Consortium and at the Institute of Neuropathology at Charité, who heads the recently published study.

"In tumor cells, the epigenetic information is altered in a characteristic way. Based on their profiles, we can differentiate between tumors and classify them."

In the case of brain tumors, even a sample of the cerebrospinal fluid is sufficient in some cases, and can be obtained relatively easily—dispensing with surgery altogether.



crossNN model architecture, training and CV. Credit: *Nature Cancer* (2025). DOI: 10.1038/s43018-025-00976-5

In order to compare an unknown fingerprint with thousands of known fingerprints of different cancers and assign it to a specific tumor type, machine learning methods, i.e., artificial intelligence, are required given that the data are very extensive and complex. What's more, different DNA sequencing methods were applied in the past. In addition, epigenetic analyses are usually limited to defined patterns and gene segments that are typical for individual tumor types.

"Consequently, our aim was to develop a model that accurately classifies tumors, even if they are only based on parts of the entire tumor epigenome or the profiles were collected by means of different techniques and varying degrees of accuracy," says bioinformatician Dr. Sören Lukassen, head of the Medical Omics working group at the Berlin Institute of Health at Charité (BIH).

Reliable and traceable

A newly developed AI model goes by the name of crossNN, whose architecture is based on a simple neural network. The model was trained with a large number of reference tumors and subsequently tested on more than 5,000 tumors.

"Our model allows a very precise diagnosis of brain tumors in 99.1% of all cases and is more accurate than the AI solutions at work to date," Euskirchen says.

"In addition, we were able to train an AI model in the same way that can differentiate between over 170 [tumor types](#) from all organs, while achieving accuracy of 97.8%. This means that it can be used for cancers of all organs, in addition to the relatively rare brain tumors."

The decisive factor for future approvals in [clinical application](#) is that the models are fully explainable, i.e., it must be possible to understand how the decisions are arrived at.

The molecular fingerprint that the AI model receives for determination can stem from a tissue sample or from body fluids. In the case of specific brain tumors, the Department of Neuropathology at Charité already offers non-invasive diagnostics based on cerebrospinal fluid, known as liquid biopsy. This allows a diagnosis to be made without a stressful operation, also in difficult situations. The patient who consulted us with double vision was one of the beneficiaries.

"We examined the cerebrospinal fluid using nanopore sequencing, a novel, very fast and efficient form of genetic analysis. The classification by our models revealed that it was a lymphoma of the central nervous system, enabling us to promptly initiate appropriate chemotherapy," Euskirchen explains.

crossNN in clinical trials

The accuracy of the methodology even took the researchers by surprise. "Although the architecture of our AI model is far more simple than previous approaches and therefore remains explainable, it delivers more precise predictions and therefore greater diagnostic certainty," says Lukassen.

Together with the German Cancer Consortium (DKTK), the research team is therefore planning clinical trials with crossNN at all eight DKTK locations in Germany. In addition, intraoperative use is also to be tested. The aim is to transfer the precise and comparatively inexpensive tumor determination based on DNA samples to routine care.

More information: Dongsheng Yuan et al, crossNN is an explainable framework for cross-platform DNA methylation-based classification of tumors, *Nature Cancer* (2025). DOI: [10.1038/s43018-025-00976-5](https://doi.org/10.1038/s43018-025-00976-5)

Journal information: [Nature Cancer](#)

Provided by [Charité - Universitätsmedizin Berlin](#)

Growing A New Body: How To Completely Restore Your Health And Well-Being At Any Age

BY MEGHAN MCDONALD

Did you know that every single organ and system in your body—from your bones to your heart, to your brain and everything in between—will regenerate itself completely multiple times over the course of your life? You grow a completely new heart every 15 years, your bones are only two years old, your lungs and skin are around two to four weeks old, and your intestines have replaced all their cells within the last three days.

So, if this is the case, why do chronic health issues persist over weeks, months and years? Why does our health typically deteriorate over time instead of improve as we age? And what are those lucky few people doing who seem to enjoy radiant, disease-free health and longevity well into old age that the rest of us aren't? Is it their genes, their eating habits, their lifestyle, their mindset, all of the above or something else entirely?

The Secrets to Longevity, DNA Repair and Deep Healing

Figuring out the answers to the above questions requires a diverse background of training, skills and experience in a number of fields—from the traditional health sciences to the cutting edges of neuroscience and epigenetics, to the far reaches of transpersonal psychology, ethnobotany, energy medicine and beyond—ones that someone like [Alberto Villoldo, Ph.D.](#) is uniquely suited for.

As a respected and prolific medical anthropologist who has spent nearly 30 years working alongside the world's leading medical researchers and doctors, as well as spending equal amounts of time studying the ancient healing modalities and alternative medicine traditions of the indigenous tribes throughout the Americas, Villoldo holds unique insight into the roots of disease and the secrets of longevity that emerge from blending the best of science, psychology and spirituality.

And what he has discovered in his 30-year journey is nothing short of extraordinary—that longevity, genetic rejuvenation and repair, and deep healing of virtually any and

all ailments of a mental, emotional or physical nature can be accomplished through a powerful five-phase process that blends the best of both ancient and modern science.

The 5 Steps to Growing a New Body

In Alberto's new book *Grow a New Body*, he dives deep into the science, supplements and daily practices needed to accomplish all of the above and far more in a remarkably short period of time—as little as 7 days if his protocols are followed steadfastly and correctly.

In the book, he dives deep into the 5 key phases anyone seeking to restore their body to optimal health and create lasting, lifelong wellness must undergo.

Phase 1. Discovering Your Inner Healer

In Alberto's own words, Phase 1 is about addressing the most foundational roots of health and disease:

"In the West we do not have a health-care system, we have a disease-care system that recognizes thousands of ailments and myriad remedies. Shamanic medicine traditions, on the other hand, identify only one ailment and one cure. The ailment is alienation—from our emotions, from our bodies, from the earth, and from Spirit. The symptoms of this ailment are physical and emotional disease. The cure is the experience of Oneness, which restores inner harmony and facilitates recovery from all maladies, regardless of origin.

When our minds, our emotions, our relationships, or our bodies go out of kilter, we tend to ignore the problem until something goes very wrong—a scary diagnosis, a broken relationship, the death of a loved one, or simply an inability to function easily in everyday life. When things are a little bad, we read a self-help book or go to a workshop. When they're really bad, we bring in experts to fix the problem—oncologists to address cancer, neurologists to repair the brain, psychologists to help us find peace and understand our family of origin. But this fragmented approach to health is merely a stopgap. To truly heal, we need to rediscover the original recipe for wellness discovered millennia ago: One Spirit Medicine.

One Spirit Medicine states that the best way to heal and maintain health is to grow a new body on a regular basis. It recognizes the body as a system, not as a collection of parts that can be medicated or replaced when they go wrong. You cannot repair the heart without attending to the gut and the brain, and vice versa. And Western science has shown us that you can begin to grow a new body in as little as seven days."

Phase 2. Shedding the Old Ways

The old ways Alberto refers to are the outdated methods of Western “disease care” medicine in favor of the new science of preventative, holistic functional medicine. In order to truly achieve radiant levels of health, wellness and longevity, we literally need to reboot our bodies and minds at the deepest levels, which requires proactive lifestyle changes that work to support health every day, not once we become sick after years of neglecting and abusing our bodies unintentionally or otherwise.

There are three keys to doing this effectively according to Alberto’s research:

1. Detoxifying and repairing the gut-grain
2. [Switching on your longevity genes with diet](#)
3. Supplementing with superfoods and other holistic power medicines

The end result of this program is a completely upgraded brain and body that kick off the epigenetic repairs needed to sustain radiant health for the rest of your life.

Phase 3. Resetting Your Death Clock

Phase 3 is all about repairing health at the cellular level, which is also intimately connected to the mental and emotional levels in a powerful mind-body feedback loop. To do that fully, you’ll need to learn to transform dysfunctional emotional patterns founded on anger and fear, and how to supplement with certain nutrients that can help to improve higher-brain function and help you manage stress.

During this phase, you’ll also learn to work with and optimize your mitochondria (the power centers of your body) along with how to reset your cells’ “death clocks” and switch on longevity proteins controlled by mitochondria that extend life exponentially. You’ll learn exactly what free radicals and inflammation do to the body and how to reverse the damage, and you’ll discover ancient shamanic techniques used for thousands of years that can upgrade your body and brain. It is infinitely less complicated than it sounds, and part of Alberto’s genius is offering easy things you can do that achieve many of these goals at once.

Phase 4. Whole-Self Rebirth

Phase 4 supports you in the process of letting go of old, unhealthy ways of thinking so that you can experience full spectrum mind-body healing. You’ll learn how to shed outworn narratives about your past and embrace a new, liberating personal story. You’ll overcome fear of loss and change, and discover a new purpose for your life journey.

Phase 5. The Longevity Lifestyle

In Phase 5, you combine everything you've learned from the previous four steps into action. This is where the magic happens—in just seven days you will initiate a radical healing process in the body that will speed up the regeneration and repair of every organ and system in your body with new, upgraded genetic information and proper functioning that works at the deepest levels to restore the optimal functioning of your body and mind.

You will learn to change what and when you eat in order to enter into ketosis to activate the regions in the brain that will allow you to have an experience of Oneness. You will also learn to work with neuro-nutrients available at your health-food store that detoxify the body and brain and switch on the genes that create health and longevity. Lastly, you will learn how to do a soul retrieval to recover the vibrant and innocent parts of yourself that were lost because of childhood trauma.

A Proven Process of Rejuvenation

[The Grow a New Body rejuvenation program](#) outlined in Alberto's new book is the result of decades of research and study both in the lab and at the front lines of holistic medicine, and has been successfully implemented and fine-tuned for the last 10 years with thousands of people from around the world.

Backed by the latest breakthroughs in neuroscience, quantum biology, psycho-neurobiology, functional and mind-body medicine, this book is a powerhouse of wisdom of all flavors and highly practical, real-world tested information that anyone can use to supercharge their health in a remarkably short period of time—just one week!

To start the 7-day Grow a New Body Protocol, order a copy of Alberto's book *Grow a New Body* for yourself at: growanewbody.com

This article is a sponsored post written in collaboration with [The Four Winds Society](#), whose products and ethos complies with Conscious Lifestyle Magazine's stringent quality and integrity guidelines.

NEWS 10 June 2025

How to spot suspicious papers: a sleuthing guide for scientists

An open collection of tips and tools could help researchers and publishers to pick up on problematic research.



Science-integrity sleuths check the text and images in published papers for errors and inconsistencies. Credit: tadamichi/Getty

A group of research-integrity experts has launched a toolkit for researchers that outlines how to spot suspicious scientific papers.

The guide's creators hope that it will be a useful resource for sleuths who review published work for signs of sloppy, fake or fraudulent science, and will perhaps inspire others to get started. "The sleuthing community is growing. And this will be the essential handbook," says Jana Christopher, an image-integrity analyst at FEBS Press in Heidelberg, Germany, who contributed to the project.



What makes an undercover science sleuth tick? Fake-paper detective speaks out

The [Collection of Open Science Integrity Guides](#) (COSIG) brings together 27 freely available resources that explain how to spot [image duplication](#), citation manipulation, [plagiarism](#), [tortured phrases](#) and other hallmarks of [paper mills](#) — businesses that produce fake papers to order. The guides also provide tips for reviewing papers in specific disciplines, including biology, chemistry, statistics and computer science.

“COSIG is really a compendium of all the tips and tricks that various sleuths have acquired over the years,” says Reese Richardson, a metascientist at Northwestern University in Evanston, Illinois, who led the project and authored a preprint on COSIG posted to the repository Zenodo on 4 June¹. “We want to arm scientists with the tools to uphold the integrity of the literature.”

“It will really help publishers, as well,” says Christopher, adding that research-integrity officers could use the guide as the basis for checklists to evaluate manuscripts.

‘Anybody can do it’

Richardson came up with the idea for COSIG last year, while at a meeting of science-integrity specialists in Paris. Concerned about the surge in formulaic research and flawed publications that continue to circulate largely unnoticed, he and other delegates wanted to broaden participation in correcting the literature and to help more people learn how to scrutinize papers.

“A lot of people assume that you need some special talent, you need eagle eyes to see things, or you need to be at your computer ten hours a day looking through the scientific literature. But really, anybody can do it,” says Richardson. “That’s one of our mantras.”



AI beats human sleuth at finding problematic images in research papers

The guides describe how to spot issues that are common across all disciplines, such as text that has been [generated using artificial intelligence](#) (AI), and publications from [hijacked journals](#) — scam websites that impersonate legitimate journals. They include instructions for how to use online tools such as the Feet of Clay Detector, which [flags papers citing retracted or problematic research](#), and Imagetwin and Proofig, which [use AI to detect integrity issues in scientific figures](#).

Some of the guides home in on areas of research that have seen a [recent surge in formulaic publications](#), and those that are frequently targeted by paper mills.

For example, one guide shows how to detect experiments that use non-existent cell lines, or contaminated ones that are no longer suitable for studying certain cancers. Another outlines how to check that tumour sizes in rodent studies fall within limits approved by institutional ethics committees (a reported size that is too large to have been approved can be a sign of manipulated or fabricated experiments).

Young AI coding startup **Anysphere** has become the face of “**vibe coding**” — a phrase coined by OpenAI cofounder Andrej Karpathy describing the use of large language models to create applications when the user doesn’t necessarily need to know how to program. The nascent startup has **raised \$900 million at a \$9.9 billion valuation** and claims to have about \$500 million in annualized revenue. The startup is betting that AI is going to dramatically transform software engineering in the next decade, making it magnitudes easier to program applications while eliminating cumbersome aspects of the process like correcting syntax or or debugging code. All a person has to do is press tab and AI completes the line of code for you and jumps to the next spot.

Also of note: Enterprise AI startup **Glean** raised **\$150 million in Series F funding at a \$7.2 billion** valuation. Employees use the company’s AI tools to search for internal information and build AI agents (software that can carry out specific tasks end-to-end) that can resolve IT tickets, write performance reviews and help prepare for meetings. Glean claimed to have passed \$100 million in annualized revenue in February.

AI Comes To Hollywood

Throngs of excited moviegoers piled into Alice Tully Hall at Lincoln Center on Thursday night to be a part of **Runway’s third annual AI film festival**. Cristobal Valenzuela, CEO of the **\$3.3 billion video and photo generation AI startup**, spoke to a crowd of hundreds, asking them to think less about the digital tools and AI software used to make the short films they were about to watch, and instead focus on their human elements.

The winning film, *Total Pixel Space* by Jacob Adler, is a **jumble of both realistic and impossible vivid landscapes** like a flying pig, people floating in a city or inside a pool, a bloom of jellyfish and a meerkat donning a bright yellow turtleneck. The 9 minute 28 second film raises the question of how many images could possibly exist in the world. The answer: Every image is composed of thousands of pixels—

coordinates of positions and colors, a coalition of numbers. The film was **selected from 6000 submissions**, up from 300 a year ago, as interest in experimenting with AI models has exploded over the years.

For all the creative perks of video generation AI software, TV networks and filmmakers are adopting the technology for a more pragmatic reason: **to produce and edit both television shows and movies quickly and more cheaply**. AMC Network, which has produced popular shows like *Breaking Bad* and *The Walking Dead*, recently announced its plans to use Runway's AI models to create marketing and TV content. Lionsgate, the studio behind blockbuster hits like *The Hunger Games* and *The Twilight Saga*, has a partnership with Runway to use its models with a goal of making films on a **fraction of the budget**.

But several studios [don't want to openly admit](#) they're using AI due to fears of **backlash from creatives**, who have voiced their concerns that these AI models are **trained on copyrighted data scraped from the internet without consent** and compensation. Runway AI is also currently facing litigation from a group of artists who claim their data was illicitly used to train its AI models.

Meta is reportedly planning to acquire **a 49% stake in data labelling behemoth Scale AI** for \$14.8 billion. The deal is slated to place Scale AI's young billionaire CEO Alexandr Wang at a top position inside Meta along with a number of Scale AI employees, who will work in a new AI lab dedicated to **developing "superintelligence"**— an AI system that outperforms human capabilities. CEO Mark Zuckerberg is reportedly closely involved in assembling the team of AI researchers and has gone to great lengths like setting up a **WhatsApp group called "Recruiting Party,"** personally reaching out to potential recruits and rearranging desks for researchers to sit near him, [Bloomberg](#) reported. The new lab is part of Meta's efforts to keep up in the cutthroat AI race while wrangling a string of internal issues including employee churn, management problems and delayed or disappointing product releases.

With on-device AI, is Apple making a move no one saw coming?

Apple isn't just sharing its AI. It's betting developers will finish the job.



(Image credit: Apple / YouTube)

On Monday, day one of [WWDC](#), [Apple](#) finally gave developers the keys to its local [AI](#) engine. Dubbed **Foundation Models**, this on-device framework lets third-party apps tap into Apple Intelligence's generative muscle without cloud access, API fees, or permission slips. Just three lines of Swift code, and you're off to the races.

On paper, it's a gift. But look closer, and it might be a white flag.

Rather than chase Google and OpenAI with its own killer apps, Apple is throwing open the App Store doors and crossing its fingers. If that sounds a little too generous from Apple, there's a reason: This might be less about Apple flexing its AI dominance and more about admitting that it's behind. It's also a move that would've been hard to predict just a year ago.

Letting developers build the next [ChatGPT](#) alternative or journal-coaching wizard is both strategic and convenient. And some may even consider it necessary.

A shortcut through Apple's AI shortcomings

Rather than try to outmaneuver [an increasingly smart Google Gemini](#) or OpenAI's GPT-5 with a headline-stealing chatbot, Apple is outsourcing the hard part.

The *Foundation* of [Foundation Models](#) isn't about competing feature-for-feature but rather enabling a wave of third-party AI apps and praying the best ones rise to the top.

That means Apple can skip the messiness of building one perfect assistant (Siri) and instead position itself as the AI foundation that everyone else can build on.

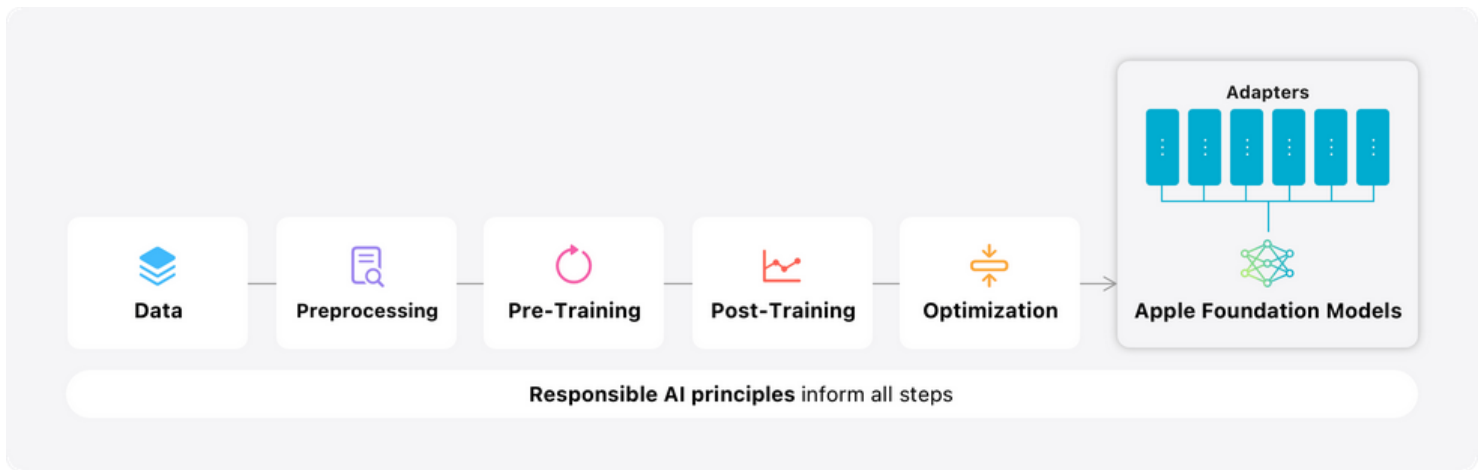
It's no secret that Siri has [struggled to evolve](#), and Apple's cloud-scale LLMs still lag competitors in size and polish. So instead of outbuilding them, Apple is out-delegating them.

Instead of outbuilding its AI rivals, Apple is out-delegating them.

This strategy plays to Apple's strengths: hardware optimization, privacy-first design, and a locked-in App Store ecosystem that still shapes what "mobile software" looks like.

If someone builds the next breakout AI experience on [iOS](#), Apple wins, whether it wrote the code or not. There's no licensing, no throttling, and no user data leaving the device. Developers are already embracing this offer. Journal app [Day One](#) is rolling out smart journaling suggestions, and [AllTrails](#) will soon auto-generate hiking routes. But those are just early ripples.

The real wave will be niche apps, side projects, and indie utilities that flood the store with personalized, privacy-first AI helpers. With zero cloud overhead and no per-query costs, even small indie teams can afford to get weird with it.



Modeling overview for the Apple foundation models. (Image credit: Apple)

Apple's Intel Inside moment?

We've seen this before. Apple didn't make the most popular fitness app; it built the [Apple Watch](#). It didn't build the App Store's most iconic games; it gave developers the tools to build them. The difference this time is that Apple needs developers to carry the load.

You can think of this as Apple taking the Intel Inside approach, but with AI.

You can think of this as Apple taking the *Intel Inside* approach, but with AI. Much like Intel once did, Apple's AI will quietly power thousands of experiences, even if it never becomes the household name.

By embedding itself as invisible infrastructure, Apple gains reach and resilience. Don't have a killer AI app? No problem. Let the App Store figure it out. Eventually, someone will.

Here's the gamble for Apple

There's no way to sugarcoat the fact that while Apple's AI rollout was elegant in presentation, it's arguably light on the finished features.

Once again, Siri was sidelined at WWDC. The flashy demos — writing replies in Messages or summarizing Notes — rely on a combination of on-device and cloud processing, and many will not arrive in beta form until the fall.

So what's Apple doing instead? It's hedging its bets.

Apple is flooding the space by letting developers plug into Foundation Models. Don't have a killer first-party AI experience? No problem. Let the App Store build five hundred.

With no API fees or usage caps, Apple is effectively inviting devs to throw ideas at the wall and see what sticks, hoping that at least a few do.

Apple is effectively inviting devs to throw ideas at the wall

This approach isn't without risk. For one, Apple can't guarantee results. A decentralized AI push means wildly inconsistent UX. Some apps will handle context and prompts gracefully. Others might deliver clunky, confusing, or outright inaccurate outputs.

Unlike Siri, which Apple can directly improve, these experiences will be fragmented and difficult to course-correct without heavy-handed store moderation.

There's also the discoverability problem. If Apple Intelligence-powered apps flood the store, users won't necessarily know which ones are useful.

Without tight integration or prominent editorial guidance, Apple risks unleashing a thousand half-baked AI clones, each with a slightly different idea of what "smart" means.

And then there's hardware load. Running a 3-billion parameter language model on an [iPhone](#) may sound cool on stage, but developers will have to wrangle with battery drain, thermal throttling, and limited RAM, especially on older devices.

Apple's own engineers were quick to highlight quantization and token efficiency, but this still isn't free. The user pays in performance, even if they're not paying in dollars.

AI, the Apple way

Still, there's brilliance in the chaos.

By sidestepping the cloud arms race, Apple is leaning into what makes its ecosystem unique: Privacy, silicon, and a developer community that (usually) figures things out. This is Apple AI as infrastructure, not as identity. That might just work.

Apple will still win if the next great AI assistant comes from a developer in Portland or Tokyo. And if not? Well, at least it didn't spend another five years trying to make Siri cool again.

"It's 5GB of pure bloatware": Apple Intelligence unlikely to win over Mac users at WWDC — Here's why June 2, 2025

WWDC 2025 probably won't include the one AI feature Mac users want



When you purchase through links on our site, we may earn an affiliate commission. [Here's how it works.](#)

(Image credit: Apple)

Apple's annual [Worldwide Developers Conference \(WWDC\)](#) takes place on Monday, June 9, and the company is set to spotlight its biggest software updates of the year — but will any of its new Apple Intelligence features win over frustrated Mac users?

Apple Intelligence may have been the star of last year's show, but the platform, pitched as "[AI](#) for the rest of us," has mostly failed to capture the awe of Mac, [iPhone](#), or iPad users, with the one update everyone *really* wants to see [is nowhere in sight](#).

There's no doubt that Apple will liberally flaunt the term AI during this year's WWDC, but a claimed lack of any killer AI features to show could highlight how far ahead Apple's competition is.

However, to some, that's a good thing. Take a look online and you'll see many Mac users souring on AI, and until Apple is ready to change their minds with something truly groundbreaking, it might be best to keep it out of the limelight a little longer.

***See also:** [Best Apple deals in June 2025](#)*

WWDC will feature a rebranded macOS, but one key update will be missing

With WWDC 2025 now just a week away, we're getting our last, and arguably most exciting, wave of rumors about what to expect. The highlight, [according to Bloomberg's Mark Gurman](#), will be a major rebrand of Apple's operating systems along with the biggest visual overhaul since [iOS 7](#).

For Mac users, that means the next OS update [will be macOS 26](#), which will likely feature tweaked app icons and potentially changes to the overall UI, like more transparent elements. Rumors indicate that Apple is taking a lot of inspiration from visionOS for this sweeping visual overhaul across its platforms, giving a rough idea of what to expect.

One year later, Apple Intelligence still hasn't won over Mac users

Apple was undoubtedly late to the game in consumer AI, falling short of the innovations Google, [Meta](#), and OpenAI have been announcing over recent years. OpenAI is even partnering with former Apple design chief Jony Ive to develop its own in-house AI device.

Considering all that, one might think Apple is in hot water, dropping the ball on AI like it has. However, that's assuming Apple users actually *want* a boatload of AI features on their devices, which might not be the case.

For instance, [a user in one Reddit thread](#) referred to Apple Intelligence as "5GB of pure bloatware," complaining that trying to disable or remove the feature can lead to bugs.

A user in another thread expressed frustration about not being able to completely remove Apple Intelligence, [commenting](#), "Some of us don't like a service shoved down our throats we never asked for, and don't need. Sure, it can be turned off, but it should be optional to be installed to begin with. I'd like to use those few gigs of storage."

One user even described Apple's AI as "all hype, no value."

Even users who don't outright dislike Apple Intelligence see little use in it, like [one commenter who said](#): "I pretty much never use writing tools, even before they were integrated, so that's been barely used. Sucks that what the majority of Apple AI is."

[Another user](#) pointed out the core of the issue: "Siri still seems dumb as a rock."

All of this effectively amounts to a lack of enthusiasm for AI in general, at least among Mac users. It makes me wonder, would Apple be better off differentiating itself by backing off on AI, at least until the new-and-improved Siri is ready?

So far, Apple Intelligence seems to be mimicking competitors' AI offerings, only with fewer features and years behind. That strategy clearly isn't winning over Mac users, though, who seem to be really looking for practical AI tools and a Siri update, not gimmicks like AI-generated emojis.

Instead of trying to convince us it's an AI brand now, maybe Apple could woo fans by showing it's focusing on launching AI tools that prioritize genuine usefulness and user privacy.

Apple's AI difference: Why it's not building a chatbot



(Image credit: Apple)

For Apple, the main message with its AI strategy is that it doesn't want to build a chatbot. Instead, it wants to “meet people where they are” with AI.

That means delivering Apple Intelligence features inside apps that are designed to make your life easier or more fun, such as with the new [Call Screening](#) and [Hold for Me](#) features in the Phone app and Live Translate in Messages, Phone and FaceTime.

So, for example, if you're in the messages app and someone sends you a message in a language that you don't speak, Live Translate will ask if you want it to start translating for you.

"It's integrated so it's there within reach whenever you need it in the way you need it with it being contextually relevant and having access to the tools necessary to accomplish what you want to accomplish at that moment," said Federighi.

"Apple's job is to figure out the right experiences that make sense in the context of what we offer to customers and to make that technology," said Joswiak. "The features that you're seeing in Apple Intelligence isn't

a destination for us. There's no app on intelligence. [It's about] making all the things you do every day better."

What's next for Apple Intelligence

For now, Apple seems focused on delivering AI features that will make an impact as part of its new suite of software rolling out this fall, including iOS 26, iPadOS 26, macOS 26 Tahoe and even the Apple Watch with watchOS 26 and the new Workout Buddy feature.

The features that you're seeing in Apple Intelligence isn't a destination for us. There's no app on intelligence. [It's about] making all the things you do every day better.

And Apple is also opening up its large language models to third-party developers so they can tap into Apple Intelligence powers on device. A good example of Apple's AI evolution is Visual Intelligence. The upgrade coming with iOS 26 will let you identify an object on your screen and then instantly buy it on Etsy, for example.

"I didn't limit it just so that we let developers adopt an app and use the intents API. To plug into that experience, so if you have I don't know an app for collecting wine and you want to look up in your wine collector app, you can easily add it to your collection," said Federighi.

Bottom line

Apple Intelligence is still very much a work in progress, but Apple seems focused on delivering the Siri that was promised along with a wide range of AI-powered features that make its ecosystem stickier — even as the competition heats up.

"In the end, people buy products right, they buy experiences, said Joswiak. "We're very proud of the fact that across each of our hero product categories, we're number one of customer satisfaction right?"

"There's a reason for that, and we're trying to make those product experiences better and better and make those products better and better and that's what customers care about."

WWDC 2025 Live: iOS 26, Liquid Glass and all the biggest Apple event announcements

Here's what Apple announced at its developer conference

Apple used WWDC 2025 to introduce us to new versions of iPhone, Mac, iPad, Apple Watch, Apple TV and Vision Pro software. And to make sure we didn't forget about Apple Intelligence.

This year's updates will deliver a unified Liquid Glass interface across different platforms as well as a common naming convention.

Instead of version numbers, [Apple's software will now reflect the year that each update will run through](#) — hence, we'll be getting iOS 26, macOS 26, watchOS 26, iPadOS 26, tvOS 26 and visionOS 26 later this year.

But more important than the new names are the new looks Apple is introducing with each update. The [Liquid Glass design](#) features translucent menus, glossy icons and rounded controls, along with streamlined menus that look to create a seamless experience as you move from one Apple device to the next.

Besides the interface overhaul, some of the software updates are gaining new features specific to their platform. Other features, like a new Games app, will appear across multiple software updates such as iOS, iPadOS and macOS.

WWDC 2025 announcements in brief

Here's what Apple had to say about each of its software updates, which are available as developer betas starting today (June 9). Public betas will follow in July ahead of the full releases coming later this fall.

iOS 26: There are changes throughout Apple's iPhone software, with the Phone app in particular getting features that will help you better screen phone calls. Messages gets a number of new tools, highlighted by Polls in group chats. The Camera app is getting more streamlined menus to make the viewfinder less crowded, and iPhones that support Apple Intelligence will benefit from improvements to Visual Intelligence and more.

watchOS 26: A Workout Buddy app now analyzes your workout data to highlight your goals and offer words of encouragement as you exercise. (Yes, there's a voice component.) Apple is also introducing improvements to the Workouts app and Smart Stacks feature while bringing a new Wrist Flick feature for dismissing notifications.

tvOS 26: Most of the changes in tvOS related to the new Liquid Glass design, though Apple is delivering profiles that make it easier to jump to your favorite shows and movies. The Apple Music Sing feature now lets you turn your iPhone into a microphone for sing-along sessions.

macOS 26: Also known as macOS Tahoe, this year's Mac software update delivers improved ways of searching for and organizing your files. The Shortcuts app is also getting a boost that will tap into AI for creating routines while Spotlight is also becoming more context aware of the app you're working in when you search for files on your Mac.

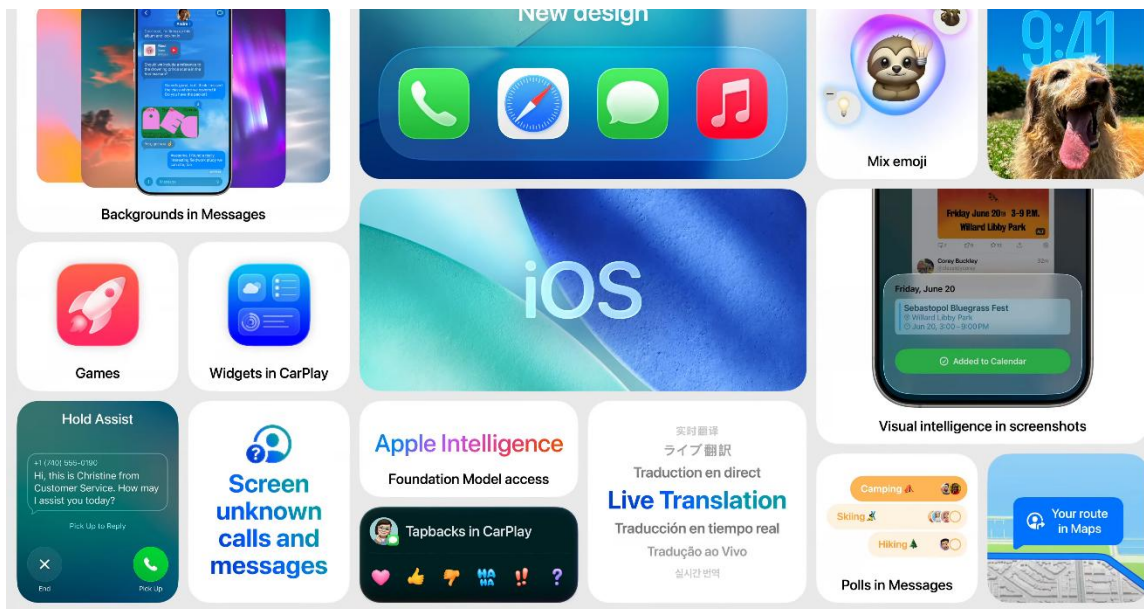
iPadOS 26: The biggest change to Apple's tablet software involves improved multitasking with a new windowing system that lets you more easily resize and switch between windows of different apps. Apple is also introducing audio and video improvements that should appeal to anyone who uses their iPad to create content like podcasts.

visionOS 26: Apple will bring widgets to its Vision Pro headset, along with shared spatial experiences that let you watch movies and play games with other Vision Pro-wearing people around you.

In addition to those software improvements, Apple plans [new Apple Intelligence updates](#), highlighted by a translation feature that will pop up in phone calls, FaceTime sessions and texts where you might need real-time translations. June 11, 2025

iOS 26 announced — here's all the new features for your iPhone

And there's a lot of big changes besides the name change



(Image credit: Apple)

Without drawing out the inevitable, Apple officially announced its big rebrand with iOS 26 taken off the wraps during its [WWDC 2025 keynote](#). Apple's choice to switch up the naming convention of its mobile software comes as no surprise, as there have been [numerous reports](#) about the big switch to a calendar year convention.

Apple's latest iPhone software has to offer, but oddly enough Apple Intelligence briefly came up in the conversation — early on with Apple saying that it'll share more news in the coming year about them. Despite this big surprise, there's the big visual makeover with iOS 26 that Apple calls Liquid Glass. It's the biggest visual change since iOS 7 and aims to unify the look and feel of its software across its entire device portfolio.

iOS 26: All the new features coming to your iPhone June 8, 2025

Whether it's called iOS 26 or iOS 19, some good stuff is expected



When you purchase through links on our site, we may earn an affiliate commission. [Here's how it works.](#)

(Image credit: Shutterstock/ Tom's Guide)

Update: Our [WWDC 2025 live blog](#) is up and running. Check it out now for all the last-minute rumors and biggest news.

Apple is gearing up to reveal iOS 26 (or iOS 19, if the [name-change rumors](#) are inaccurate) at [WWDC 2025](#).

While a potential name change is interesting, what'll make or break the company's next significant OS update are the new features. Apple needs to offer something compelling with iOS 26, as the last few updates have been relatively minor ([Apple Intelligence](#) being the most significant new feature, which has been [plagued with problems](#)).

WWDC 2025 Last-Minute Rumors: iOS 26, Apple Intelligence, SURPRISE Hardware?! - YouTube

[Watch On](#)

Thankfully, the rumor mill has been churning out potential new iOS 26 features regularly, and it sounds exciting. Here are the most significant new features we expect to see when Apple launches iOS 26 at [WWDC 25](#).

Massive interface redesign



(Image credit: Front Page Tech / YouTube)

The centerpiece of the next version of iOS is reportedly a massive visual overhaul. According to [Bloomberg's Mark Gurman](#), Apple has three key focuses for the new interface: simpler to use, faster to navigate and easier to learn.

"The big thing with this redesign is consistent controls, icons, buttons, user interface dynamics and mechanisms across both iOS and macOS," said Gurman in a report.

The alleged codename for this new interface design is [Solarium](#), and it will reportedly be consistent across all of Apple's new software and borrow the look of visionOS.

A big look at what iOS 26 could look like came from a Jon Prosser video. This isn't a final look at the OS, as it's leaked mockups, but it could give us a strong indication of what to expect from the new design.

This is expected to be the biggest shakeup to the iOS look and feel since iOS 7, which launched 12 years ago. That's a long time for a popular OS to remain relatively stagnant, so I'm very excited to see what Apple does.



Improved Apple Intelligence



(Image credit: Future)

Apple has to knock it out of the park with [Apple Intelligence](#). Whether the company calls it Apple Intelligence 2.0 or something else entirely, the original launch wasn't good. Apple typically doesn't launch a half-baked product, but that's precisely what happened with the company's race to catch up in the AI space.

Some new [changes expected to arrive on iOS 26 include](#): Siri's AI upgrades, new third-party AI integrations, iPhone battery improvements and expanding AI into more apps.

Siri in particular is way behind Google Gemini Live and ChatGPT. Apple isn't rumored to have a true competitor that can handle live conversations and multi-modal support for audio and video until next year. But we could get other promised upgrades like personal context awareness (think emails, photos, and calendar), on-screen awareness, and deeper app integration in iOS 26. It just might not happen right away.

The biggest thing Apple needs to do is make Apple Intelligence good. As the main selling point of [iOS 18](#), it underdelivered. The company has another chance to make a good impression, as users want the AI built into their phone

to work well. However, if Apple drops the ball again, it might be tough to win users over.

Charging estimate on Lock Screen

While this might seem like a minor change, knowing how long it'll take your phone to charge is a helpful feature when you're in a hurry. And this should be an iOS 26 feature.

According to some reputable leakers, you'll be able to look at your [iPhone's lock screen and see how long you must wait for a full charge](#). Knowledge is power, and this feature is most definitely a powerful one that's worth getting excited about.

Polls in Messages



(Image credit: Shutterstock)

According to a report in [9to5Mac](#), there are two notable upgrades coming to Messages in iOS 26. The first is polls. So, for example, you could insert a poll in a text thread about possible restaurant options. Apple may even suggest and format polls based on the context of the conversation.

Automatic translation in Messages



(Image credit: Future)

The second upgrade is more niche. Apparently Apple is working on automatic translation for incoming and outgoing messages if you happen to be texting someone who speaks a different language.

Notes

Apple is reportedly going to add Markdown support to the Apple Notes app. This would enable users to more easily export notes, as Markdown is a plain text format. It would presumably make keyboard-driven formatting easier and make it simpler to switch to other note-taking apps.

If you happen to use Apple Music, you'll be happy to know that Apple may be bringing full-screen animated album art to your lock screen on the iPhone. This could look pretty slick, and I hope Spotify is allowed to do the same thing, as it makes listening to music part of the OS experience.

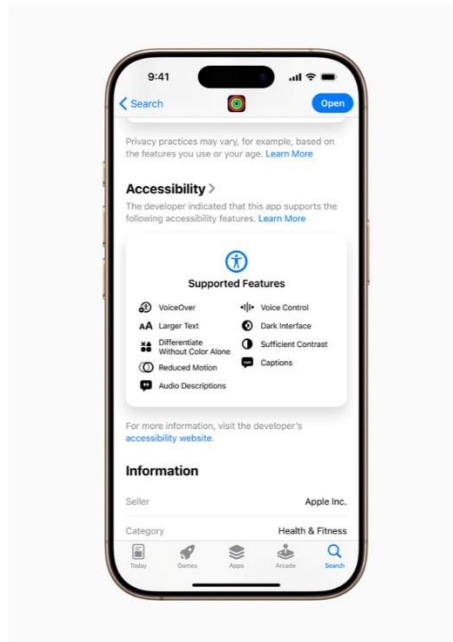
iOS 26 CarPlay



(Image credit: Apple via YouTube)

Apple CarPlay could be getting a major facelift with iOS 26, as it's rumored to mirror the same Solarium AI overhaul as the iPhone itself. Not much else is known at this point, but for people who use their iPhone with their car, this could be a big deal.

Accessibility Nutrition Labels



(Image credit: Apple)

According to Apple, the Accessibility Nutrition Labels will bring a new section to App Store product pages that highlight accessibility features within apps and games. Basically, you can use these labels to find out if an app is accessible or if it won't be.

iOS 26 Accessibility features

Like any major OS update, Apple is expected to add numerous accessibility features to iOS 26. Some of these features will only be useful for specific individuals, but others can significantly speed up workflows for everyone.

Apple announced that it will add a [system-wide Reader mode and more](#). Here's the complete list of features coming to the accessibility side of things:

- Accessibility nutrition labels on the App Store
- All-New magnifier for Mac
- New braille experience
- Accessibility reader
- Live captions on Apple Watch
- Enhanced view with Apple Vision Pro
- Background sounds
- Faster personal voice

- Vehicle motion cues come to Mac
- Eye Tracking on iOS improved switch control
- Improved head tracking
- Support Switch Control for Brain Computer Interfaces (BCIs) on iOS, iPadOS and visionOS
- Assistive Access adds a new custom Apple TV app
- Music Haptics on iPhone becomes more customizable
- Sound Recognition adds Name Recognition
- Voice Control introduces a new programming mode in Xcode
- Live Captions adds new languages
- CarPlay gets support for Large Text
- CarPlay sound recognition gets new sounds
- Share Accessibility Settings lets users share accessibility settings with another iPhone or iPad

iOS 26: New and updated official apps

We've also heard rumblings that Apple will introduce new first-party apps and make significant updates to some existing ones.

There will be a new [Apple Games app](#), according to reports. Essentially, it will replace the often-forgotten Game Center app and add new functionality that was previously unavailable. For mobile gamers, this alone is worth getting excited about, as it signals that Apple still considers games an essential part of iOS.

Other rumored updates include live translation in the Translate app, allowing it to further compete with Google Translate (which [can now be made the default on iOS](#)).

Apple is also expected to enhance the Health App with Apple Health+. Features include food logging, educational videos, and AI coaching are on the table.

Finally, Messages is expected to play better with RCS, further bridging the divide between Android and iOS.

Other small features

Here's a rundown of other small tweaks you might expect to see on iOS 19:

- Stage Manager on external display
- Captive Wi-Fi sync to make connecting easier
- Background Sounds enhancements
- eSIM transfer to Android

WWDC 2025 is make or break for Apple Intelligence — here's why June 1, 2025

Apple's developer conference is going to be a turning point for Siri — but it could be good or bad news



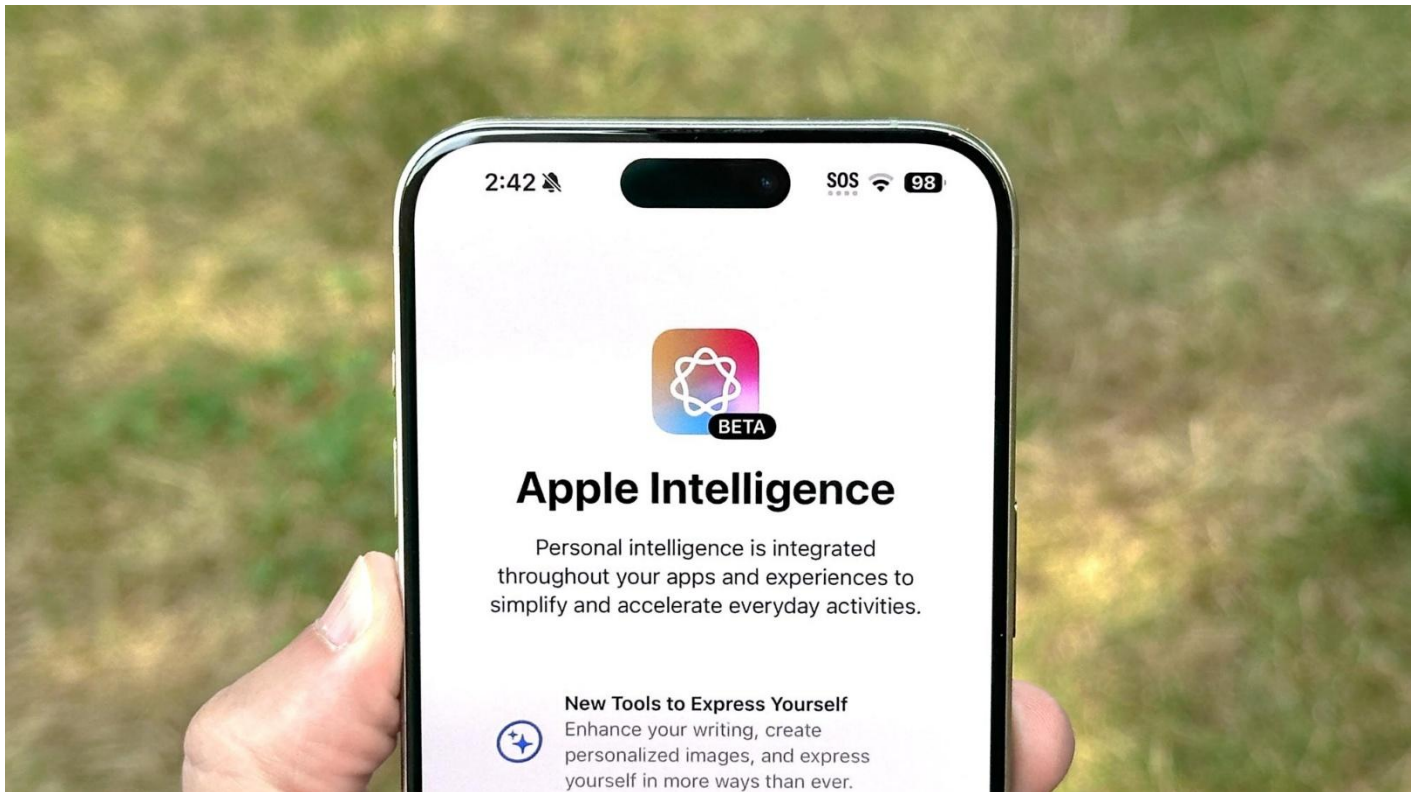
(Image credit: Tom's Guide / Shutterstock)

[WWDC 2025](#) is going to be a big deal for Apple users. Not only are we expecting to see a big redesign for [iOS 19](#)/ (or [iOS 26](#)), but it also marks one year since Apple went all in on AI and announced Apple Intelligence. Of course Apple Intelligence hasn't really been the resounding success that Apple probably hoped. It's not been a disaster, but [WWDC 2024](#) turned out to be the one thing Apple typically tries to avoid doing — overpromising and underdelivering. Nearly a year later, many of the promised [Siri](#) features are still missing in action.

Considering Apple was already late to the party with AI, and the troubles it's had, the pressure is on at [WWDC 2025](#). It's make or break, and if Apple doesn't ease the biggest concerns about Apple Intelligence then [it risks it ending up like Siri did 10 years ago](#).

YOU MAY LIKE

The problems with Apple Intelligence so far



(Image credit: Future)

The biggest issue with Apple Intelligence is that Apple realized AI was going to be a big deal much later than everyone else. Apple wasn't ignoring AI, but in the years before [ChatGPT](#) exploded in popularity, the company wasn't that interested in investing large amounts of money into AI development — especially with no clear end goal.

[According to a report from Bloomberg](#), it wasn't until after ChatGPT arrived that Apple's software chief Craig Federighi used generative AI for himself and realized how useful a tool generative AI could be. But by that point Apple was seriously far behind its rivals, and wouldn't be able to catch up easily.

This is apparently where the main problems with Siri come in, since Apple attempted to catch up by tacking the new [LLM-powered Siri](#) onto the older voice assistant. This hasn't worked out, not only because of the delays but also because it apparently caused a bunch of problems that have been described as "whack-a-mole."

All that inevitably made the controversial rollout of Apple Intelligence even more problematic. Not because the features that were released were bad, though things like [news summaries proved too problematic to keep around](#). Apple Intelligence itself didn't land until [iOS 18.1 arrived in late October](#), a month after [iOS 18](#) and the [iPhone 16](#) were released. iOS 18.2 was where the real improvements came into play, and that [didn't arrive until late December](#). iOS 18.3 and 18.4 landed throughout the first few months of 2025, but by that point the number of useful new features had dropped dramatically.

The problem wasn't the state of Apple Intelligence, though, and more of how Apple handled it. Simply put, it looked like Apple didn't want to be seen lagging behind its rivals, then overestimated what it could accomplish.

WWDC needs to set the stage for Apple Intelligence's future



(Image credit: Tom's Guide/Shutterstock)

WWDC is where Apple tells us what's going on with all its software, and it would be a mistake not to give Apple Intelligence the attention it needs. This is the first anniversary of its reveal, and despite all the problems Apple can't afford to be seen ignoring it.

I'm not saying that WWDC needs to be an all-Apple Intelligence show. [Google I/O](#) did that, and it was far too much AI for any normal person to handle. But

that doesn't mean Apple can brush AI to the wayside and treat it like Siri was treated for so many years.

If that happens, Apple might as well be throwing in the towel on the AI race. We all know that the company is behind the likes of [Google](#) and [OpenAI](#), but that doesn't mean the company's AI ambitions are dead. There's plenty of time to improve, and potentially catch up.

In a best-case scenario Apple would admit that it dropped the ball with Apple Intelligence, and pledges to do better going forward. I don't see that happening. Apple is not known for willingly admitting its mistakes.

But I also don't see Apple spending a great deal of time on AI either. Not just because it has a bunch of major design revamps to get us through in a keynote that can only be so long. But also because I'm sure Apple doesn't want to risk making the same mistakes as last year. No doubt we'll be hearing a lot of impressive specs about Apple Intelligence and its adoption, and maybe some reveals on different smaller features that may be on the way.

And that should be enough. AI isn't the focus of this year's releases based on what we've heard, and it shouldn't dominate the show. But it does still need attention and improvements so it can continue to grow.

Bottom line



(Image credit: Tom's Guide)

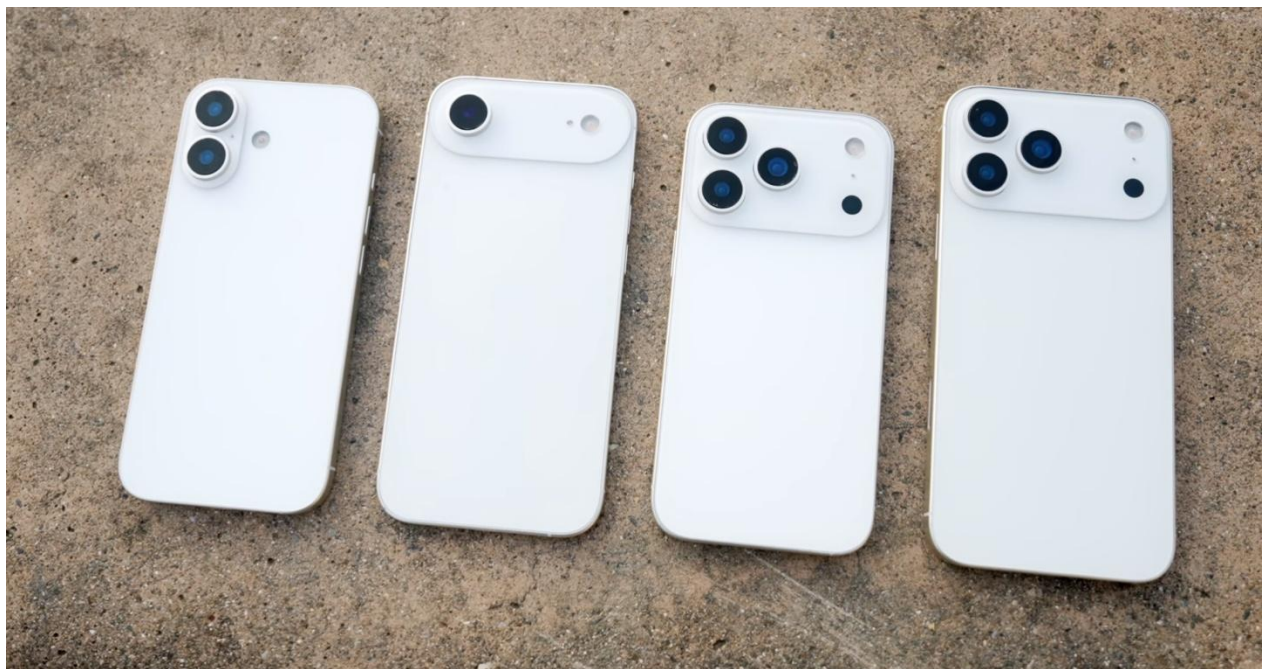
Apple has already made plenty of mistakes with AI, from jumping on the bandwagon late to screwing up the launch of the features when they were ready. So it's imperative that the company get itself into gear, and come up with an adequate strategy for future updates and AI features.

WWDC is going to be the starting point for all of that, and the attention Apple Intelligence gets at the show is going to lay the groundwork for the next few years of Apple AI rollouts. And while we can't expect Apple to roll out another wave of announcements like the ones we saw last year, it needs to avoid ignoring the topic completely.

Otherwise, if AI is just going to get tossed to the side because of some early hurdles, then Apple probably shouldn't have bothered investing in it in the first place.

Forget the iPhone 17 Air for a moment — these are the 5 most exciting rumors about the other iPhone 17 models May 23, 2025

Expect some important changes with this year's iPhones



(Image credit: AppleTrack)

When it comes to rumors about this year's iPhone releases, the [iPhone 17 Air](#) is the belle of the ball. And it's easy to understand why — the rumored ultra-thin model would mark a departure for an iPhone lineup that's been set in stone for several years now. Naturally, people want to hear about what an entirely new iPhone will be like, and rumor mongers are eager to fill in the blanks.

But all the excitement over the iPhone 17 Air threatens to overshadow the rest of the [iPhone 17](#) lineup. And that would be a mistake as this year's iPhones figure to deliver some key changes to Apple's phones, even if they're not as headline-grabbing as an ultra-thin frame that is rumored to be slimmer than even the 5.8mm-thin [Galaxy S25 Edge](#).

We're keeping tabs on *all* the iPhone 17 rumors here — not just the latest on the iPhone 17 Air, but what's going to happen to all the models Apple is likely to introduce later this year. So in case you only have eyes for the Air, here's a look at the most exciting rumors we've heard so far about the iPhone 17, the [iPhone 17 Pro](#) and the [iPhone 17 Pro Max](#).

Finally, a fast refreshing display



(Image credit: Future)

The entry-level iPhone has stood out for years when compared to the [best phones](#) you can buy — and not for a good reason. Rather, it's because the iPhone 16 is one of the few flagship models currently available where you're still stuck with a 60Hz display.

In recent years, smartphone makers have shifted to displays with adaptive refresh rates, which can ramp up to 120Hz when there's a lot of on-screen activity. A faster refresh rate means smoother scrolling and more immersive graphics on games that can take advantage of fast-refreshing displays. It's a dramatically better experience when you see a phone with a fast-refreshing display next to one with the screen locked at 60Hz.

Apple's been releasing iPhone Pro models with fast refreshing displays since the iPhone 13 Pro came out in 2021, but it's kept giving the standard iPhone a 60Hz display for the three iPhone launches since then. In the meantime, not only do phones that share the iPhone 16's \$799 starting price offer fast-refreshing displays of their own, so do many of the best cheap phones under \$500. The iPhone is way behind the curve here.

Things should change this fall. Multiple reports claim that [Apple is switching to an LTPO display for its entry-level flagship](#) — that's the kind of panel that supports faster refresh rates. The Pro models may still have a wider refresh rate range — the current [iPhone 16 Pro](#) offerings can scale all the way down to 1Hz when the on-screen action is static, for example — but it seems nailed on that the iPhone 17 will offer a much more capable screen than its predecessor.

A telephoto lens upgrade on the iPhone 17 Pro



(Image credit: Apple/ Tom's Guide)

The last couple iPhone upgrade cycles have seen a number of improvements to the cameras on Apple's Pro models. The [iPhone 15 Pro Max](#) added a longer optical zoom — a feature adopted by the iPhone 16 Pro last fall. And both Pro models saw their ultrawide camera sensors boosted to 48MP.

Apparently, Apple isn't done with the camera array upgrades. The iPhone 17 Pro and Pro Max are both in line to get a 48MP telephoto lens this fall. And yes, that would mean three 48MP sensors making up the rear cameras on the iPhone 17 Pro.

The [improved telephoto lens could come at a cost](#). Currently, the iPhone 16 Pro models support a 5x optical zoom, while one rumor suggests the higher resolution sensor will reduce zoom capabilities to 3.5x. Nevertheless, the higher resolution camera could mean more detailed images even if longer zooms may not look as sharp as before.

Lots of memory



(Image credit: Tom's Guide)

Apple is all about AI these days, and you'd imagine a whole new batch of Apple Intelligence tools are in the works with this fall's phone release in mind. And there's still the matter of the AI-based Siri improvements Apple promised with [iOS 18](#) before pushing them back to a later update.

All this AI requires a lot of memory. The current iPhone lineup comes with 8GB of RAM, based on teardowns. (Memory and battery size are two iPhone specs Apple doesn't officially reveal.) But reports claim that [the new iPhones could bump that number to 12GB](#) with an eye toward future-proofing the phones for future AI enhancements.

It's worth noting that some of the rumors suggest that only the iPhone 17 Pro models are in line to see an increase in memory, with the iPhone 17 potentially sticking at 8GB. Still, if you're excited about the prospect of better Apple Intelligence performance, you'll want Apple to add as much memory as possible to whatever models it can.

New iPhone 17 Pro design



(Image credit: AppleTrack)

It doesn't sound like a major redesign, and not everyone's going to love the rumored look, but from the renders that have emerged so far, the iPhone 17 Pro models should have a different look from their predecessors. So if you think Apple's designs have gone stale, you'll want to at least take a look at what this fall's releases could bring.

Specifically, Apple sounds like it's taking a page out of Google's design manual and [extending the camera array for the iPhone 17 Pro and iPhone 17 Pro Max](#). The triangular setup of the main, ultrawide and zoom lenses will remain in place, but the flash, LiDAR depth sensor and microphone will move to the other side of the phone.

There could be a change up front on the iPhone 17 Pro models, too. Reports suggest [Apple is going to shrink the Dynamic Island for the Pro and Pro Max](#), which will free up a little screen real estate.



(Image credit: Future)

The tweaks don't appear to be as prominent on the iPhone 17 renders we've seen. Leaked iPhone 17 designs show a pair of vertically stacked lenses housed in a rectangular camera bump rather than the oval array Apple used for the [iPhone 16](#).

Battery life improvements

it doesn't seem like any of the iPhone 17 models are due big boosts to battery size, though we might see some increased capacity here and there. The bigger advance could be a vapor cooling chamber in the iPhone 17 Pro models aimed at keeping the phones from overheating and suffering dips in performance. And as always, advances in power management from the new A19 silicon expected to power this year's phones could improve battery life over the current models.

Really, the battery capacity of the conventional iPhones could wind up contrasting with that of the iPhone 17 Air, which could have a [2,800 mAh power pack](#) — a decidedly smaller power pack than the 3,561 mAh cell in the 6.1-inch iPhone 16. Even with a report that claims Apple could turn to a new generation of silicon-anode battery from TDK for the iPhone 17 Air, that thin

phone could struggle to last a long time on a charge, making the endurance of the other iPhone 17 models seem even better by comparison.

iPhone 17 Air may not have all-day battery life — and the solutions defeat the purpose of a skinny phone May 19, 2025

Style over substance is a bad way to market a phone



(Image credit: Apple/ Tom's Guide)

2025 is shaping up to be the year of the super-skinny phone. Not only is the [Samsung Galaxy S25 Edge](#) going on sale at the end of the month, we're also expecting the [iPhone 17 Air](#) to be revealed later on this year. The hook these phones have is that they're thin and lightweight, emphasizing an almost unique style that we haven't seen on a smartphone for some time. The problem is that making a phone as thin as possible inevitably means trade-offs occur, and accounting for them could make the whole endeavor rather futile.

Big batteries need space



(Image credit: AppleTrack)

According to one recent report, the [iPhone 17 Air isn't likely to last a full day on a single charge](#) — with only 60-70% of users predicted to be able to do so. That's compared to 80 and 90% for the rest of the [iPhone 17](#) family.

We don't know how big the battery is going to be, and Apple isn't likely to ever tell us, but things are apparently so bad that it's going to resurrect a long-dead iPhone accessory — the battery case.

The battery case was discontinued after the launch of [iPhone 12](#), when [MagSafe](#) meant users could rely on magnetic battery packs to do the same job. This suggests two things about the iPhone 17 Air. The first is that the battery is so much weaker than other iPhone 17 that it needs a special accessory to help keep it going. The other is that we may not see MagSafe on the super-thin iPhone — but wireless charging is likely.

After running the Tom's Guide battery test on the Samsung Galaxy S25 Edge, that phone's 3,900 mAh battery lasted just 12 hours and 38 minutes. By comparison, the

While not tiny, it's still smaller than the rest of the [Galaxy S25](#), which has a 4,400 mAh battery, lasted 15 hours and 43 minutes, while the [S25 Ultra](#)'s 5,000 mAh battery lasted 17 hours and 14 minutes.

The problem is that phone batteries take up a lot of space, and the bigger the battery the more space is needed to hold it. A lot of work has happened in [compressing batteries](#) and [improving energy efficiency](#), which is why we've seen battery life steadily climb without huge leaps in capacity.

By making the phone thinner, you're limiting the amount of space for a large battery. Which, regardless of how much power efficiency you can score, is never going to produce record-breaking numbers. Not good ones anyway.

And if you have to slap a battery onto the back to keep your phone alive for the entire day, then you've defeated the purpose of actually having a thin phone.

How durable is a thin phone going to be?



(Image credit: @LusiRoy8/X)

Remember back in 2014 when Apple released the [iPhone 6](#)? It was, and still is, the thinnest iPhone ever made at 6.9mm thick. Unfortunately that thickness was a problem, because the phone ended up susceptible to bending — a problem that became known as "[Bendgate](#)."

The iPhone 17 Air is set to be even thinner at 5.5mm, almost a millimeter and a half thinner.

I'm certain that Apple has learned a lot about durability over the past 11 years, and will be working tirelessly to make sure Bendgate 2: Cupertino Boogaloo doesn't happen once the iPhone 17 Air is released. We noted that the 5.8mm S25 Edge felt surprisingly durable in our [Galaxy S25 Edge hands-on](#), so it's a lesson that others took to heart. But no matter how durable super thin phones actually are, that bias is still likely to exist in the minds of many. To make matters worse, even if skinny phones can match the durability of their thicker counterparts, that still isn't particularly durable.

Which means that these phones are probably going to end up [wrapped in some sort of case](#). Those cases will then add extra thickness and weight to the phone, depending on the kind of case people opt for, and there gets to a point where you may as well have not bought the thin model at all.

And for those that prefer a naked phone, are you willing to gamble the higher price tag on the phone being as durable as its counterparts? Because I wouldn't be surprised if Apple gets a lot of Apple Care Plus claims in the months following the Air's release.

Bottom line



(Image credit: Future)

I feel like I'm getting very mixed messages with the thin phones that are on the way. The idea behind them seems to be to emphasize the design and style, showing how phones could look with a few changes . And yet the compromises that have to be made seem to cause their own problems.

I can't escape the prospect of durability being an issue as well. It's not like smartphones are totally impervious to damage as it is, and potentially making them more fragile seems like an issue to me. And

sure, a case can solve the problem, but it defeats the point. More so if you have to take weak batteries into account, and need to slap a battery pack onto the back of your phone.

And the more I think about it, the less appealing these phones seem to be.

Sam Altman The Gentle Singularity

We are past the event horizon; the takeoff has started. Humanity is close to building digital superintelligence, and at least so far it's much less weird than it seems like it should be.

Robots are not yet walking the streets, nor are most of us talking to AI all day. People still die of disease, we still can't easily go to space, and there is a lot about the universe we don't understand.

And yet, we have recently built systems that are smarter than people in many ways, and are able to significantly amplify the output of people using them. The least-likely part of the work is behind us; the scientific insights that got us to systems like GPT-4 and o3 were hard-won, but will take us very far.

AI will contribute to the world in many ways, but the gains to quality of life from AI driving faster scientific progress and increased productivity will be enormous; the future can be vastly better than the present. Scientific progress is the biggest driver of overall progress; it's hugely exciting to think about how much more we could have.

In some big sense, ChatGPT is already more powerful than any human who has ever lived. Hundreds of millions of people rely on it every day and for increasingly important tasks; a small new capability can create a hugely positive impact; a small misalignment multiplied by hundreds of millions of people can cause a great deal of negative impact.

2025 has seen the arrival of agents that can do real cognitive work; writing computer code will never be the same. 2026 will likely see the arrival of systems that can figure out novel insights. 2027 may see the arrival of robots that can do tasks in the real world.

A lot more people will be able to create software, and art. But the world wants a lot more of both, and experts will probably still be much better than novices, as long as they embrace the new tools. Generally speaking, the ability for one person to get much more done in 2030 than they could in 2020 will be a striking change, and one many people will figure out how to benefit from.

In the most important ways, the 2030s may not be wildly different. People will still love their families, express their creativity, play games, and swim in lakes.

But in still-very-important-ways, the 2030s are likely going to be wildly different from any time that has come before. We do not know how far beyond human-level intelligence we can go, but we are about to find out.

In the 2030s, intelligence and energy—ideas, and the ability to make ideas happen—are going to become wildly abundant. These two have been the fundamental limiters on human progress for a long time; with abundant intelligence and energy (and good governance), we can theoretically have anything else.

Already we live with incredible digital intelligence, and after some initial shock, most of us are pretty used to it. Very quickly we go from being amazed that AI can generate a beautifully-written paragraph to wondering when it can generate a beautifully-written novel; or from being amazed that it can make live-saving medical diagnoses to wondering when it can develop the cures; or from being amazed it can create a small computer program to wondering when it can create an entire new company. This is how the singularity goes: wonders become routine, and then table stakes.

We already hear from scientists that they are two or three times more productive than they were before AI. Advanced AI is interesting for many

reasons, but perhaps nothing is quite as significant as the fact that we can use it to do faster AI research. We may be able to discover new computing substrates, better algorithms, and who knows what else. If we can do a decade's worth of research in a year, or a month, then the rate of progress will obviously be quite different.

From here on, the tools we have already built will help us find further scientific insights and aid us in creating better AI systems. Of course this isn't the same thing as an AI system completely autonomously updating its own code, but nevertheless this is a larval version of recursive self-improvement.

There are other self-reinforcing loops at play. The economic value creation has started a flywheel of compounding infrastructure buildout to run these increasingly-powerful AI systems. And robots that can build other robots (and in some sense, datacenters that can build other datacenters) aren't that far off.

If we have to make the first million humanoid robots the old-fashioned way, but then they can operate the entire supply chain—digging and refining minerals, driving trucks, running factories, etc.—to build more robots, which can build more chip fabrication facilities, data centers, etc, then the rate of progress will obviously be quite different.

As datacenter production gets automated, the cost of intelligence should eventually converge to near the cost of electricity. (People are often curious about how much energy a ChatGPT query uses; the average query uses about 0.34 watt-hours, about what an oven would use in a little over one second, or a high-efficiency lightbulb would use in a couple of minutes. It also uses about 0.000085 gallons of water; roughly one fifteenth of a teaspoon.)

The rate of technological progress will keep accelerating, and it will continue to be the case that people are capable of adapting to almost anything. There will be very hard parts like whole classes of jobs going away, but on the other hand the world will be getting so much richer so quickly that we'll be able to seriously entertain new policy ideas we never could before. We probably won't adopt a new social contract all at once, but when we look back in a few decades, the gradual changes will have amounted to something big.

If history is any guide, we will figure out new things to do and new things to want, and assimilate new tools quickly (job change after the industrial revolution is a good recent example). Expectations will go up, but capabilities will go up equally quickly, and we'll all get better stuff. We will build ever-more-wonderful things for each other. People have a long-term important and curious advantage over AI: we are hard-wired to care about other people and what they think and do, and we don't care very much about machines.

A subsistence farmer from a thousand years ago would look at what many of us do and say we have fake jobs, and think that we are just playing games to entertain ourselves since we have plenty of food and unimaginable luxuries. I hope we will look at the jobs a thousand years in the future and think they are very fake jobs, and I have no doubt they will feel incredibly important and satisfying to the people doing them.

The rate of new wonders being achieved will be immense. It's hard to even imagine today what we will have discovered by 2035; maybe we will go from solving high-energy physics one year to beginning space colonization the next year; or from a major materials science breakthrough one year to true high-bandwidth brain-computer interfaces the next year. Many people will choose to live their lives in much the same way, but at least some people will probably decide to "plug in".

Looking forward, this sounds hard to wrap our heads around. But probably living through it will feel impressive but manageable. From a relativistic perspective, the singularity happens bit by bit, and the merge happens slowly. We are climbing the long arc of exponential technological progress; it always looks vertical looking forward and flat going backwards, but it's one smooth curve. (Think back to 2020, and what it would have sounded like to have something close to AGI by 2025, versus what the last 5 years have actually been like.) There are serious challenges to confront along with the huge upsides. We do need to solve the safety issues, technically and societally, but then it's critically important to widely distribute access to superintelligence given the economic implications. The best path forward might be something like:

Solve the alignment problem, meaning that we can robustly guarantee that we get AI systems to learn and act towards what we collectively really want over the long-term (social media feeds are an example of misaligned AI; the algorithms that power those are incredible at getting you to keep scrolling and clearly understand your short-term preferences, but they do so by exploiting something in your brain that overrides your long-term preference).

Then focus on making superintelligence cheap, widely available, and not too concentrated with any person, company, or country. Society is resilient, creative, and adapts quickly. If we can harness the collective will and wisdom of people, then although we'll make plenty of mistakes and some things will go really wrong, we will learn and adapt quickly and be able to use this technology to get maximum upside and minimal downside. Giving users a lot of freedom, within broad bounds society has to decide on, seems very important. The sooner the world can start a conversation about what these broad bounds are and how we define collective alignment, the better.

We (the whole industry, not just OpenAI) are building a brain for the world. It will be extremely personalized and easy for everyone to use; we will be limited by good ideas. For a long time, technical people in the startup industry have made fun of "the idea guys"; people who had an idea and were looking for a team to build it. It now looks to me like they are about to have their day in the sun. OpenAI is a lot of things now, but before anything else, we are a superintelligence research company. We have a lot of work in front of us, but most of the path in front of us is now lit, and the dark areas are receding fast. We feel extraordinarily grateful to get to do what we do.

Intelligence too cheap to meter is well within grasp. This may sound crazy to say, but if we told you back in 2020 we were going to be where we are today, it probably sounded more crazy than our current predictions about 2030.

May we scale smoothly, exponentially and uneventfully through super intelligence.

AI models aren't as smart as first thought due to major reason

June 11, 2025

This limitation stems from the fact that AI lacks true consciousness and understanding of context. As a result, AI systems rely on predefined rules and patterns to make decisions. While AI can process vast amounts of data quickly, it struggles with abstract thinking and complex problem-solving. Overcoming the challenge of reasoning remains a key area of research in the field of artificial intelligence.



Representational image

Nowadays, everywhere you turn online you'll be getting AI shoved down your throat. From your social media feed churning out AI

pictures to politicians hailing it as the next panacea to society's problems, it's clear this new technology is here to stay.

Fears of AI

Fears of AI causing some kind of world war or taking over and becoming our overlords have been around for decades but have ramped up in recent years. Except some fears are overblown with AI failing at some very simple tasks, especially when it comes to reasoning. Researchers at Apple say that the top AI models, including OpenAI's o3, Anthropic's Claude 3.7, and Google's Gemini, go through near total collapse when faced by complex reason tasks.

US territory wants to rejoin Spain and become 'like the Canary Islands' Trump warns 'we are going to have troops everywhere' after ICE protests. The new study was published June 7 on Apple's Machine Learning Research website. "While these models demonstrate improved performance on reasoning benchmarks, their fundamental capabilities, scaling properties, and limitations remain insufficiently understood," the boffins behind the paper wrote.

They used "controllable puzzle environments" to estimate the AI models' ability to "think" and found something that undermines everyone's obsession with the new tech. "Through extensive experimentation across diverse puzzles, we show that frontier [large reasoning models] face a complete accuracy collapse beyond certain complexities," they wrote.

The problem is that the models appear to be 'overthinking,' when given complex problems, even ones with what should be sufficient training. Although some will point at Apple lagging behind in the AI race as a key driver of a report like this, it does sit within a wider trend of the latest generations of reasoning models being more prone to 'hallucinating' - which is when they just make up information and

deliver it as fact. An OpenAI technical report has highlighted that reasoning models are much more likely to be derailed by hallucinations than their generic counterparts, with the problem only getting worse as models advance. It could indicate that the technology is heading in the wrong direction.

Implications of AI

This comes as society grapples with the implications of AI. The rise of artificial intelligence and the popularity of online classes have led to an explosion of financial aid fraud. Fake college enrollments have been surging as crime rings deploy “ghost students” — chatbots that join online classrooms and stay just long enough to collect a financial aid check. In some cases, professors discover almost no one in their class is real. Students get locked out of the classes they need to graduate as bots push courses over their enrollment limits. And victims of identity theft who discover loans fraudulently taken out in their names must go through months of calling colleges, the Federal Student Aid office and loan servicers to try to get the debt erased.

ing Data and AI to Every Corner of the Business

Databricks Unveils Databricks One: A New Experience to Bring Data and AI to Every Corner of the Business

June 11, 2025

Databricks has unveiled a new product called Databricks One, designed to provide data intelligence to executives and decision-makers. This new offering aims to empower both executives and frontline workers with insights derived from data and artificial intelligence. Databricks One promises to enhance decision-making processes by making data analytics more accessible across all levels of an organization. The launch of Databricks One marks a significant step towards democratizing data intelligence within businesses. This product is anticipated to revolutionize the way organizations leverage data for strategic decision-making.

New product brings the power of data intelligence to executives and frontline decision-makers SAN FRANCISCO , June 11, 2025

/PRNewswire/ -- Data + AI Summit -- Databricks , the Data and AI company, today announced the launch of Databricks One , a new

experience that gives all business users simple and secure access to the data and AI capabilities offered by the Data Intelligence Platform . For the first time, business users will be able to interact with AI/BI dashboards, ask their data questions in natural language through AI/BI Genie powered by deep research, quickly find relevant dashboards and use custom-built Databricks apps — all in an elegant, code-free environment built for their needs. This marks a major evolution for Databricks to extend data intelligence beyond technical users to business teams across the enterprise. Databricks One is launching in Private Preview starting today.

Databricks Empowering Business Users

Databricks has long empowered engineers and data scientists to deliver advanced analytics, build robust data pipelines and unlock deep organizational insights. With AI/BI Dashboards, Genie and Databricks Apps, the Databricks Data Intelligence Platform has much to offer non-technical business users from executive teams to finance, marketing, sales and more, but the technical environment of the Databricks workspace can be a barrier. At the heart of Databricks One is a completely redesigned interface tailored for business users — a simple, intuitive home for non-technical users to access exactly what they need. Anyone can now talk to their data to understand the what, why and how, such as "What region is growing fastest?" "Why did sales spike in April?" and "How can we continue accelerating growth?"

Databricks One Mission

"Our mission at Databricks is to democratize data + AI," said Ali Ghodsi , Co-founder and CEO of Databricks. "Every person of every skill level should have equal access to work with data and use AI. With

Databricks One, we want to make our experience for non-technical users as amazing as our experience for technical users. This is our first step of making this true so that everyone across the organization can unlock the full value of their data and drive innovation."

Simple but Powerful Analytics with AI/BI

With AI/BI Dashboards, teams can explore beautifully rendered, lightning-fast visualizations and perform advanced analytics like forecasting and key driver analysis — all without writing a single line of code. The experience features AI/BI Genie, a conversational assistant that delivers deep, contextual insights on demand. Ask questions in plain language and get instant, accurate visual answers backed by data. Built on the Data Intelligence Platform, Genie understands business-specific semantics and will soon support "Deep Research," helping users explore not just what happened, but why and how, by analyzing root causes and summarizing findings.

Go beyond analytics with Databricks Apps

Users will also have secure access to Databricks Apps through Databricks One. Databricks Apps provides a simple and versatile way for organizations to package up complex workflows that interweave analytics, AI and transactional processing in a custom app that's built for a particular problem.

Built on the Data Intelligence Platform

Databricks One is built on the proven foundations of the Databricks Data Intelligence Platform. This offers a number of key benefits:

- **Unified Governance and Security:** Databricks One's governance model is built on Unity Catalog. Administrators can

centrally define fine-grained access control policies such as RLS/CLS, organize content by business domains and audit consumption activity to confidently provide access to this expanded audience.

- **Simplified Identity Management:** The Databricks IAM platform for Databricks One integrates directly with IDPs like Entra AD or Okta and does not have seat-based restrictions, so administrators can easily add anyone within their organization without having to worry about procuring new licenses.
- **Industry-Leading Price/Scale and Performance:** Databricks One is powered by Databricks' performant and efficient serverless compute platform, which delivers interactive performance at the scale of the largest enterprises at a market-leading price/performance.

Customer Momentum

"Through Databricks AI/BI, we developed a suite of dashboards—each of which could benefit any member of Wineshipping. With Databricks One, we will be able to securely deliver these dashboards in an environment that is intuitive and accessible for all users." – Christina Mottolo , VP, Business Applications, Wineshipping

"Genie is changing how we empower teams at HP. With its intuitive natural language interface, we're enabling users to instantly access critical data, no coding required, making data-driven insights faster and more efficiently." — Bruce Hillsberg , VP, Data Engineering & Insights, Technology & Innovation, HP Inc

"Our business teams have always relied on analysts for sales data and insights. Now, with Unity Catalog and Genie, we're empowering them to access data directly." — Shahmeer Mirza, Senior Director of Data, AI/ML, and R&D, 7-Eleven

Availability

Databricks One is available to customers of the Databricks Data Intelligence Platform starting today in Private Preview and will be available to all customers in Beta later this summer. Databricks One is available to all customers at no extra cost.

About Databricks

Databricks is a Data and AI company. More than 15,000 organizations worldwide — including Block, Comcast, Condé Nast, Rivian, Shell, and over 60% of the Fortune 500 — rely on the Databricks Data Intelligence Platform to take control of their data and put it to work with AI. Databricks is headquartered in San Francisco, with offices around the globe, and was founded by the original creators of Lakehouse, Apache Spark™, Delta Lake, MLflow, and Unity Catalog. To learn more, follow Databricks on X and LinkedIn.

Open Source

Introducing the V-JEPA 2 world model and new benchmarks for physical reasoning

June 11, 2025•

Takeaways

- Meta Video Joint Embedding Predictive Architecture 2 (V-JEPA 2) is a world model that achieves state-of-the-art performance on visual understanding and prediction in the physical world. Our model can also be used for zero-shot robot planning to interact with unfamiliar objects in new environments.
- V-JEPA 2 represents our next step toward our goal of achieving advanced machine intelligence (AMI) and building useful AI agents that can operate in the physical world.
- We're also releasing three new benchmarks to evaluate how well existing models can reason about the physical world from video.

Today, we're excited to share V-JEPA 2, the first world model trained on video that enables state-of-the-art understanding and prediction, as well as zero-shot planning and robot control in new environments. As we work toward our goal of achieving advanced machine intelligence (AMI), it will be important that we have AI systems that can learn about the world as humans do, plan how to execute unfamiliar tasks, and efficiently adapt to the ever-changing world around us.

V-JEPA 2 is a 1.2 billion-parameter model that was built using [Meta Joint Embedding Predictive Architecture](#) (JEPA), which we first shared in 2022. Our previous work has shown that JEPA performs well for modalities like [images](#) and [3D point clouds](#). Building on [V-JEPA](#), our first model trained on video that we released last year, V-JEPA 2 improves action prediction and world modeling capabilities that enable robots to interact with unfamiliar objects and environments to complete a task. We're also sharing three new benchmarks to help the research community evaluate how well their existing models learn and reason about the world using video. By sharing this work, we aim to give researchers and developers access to the best models and benchmarks to help accelerate research and progress—ultimately leading to better and more capable AI systems that will help enhance people's lives.

What are world models?

We all know that if you toss a tennis ball into the air, gravity will pull it back down. It would be surprising if it hovered, suddenly pivoted mid-air and went flying in a different direction, or spontaneously changed into an apple. That kind of physical intuition isn't something adults obtain after years of education—young children develop this intuition by observing the world around them before they can even speak in full sentences.

The ability to predict how the world will respond to our actions—or the actions of others—is something humans use all the time, especially when planning what actions to take and how to best navigate new situations. Consider all the ways this physical intuition shows up in our everyday lives. When we walk through an unfamiliar crowded area, we're making moves toward our destination while also trying not to bump into people or obstacles along the path. When playing hockey, we skate to where the puck is going, not where it currently is. And when preparing a meal using a stove, we think about how much longer to leave the pot on the flame or whether to turn down the heat. Our internal model of the world provides us with this intuition and also acts as an internal simulator, allowing us to predict

the outcome of a hypothetical action, so we can ultimately choose the best action based on what we believe will best achieve our goal.

Before taking action, we use our world model to imagine the potential consequences. As we work toward building AI agents that can similarly think before they act, it's important that they learn world models that enable the following capabilities:

- **Understanding:** A world model should be able to understand observations of the world, including things like recognizing objects, actions, and motions in a video.
- **Predicting:** A world model should be able to make predictions about how the world will evolve, and how the world will change if the agent takes an action.
- **Planning:** Building on the ability to make predictions, a world model should be useful for planning sequences of actions that achieve a given goal.

Introducing V-JEPA 2

Our long-term vision is that world models will enable AI agents to plan and reason in the physical world. As the next step towards this vision, we're releasing V-JEPA 2, a world model trained primarily on video—a rich and readily available source of information about the world. By making V-JEPA 2 code and model checkpoints available for commercial and research applications, we hope to build a broad community around this research, driving progress toward our ultimate goal of developing world models that can transform the way AI interacts with the physical world.

Built using a joint-embedding predictive architecture (JEPA), V-JEPA 2 has two main components:

- An **encoder**, which takes in raw video and outputs embeddings that capture useful semantic information about the state of the observed world.
- A **predictor**, which takes in a video embedding and additional context about what to predict and outputs predicted embeddings.

We train V-JEPA 2 using self-supervised learning from video, which allows us to train on video without requiring additional human annotation. V-JEPA 2 training involves two stages: actionless pre-training, followed by additional action-conditioned training.

In the first stage—pre-training—we use more than 1 million hours of video and 1 million images from diverse sources. This rich visual data helps the model learn a lot about how the world works, including how people interact with objects, how objects move in the physical world, and how objects interact with other objects. We find that the model already demonstrates key capabilities related to understanding and prediction after the pre-training stage. For example, by training a lightweight attentive read-out on top of the frozen encoder features, V-JEPA 2 achieves exceptional performance on the Something-Something v2 action recognition task, which relies on motion understanding. Similarly, by training an attentive read-out on top of the frozen encoder and predictor features, V-JEPA 2 sets a new state-of-the-art on the Epic-Kitchens-100 action anticipation task of predicting what action (comprised of a noun and a verb) will be performed 1 second into the future from egocentric video. Finally, aligning V-JEPA 2 with a language model results in state-of-the-art performance on video question answering benchmarks such as Perception Test and TempCompass.

After the actionless pre-training stage, the model can make predictions about how the world might evolve—however, these predictions don’t directly take into account specific actions that an agent would take. In the second stage of training, we focus on making the model more useful for planning by using robot data, which includes visual observations (video) and the control actions that the robot was executing. We incorporate this data into the JEPA training procedure by providing the action information to the predictor. After training on this additional data, the predictor learns to account for specific actions when making predictions and can then be used for control. We don’t need a lot of robot data for this second phase—in our technical report, we show that training with only 62 hours of robot data already results in a model that can be used for planning and control.

We demonstrate how V-JEPA 2 can be used for zero-shot robot planning in new environments and involving objects not seen during training. Unlike other robot foundation models—which usually require that some training data come from the specific robot instance and environment where the model is deployed—we train the model on the open source [DROID dataset](#) and then deploy it directly on robots in our labs. We show that the V-JEPA 2 predictor can be used for foundational tasks like reaching, picking up an object, and placing it in a new location.

For short-horizon tasks, such as picking or placing an object, we specify a goal in the form of an image. We use the V-JEPA 2 encoder to get embeddings of the current and goal states. Starting from its observed current state, the robot then plans by using the predictor to

imagine the consequences of taking a collection of candidate actions and rating the candidates based on how close they get to the desired goal. At each time step, the robot re-plans and executes the top-rated next action toward that goal via model-predictive control. For longer horizon tasks, such as picking up an object and placing it in the right spot, we specify a series of visual subgoals that the robot tries to achieve in sequence, similar to visual imitation learning observed in humans. With these visual subgoals, V-JEPA 2 achieves success rates of 65% – 80% for pick-and-placing new objects in new and unseen environments.

Download the V-JEPA artifacts:

[GitHub](#)

[Hugging Face](#)

[Visit the V-JEPA 2 website](#)

[Read the paper](#)

Benchmarking physical understanding

As we continue to make advancements in the field of world models, we're excited to share our work and support progress in the open source community. We're releasing three new benchmarks to evaluate how well existing models can understand and reason about the physical world from video. While humans perform well on all three benchmarks (85% – 95% accuracy), there's a notable gap between human performance and that of top models including V-JEPA 2, indicating important directions for models to improve in.

IntPhys 2 is specifically designed to measure the ability of models to distinguish between physically plausible and implausible scenarios, building and expanding upon the earlier [IntPhys benchmark](#). We designed IntPhys 2 similar to the way developmental cognitive scientists evaluate when young humans acquire intuitive physics, via the violation of expectations paradigm. We achieve this using a game engine that generates pairs of videos, where the two videos are identical up to a certain point, and then a physics-breaking event occurs in one of the two videos. The model must then identify which video

has the physics-breaking event. While humans achieve near-perfect accuracy on this task across a range of scenarios and conditions, we find that current video models are at or close to chance.

Download IntPhys 2:

[GitHub](#)

[Hugging Face](#)

[Read the paper](#)

Minimal Video Pairs (MVPBench) measures the physical understanding abilities of video-language models via multiple choice questions. Unlike other video question-answering benchmarks in the literature, MVPBench is designed to mitigate common shortcut solutions that have been observed in video-language models, such as relying on superficial visual or textual cues and biases. Each example in MVPBench has a minimal-change pair: a visually similar video together with the same question but with an opposing answer. In order to get credit for one example, a model must also get its minimal-change pair correct.

Download Minimal Video Pairs (MVPBench):

[GitHub](#)

[Hugging Face](#)

[Read the paper](#)

CausalVQA measures the ability of video-language models to answer questions related to physical cause-and-effect. The benchmark is designed to focus on causal understanding in physical-world videos, including questions about counterfactuals (*what would have happened if...*), anticipation (*what might happen next*), and planning (*what action should occur next to accomplish a goal*). We find that while large multimodal models are increasingly capable of answering questions about “what happened” in the

video, they still struggle to answer questions about “what could have happened” and “what might happen next,” revealing a substantial gap with respect to human performance on predicting how the physical world will likely evolve given the space of actions and events.

Download CausalVQA:

[GitHub](#)

[Read the paper](#)

We’re also publishing a Leaderboard on Hugging Face to help the community track model progress against these new benchmarks.

[View the Hugging Face Leaderboard](#)

Next steps along the path to advanced machine intelligence

There are several areas we plan to explore further as we continue our work on world models. Currently, V-JEPA 2 learns and makes predictions at a single time scale. However, many tasks require planning across multiple time scales. Think of breaking down a high-level task into smaller steps, such as loading the dishwasher or baking a cake. We want to focus on training hierarchical JEPA models that are capable of learning, reasoning, and planning across multiple temporal and spatial scales. Another important direction will be multimodal JEPA models that can make predictions using a variety of senses, including vision, audio, and touch. As always, we look forward to sharing more in the future and continuing the important discussions we’re having with the research community.

JUNE 10, 2025

AI-enabled control system helps autonomous drones stay on target in uncertain environments

by Adam Zewe, [Massachusetts Institute of Technology](#)

edited by [Sadie Harley](#), reviewed by [Robert Egan](#)

An autonomous drone carrying water to help extinguish a wildfire in the Sierra Nevada might encounter swirling Santa Ana winds that threaten to push it off course. Rapidly adapting to these unknown disturbances in flight presents an enormous challenge for the drone's flight control system.

To help such a [drone](#) stay on target, MIT researchers developed a new, machine learning-based adaptive control algorithm that could minimize its deviation from its intended trajectory in the face of unpredictable forces like gusty winds.

The study is [published](#) on the *arXiv* preprint server.

Unlike standard approaches, the new technique does not require the person programming the autonomous drone to know anything in advance about the structure of these uncertain disturbances.

Instead, the control system's artificial intelligence model learns all it needs to know from a small amount of observational data collected from 15 minutes of flight time.

Importantly, the technique automatically determines which [optimization algorithm](#) it should use to adapt to the disturbances, which improves tracking performance. It chooses the algorithm that best suits the geometry of specific disturbances this drone is facing.

The researchers train their control system to do both things simultaneously using a technique called meta-learning, which teaches the system how to adapt to different types of disturbances.

Taken together, these ingredients enable their adaptive control system to achieve 50% less trajectory tracking error than baseline methods in simulations and perform better with new wind speeds it didn't see during training.

In the future, this adaptive control system could help autonomous drones more efficiently deliver heavy parcels despite strong winds or monitor fire-prone areas of a national park.

"The concurrent learning of these components is what gives our method its strength. By leveraging meta-learning, our controller can automatically make choices that will be best for quick adaptation," says Navid Azizan, who is the Esther and Harold E. Edgerton Assistant Professor in the MIT Department of Mechanical Engineering and the Institute for Data, Systems, and Society (IDSS), a principal investigator of the Laboratory for Information and Decision Systems (LIDS), and the senior author of the paper.

Azizan is joined on the paper by lead author Sunbochen Tang, a graduate student in the Department of Aeronautics and Astronautics, and Haoyuan Sun, a graduate student in the Department of Electrical Engineering and Computer Science. The research was also presented at the Learning for Dynamics and Control Conference.

Finding the right algorithm

Typically, a control system incorporates a function that models the drone and its environment, and includes some existing information on the structure of potential disturbances. But in a real world filled with uncertain conditions, it is often impossible to hand-design this structure in advance.

Many control systems use an adaptation method based on a popular optimization algorithm, known as gradient descent, to estimate the unknown parts of the problem and determine how to keep the drone as close as possible to its target trajectory during flight. However, gradient descent is only one algorithm in a larger family of algorithms available to choose, known as mirror descent.

"Mirror descent is a general family of algorithms, and for any given problem, one of these algorithms can be more suitable than others. The name of the game is how to choose the particular algorithm that is right for your problem. In our method, we automate this choice," Azizan says.

In their control system, the researchers replaced the function that contains some structure of potential disturbances with a neural network model that learns to approximate them from data. In this way, they don't need to have an a priori structure of the wind speeds this drone could encounter in advance.

Their method also uses an algorithm to automatically select the right mirror-descent function while learning the neural network model from data, rather than assuming a user has the ideal function picked out already. The researchers give this algorithm a range of functions to pick from, and it finds the one that best fits the problem at hand.

"Choosing a good distance-generating function to construct the right mirror-descent adaptation matters a lot in getting the right algorithm to reduce the tracking error," Tang adds.

Learning to adapt

While the wind speeds the drone may encounter could change every time it takes flight, the controller's neural network and mirror function should stay the same so they don't need to be recomputed each time.

To make their controller more flexible, the researchers use meta-learning, teaching it to adapt by showing it a range of wind speed families during training.

"Our method can cope with different objectives because, using meta-learning, we can learn a shared representation through different scenarios efficiently from data," Tang explains.

In the end, the user feeds the control system a target trajectory and it continuously recalculates, in real-time, how the drone should produce thrust to keep it as close as possible to that trajectory while accommodating the uncertain disturbance it encounters.

In both simulations and real-world experiments, the researchers showed that their method led to significantly less trajectory tracking error than baseline approaches with every wind speed they tested.

"Even if the wind disturbances are much stronger than we had seen during training, our technique shows that it can still handle them successfully," Azizan adds.

In addition, the margin by which their method outperformed the baselines grew as the wind speeds intensified, showing that it can adapt to challenging environments.

The team is now performing hardware experiments to test their control system on real drones with varying wind conditions and other disturbances.

They also want to extend their method so it can handle disturbances from multiple sources at once. For instance, changing [wind](#) speeds could cause the weight of a parcel the drone is carrying to shift in flight, especially when the drone is carrying sloshing payloads.

They also want to explore continual learning, so the drone could adapt to new disturbances without the need to also be retrained on the data it has seen so far.

"Navid and his collaborators have developed breakthrough work that combines meta-learning with conventional adaptive control to learn nonlinear features from data," says Babak Hassibi, the Mose and Lillian S. Bohn Professor of Electrical Engineering and Computing and Mathematical Sciences at Caltech, who was not involved with this work.

"Key to their approach is the use of mirror descent techniques that exploit the underlying geometry of the problem in ways prior art could not. Their work can contribute significantly to the design of autonomous systems that need to operate in complex and uncertain environments."

More information: Sunbochen Tang et al, Meta-Learning for Adaptive Control with Automated Mirror Descent, *arXiv* (2024). DOI: [10.48550/arxiv.2407.20165](https://doi.org/10.48550/arxiv.2407.20165)

Journal information: [arXiv](#)

Provided by [Massachusetts Institute of Technology](#)

JUNE 10, 2025

Multi-modal AI agent mimics human thinking for long video analysis and reasoning

by Hazel Cheng, Edmond Siu, [Hong Kong Polytechnic University](#)

edited by [Gaby Clark](#), reviewed by [Robert Egan](#)

Credit: GitHub: <https://github.com/yeliudev/VideoMind>

While Artificial Intelligence (AI) technology is evolving rapidly, AI models still struggle with understanding long videos. A research team from The Hong Kong Polytechnic University (PolyU) has developed a novel video-language agent, VideoMind, that enables AI models to perform long video reasoning and question-answering tasks by emulating humans' way of thinking.

The [VideoMind](#) framework incorporates an innovative Chain-of-Low-Rank Adaptation (LoRA) strategy to reduce the demand for computational resources and power, advancing the application of generative AI in video analysis. The findings have been submitted to the world-leading AI conferences.

Videos, especially those longer than 15 minutes, carry information that unfolds over time, such as the sequence of events, causality, coherence and scene transitions. To understand the [video content](#), AI models therefore need not only to identify the objects present, but also take into account how they change throughout the video. As visuals in videos occupy a large number of tokens, video understanding requires vast amounts of computing capacity and memory, making it difficult for AI models to process long videos.

Prof. Changwen Chen, Interim Dean of the PolyU Faculty of Computer and Mathematical Sciences and Chair Professor of Visual Computing, and his team have achieved a breakthrough in research on long video reasoning by AI. In designing VideoMind, they made reference to a human-like process of video understanding, and introduced a role-based workflow. The four roles included in the framework are:

- the Planner, to coordinate all other roles for each query;
- the Grounder, to localize and retrieve relevant moments;
- the Verifier, to validate the information accuracy of the retrieved moments and select the most reliable one;

- and the Answerer, to generate the query-aware answer.

This progressive approach to video understanding helps address the challenge of temporal-grounded reasoning that most AI models face.

Another core innovation of the VideoMind framework lies in its adoption of a Chain-of-LoRA strategy. LoRA is a finetuning technique that has emerged in recent years. It adapts AI models for specific uses without performing full-parameter retraining. The innovative chain-of-LoRA strategy pioneered by the team involves applying four lightweight LoRA adapters in a unified model, each of which is designed for calling a specific role.

With this strategy, the model can dynamically activate role-specific LoRA adapters during inference via self-calling to seamlessly switch among these roles, eliminating the need and cost of deploying multiple models while enhancing the efficiency and flexibility of the single model.

VideoMind is [open source](#) on GitHub and Huggingface, and the related research is [available](#) on the *arXiv* preprint server. Details of the experiments conducted to evaluate its effectiveness in temporal-grounded video understanding across 14 diverse benchmarks are also available. Comparing VideoMind with some state-of-the-art AI models, including GPT-4o and Gemini 1.5 Pro, the researchers found that the grounding accuracy of VideoMind outperformed all competitors in challenging tasks involving videos with an average duration of 27 minutes.

Notably, the team included two versions of VideoMind in the experiments: one with a smaller, 2 billion (2B) parameter model, and another with a bigger, 7 billion (7B) parameter model. The results showed that, even at the 2B size, VideoMind still yielded performance comparable with many of the other 7B size models.

Prof. Chen said, "Humans switch among different thinking modes when understanding videos: breaking down tasks, identifying relevant moments, revisiting these to confirm details and synthesizing their observations into coherent answers. The process is very efficient with the human brain using only about 25 watts of power, which is about a million times lower than that of a supercomputer with equivalent computing power.

"Inspired by this, we designed the role-based workflow that allows AI to understand videos like human, while leveraging the chain-of-LoRA strategy to minimize the need for computing power and memory in this process."

AI is at the core of global technological development. The advancement of AI models is however constrained by insufficient computing power and excessive power consumption. Built upon a unified, open-source model Qwen2-VL and augmented with additional optimization tools, the VideoMind framework has lowered the technological cost and the threshold for deployment, offering a feasible solution to the bottleneck of reducing power consumption in AI models.

Prof. Chen added, "VideoMind not only overcomes the performance limitations of AI models in video processing, but also serves as a modular, scalable and interpretable multimodal

reasoning framework. We envision that it will expand the application of generative AI to various areas, such as intelligent surveillance, sports and entertainment video analysis, video search engines and more."

More information: Ye Liu et al, VideoMind: A Chain-of-LoRA Agent for Long Video Reasoning, *arXiv* (2025). DOI: [10.48550/arxiv.2503.13444](https://doi.org/10.48550/arxiv.2503.13444)

Journal information: [arXiv](#)

Provided by [Hong Kong Polytechnic University](#)

JUNE 10, 2025

'Optical neural engine' can solve partial differential equations

by [University of Utah](#) edited by [Robert Egan](#)

The researchers' device applies principles of neural networking to an optical framework. As a wave encoded with a PDE passes through the ONE's series of components, its properties gradually shift and change, until they ultimately represent the solution to the given PDE.

Credit: Gao lab, University of Utah

Partial differential equations (PDEs) are a class of mathematical problems that represent the interplay of multiple variables, and therefore have predictive power when it comes to complex physical systems. Solving these equations is a perpetual challenge, however, and current computational techniques for doing so are time-consuming and expensive.

Now, research from the University of Utah's John and Marcia Price College of Engineering is showing a way to speed up this process: encoding those equations in light and feeding them into their newly designed "optical neural engine," or ONE.

The researchers' ONE combines diffractive optical neural networks and optical matrix multipliers. Rather than representing PDEs digitally, the researchers represented them optically, with variables represented by the various properties of a light wave, such as its intensity and phase. As a wave passes through the ONE's series of optical components, those properties gradually shift and change, until they ultimately represent the solution to the given PDE.

The research was led by Weilu Gao, assistant professor in the Department of Electrical & Computer Engineering, and Ruiyang Chen, a Ph.D. candidate in Gao's research group. They published a study demonstrating this optical neural engine in the journal [Nature Communications](#).

"Partial differential equations are a powerful computational tool to simulate physics problems instead of performing expensive and time-consuming real-world experiments," Chen said. "However, the current numerical simulation method is slow and requires a lot of computing resources, and even electronic machine learning techniques are not fast enough."

The machine learning techniques and digital neural networks currently used to solve PDEs involve passing the equation through a network of computational nodes, each of which weights its output as it passes it on to the next node. As the signal makes its way through the network, the correct solution becomes weighted most heavily and ends up as the output.

The researchers' ONE takes this concept and applies it to photonic devices.

"The ONE takes the spatiotemporal data of an input physical quantity, which is a function of positions and time, to predict the spatiotemporal data of an output physical quantity as a function of positions and time," Gao said.

Electronic machine learning techniques can produce a similar output, but at [slower speeds](#) and higher energy costs.

"This optical approach accelerates the machine learning process and requires less energy as compared to an electronic approach," said lead author and former Gao lab member Yingheng Tang, now a research scientist with Lawrence Berkeley National Laboratory.

The researchers demonstrated their ONE's capabilities on a number of PDEs, including the Darcy flow equation, the magnetostatic Poisson's equation in demagnetization, and the Navier-Stokes equation in incompressible fluid.

"The Darcy flow equation, for example, describes a fluid flow through a porous medium," Gao said. "Given data about the permeability and pressure fields inside a given medium, the ONE architecture essentially learns the mapping between those qualities, and can predict flow properties without having to experiment."

"This research offers a versatile and powerful platform for large-scale scientific and engineering computation, such as geology and chip design," Gao said.

More information: Yingheng Tang et al, Optical neural engine for solving scientific partial differential equations, *Nature Communications* (2025). DOI: [10.1038/s41467-025-59847-3](https://doi.org/10.1038/s41467-025-59847-3)
Journal information: [Nature Communications](#)
Provided by [University of Utah](#)

JUNE 11, 2025

Bio-mimetic robotic hand seamlessly integrates tactile feedback to outperform predecessors

by [Ingrid Fadelli](#), Phys.org edited by [Lisa Lock](#), reviewed by [Robert Egan](#)

Comparison between F-TAC Hand and human hand. Credit: Zhao et al.

Over the past decades, roboticists have developed increasingly advanced systems that can emulate some human capabilities and effectively tackle various real-world tasks. To reliably grasp, manipulate and utilize objects in their surroundings, robots should be able to detect and process tactile information, replicating the processes underpinning the human sense of touch.

Researchers at Peking University, Beijing Institute for General Artificial Intelligence and Queen Mary University of London recently developed F-TAC hand, a bio-mimetic robotic hand designed to imitate human tactile perception. This hand, introduced in a paper [published](#) in *Nature Machine Intelligence*, was found to outperform various existing robotic hands on tactile manipulation tasks, by collecting detailed [tactile information](#) that can be used to plan the hand's movements and trajectories.

"This research was driven by the goal of enabling machines to manipulate objects with precision and adaptability akin to human capabilities," Yixin Zhu, the corresponding author of the paper, told Tech Xplore. "Our inspiration to address this problem stemmed from the intricate tactile sensing capabilities observed in the human hand and the specialized neural processing within the primary somatosensory cortex, which efficiently interprets and integrates vast amounts of sensory input in real-time."

The main goal of this recent study by Zhu and his colleagues was to develop a new robotic hand that can seamlessly integrate tactile feedback collected across its surface with sensory-motor mechanisms. As most existing robotic hands fail to do so, such a hand could perform better on various real-world manipulation tasks.

Comprehensive tactile sensing capabilities of F-TAC Hand. Credit: *Nature Machine Intelligence* (2025). DOI: 10.1038/s42256-025-01053-3

"The F-TAC Hand replicates the biological structure of the human hand by integrating 17 high-resolution tactile sensors across 70% of the palm's surface, achieving an impressive spatial resolution of 0.1 millimeters—equivalent to 10,000 tactile pixels per square centimeter," explained Zhu.

"A probabilistic algorithm enables the generation of diverse, human-like grasping strategies spanning 19 common grasp types. This allows the F-TAC Hand to adapt dynamically, switching strategies within just 100 milliseconds when faced with unexpected scenarios."

The most notable advantage of the bio-mimetic hand developed by this research team is that it can pick-up tactile data across its entire surface, utilizing this data to plan its own movements. To assess its performance, Zhu and his colleagues carried out extensive experiments, during which the F-TAC Hand completed a wide range of manual tasks in real-world scenarios.

High-resolution tactile sensing of F-TAC Hand. Credit: Zhao et al.

"In 600 real-world experiments, the F-TAC Hand demonstrated a significant improvement in multi-object grasping success rates, increasing from 53.5% to 100% compared to systems lacking tactile-informed feedback," said Zhu. "Potential practical applications include surgical assistance, high-precision assembly, aerospace operations, and emergency response scenarios that demand exceptional operational precision."

The researchers' initial findings highlight the advantages of the new hand they developed, suggesting that it could help to boost the sensory and motor capabilities of robots. In the future, the hand could be improved further, integrated with different humanoid robots, and tested in further experiments to assess its potential for specific applications.

"We now plan to continue deepening the integration of tactile sensing and robotic control, exploring more intelligent somatosensory interaction paradigms," added Zhu. "We aim to advance embodied intelligence by combining high-fidelity physical sensing with intelligent control systems, contributing to the development of truly [general artificial intelligence](#)."

*Written for you by our author [Ingrid Fadelli](#), edited by [Sadie Harley](#), and fact-checked and reviewed by [Robert Egan](#)—this article is the result of careful human work. We rely on readers like you to keep independent science journalism alive. If this reporting matters to you, please consider a [donation](#) (especially monthly). You'll get an **ad-free** account as a thank-you.*

More information: Zihang Zhao et al, Embedding high-resolution touch across robotic hands enables adaptive human-like grasping, *Nature Machine Intelligence* (2025). DOI: [10.1038/s42256-025-01053-3](https://doi.org/10.1038/s42256-025-01053-3)

Journal information: [Nature Machine Intelligence](#)

© 2025 Science X Network

JUNE 11, 2025

Chinese researchers debut world's first AI-based processor chip design system

by [Bob Yirka](#), Phys.org edited by [Lisa Lock](#), reviewed by [Robert Egan](#)

Left: Multimodal architecture of LPCM capable of understanding, representing, and generating both text and graph data. Right: Feedback-driven inference of LPCM with a dual-loop mechanism, consisting of an outer performance feedback loop and an inner functional correctness feedback loop. Credit: *arXiv* (2025). DOI: 10.48550/arxiv.2506.05007

A team of engineers, AI specialists and chip design researchers at the Chinese Academy of Sciences has designed, built and tested what they are describing as the first AI-based chip design system. The group has published a [paper](#) describing their system, called QiMeng, on the *arXiv* preprint server.

Over the past several decades, integrated circuit makers have developed systems for developing processor chips for computers, smartphones and other [electronic devices](#). Such systems tend to be made up of large teams of highly skilled people who can take design ideas (such as faster computing or running AI apps) and turn them into physical designs that can be fabricated in specially designed factories. The process is notoriously slow and expensive.

More recently, computer and device makers have been looking for ways to speed up the process and to allow for more flexibility—some may want a chip that can do just one thing, for example, but do it really well. In this new study, the team in China has applied AI to the problem.

The work involved using an LLM to take user requests regarding [performance standards](#) and turn them into architectural plans for a [processor chip](#) that would meet the specifications and also create the software that runs on it.

The initiative was reportedly begun as scientists in China faced pressure to ramp up their chip-making abilities as Western countries have become less willing to share technological gains with them.

The new system has three interconnected parts. The first is a domain-specific chip model. The second is the design agent responsible for most of the work in building a design. The third part is an assemblage of design applications that are made available to the design agent. As with other AI-based learning applications, the system learns what a processor looks like and how it works by being exposed to a [large database](#) of existing technology.

The system has proved itself capable of creating effective designs—it came up with chips similar in capability to Intel's 486 and Arm's Cortex A53. Both chips are notably quite dated and lag far behind what is in use today. Still, the researchers are optimistic that QiMeng will

evolve as it learns and will eventually be capable of designing chips on a par with those now being developed in the West at far less expense and over a much shorter timeline.

More information: Rui Zhang et al, QiMeng: Fully Automated Hardware and Software Design for Processor Chip, *arXiv* (2025). DOI: [10.48550/arxiv.2506.05007](https://doi.org/10.48550/arxiv.2506.05007)

QiMeng Project Website: qimeng-ict.github.io/

Journal information: [arXiv](#)

© 2025 Science X Network

JUNE 10, 2025

Apple plays it safe on AI despite Wall Street pressure

by Glenn CHAPMAN with Alex PIGMAN in Washington edited by [Andrew Zinin](#)

Apple CEO Tim Cook speaks during the Apple Worldwide Developers Conference. Apple on Monday remained on its cautious path to embracing generative AI even as rivals race ahead with the technology and Wall Street expresses doubts over its strategy.

The pressure was on Apple not to disappoint at its annual Worldwide Developers Conference (WWDC) a year after the iPhone juggernaut made a promise it failed to keep—to improve its Siri voice assistant with generative AI.

The annual WWDC is addressed to developers who build apps and tools to run on the company's products.

Despite last year's disappointment, Apple insisted on Monday it was still very much in the AI race, announcing incremental updates to its Apple Intelligence software, including the ability for app makers to directly access a device's AI capabilities.

This would allow users to engage with apps using generative AI while offline, letting them interact ChatGPT-style with a hiking app, for example, while in remote areas without a connection.

Apple CEO Tim Cook briefly mentioned that Siri's AI makeover was still under development and "needed more time to meet our high quality bar," which includes Apple's standards on privacy and data security.

"We are making progress, and we look forward to getting these features into customers' hands," he added.

For Gadjo Sevilla, senior analyst for Emarketer, "the delays to Apple's in-house AI efforts will continue to draw scrutiny."

"Especially since rivals like Google and Samsung are moving ahead by introducing new on-device AI capabilities, or partnering with AI startups like Perplexity (in Samsung's case) to provide users with AI features," he added.

The biggest announcement at the event was the renaming of Apple's operating systems so that releases better match their release year.

The next operating system will be iOS 26 and will be available across all of Apple's devices—including the Mac, Watch and Vision Pro headset—in the fall, in time for the likely release of the next iPhone 17.

Today, Apple's operating systems have vastly different nomenclatures across devices, including the current iOS 18 for iPhone or macOS 15 for Mac computers.

Apple also announced that the new operating system will be the first major iOS redesign since 2013, calling the new look "Liquid Glass."

Wall Street divided

The relationship between Apple and app-making developers has been strained in recent years, with developers chafing at the iPhone maker's high fees for getting access to the App Store.

A marathon lawsuit by Fortnite maker Epic Games ended with Apple being ordered to allow outside payment systems to be used in the US App Store.

Adding to doubts about Apple's direction is the fact that the legendary designer behind the iPhone, Jony Ive, has joined with ChatGPT maker OpenAI to create a potential rival device for engaging with AI.

Apple also has to deal with tariffs imposed by US President Donald Trump in his trade war with China, a key market for sales growth and the place where most iPhones are manufactured.

Trump has also threatened to hit Apple with tariffs if iPhone production wasn't moved to the US, a change which analysts say would be impossible given the costs and capabilities required.

Wall Street analysts remain divided on Apple's prospects, with the stock down about 17% since the start of the year, wiping over \$600 billion from its [market value](#) and far outshone by its Big Tech rivals.

While some analysts remain optimistic about Apple's long-term AI monetization potential, others worry the company's cautious approach may prove costly in the longer term.

WWDC "was void of any major Apple Intelligence progress as Cupertino is playing it safe and close to the vest after the missteps last year," said Dan Ives of Wedbush Securities.

"We have a high level of confidence Apple can get this right, but they have a tight window to figure this out," he added.

© 2025 AFP

JUNE 11, 2025

The AI hype is just like the blockchain frenzy—here's what happens when the hype dies

by Gediminas Lipnickas, [The Conversation](#) edited by [Lisa Lock](#), reviewed by [Andrew Zinin](#)

In recent years, artificial intelligence (AI) has taken center stage across various industries. From AI-generated art to chatbots in customer service, every sector is seemingly poised for disruption.

It's not just in your news feed every day—[venture capital](#) is pouring in, while CEOs are eager to declare their companies "[AI-first](#)." But for those who remember the lofty promises of other technologies that have since faded from memory, there's an uncanny sense of déjà vu.

In 2017, it was [blockchain](#) that promised to transform every industry. Companies added "[blockchain](#)" to their name and watched [stock prices](#) skyrocket, regardless of whether the technology was actually used, or how.

Now, a similar trend is emerging with AI. What's unfolding is not just a wave of innovation, but a textbook example of a tech hype cycle. We've been here many times before.

Understanding the hype cycle

The [tech hype cycle](#), first defined by the research firm Gartner, describes how emerging technologies rise on a wave of inflated promises and expectations, crash into disillusionment and, eventually, find a more realistic and useful application.

Recognizing the signs of this cycle is crucial. It helps in distinguishing between genuine technological shifts and passing fads driven by speculative investment and good marketing.

It can also mean the difference between making a good business decision and a very costly mistake. Meta, for example, invested more than [US\\$40 billion into the metaverse](#) idea while seemingly chasing their own manufactured tech hype, only to [abandon it later](#).

When buzz outpaces reality

In 2017, blockchain was everyone's focus. Presented as a revolutionary technology, blockchain offered a decentralized way to record and verify transactions, unlike traditional systems that rely on central authorities or databases.

US soft drinks company Long Island Iced Tea Corporation became Long Blockchain Corporation and saw its stock rise [400% overnight](#), despite having no blockchain product. Kodak launched a vague cryptocurrency called KodakCoin, sending its [stock price soaring](#).

These developments were less about innovation and more about speculation, chasing short-term gains driven by hype. Most blockchain projects never delivered real value. Companies rushed in, driven by fear of missing out and the promise of technological transformation.

But the tech wasn't ready, and the solutions it supposedly offered were often [misaligned with real industry problems](#). Companies tried everything, from [tracking pet food ingredients](#) on blockchain, to launching [loyalty programs with crypto tokens](#), often without clear benefits or better alternatives.

In the end, about [90% of enterprise blockchain solutions failed](#) by mid-2019.

The generative AI déjà vu

Fast-forward to 2023, and the same pattern started playing out with AI. Digital media company BuzzFeed saw its stock jump more than 100% after [announcing it would use AI to generate quizzes and content](#). Financial services company Klarna [replaced 700 workers with an AI chatbot](#), claiming it could handle millions of customer queries.

The results were mostly negative. Klarna soon saw a decline in customer satisfaction and had to [walk back its strategy](#), rehiring humans for customer support this year. BuzzFeed's AI content push failed to save its struggling business, and its news division later [shut down](#). Tech media company CNET published AI-generated articles [riddled with errors](#), damaging its credibility.

These are not isolated incidents. They're signals that AI, like blockchain, was being over hyped.

Why do companies chase tech hype?

There are three main forces at play: inflated expectations, short-term view and flawed implementation. Tech companies, under pressure from investors and media narratives, [overpromise what AI can do](#).

Leaders pitch vague and utopian concepts of "transformation" without the infrastructure or planning to back them up. And many rush to implement, riding the hype wave.

They are often hindered by a short-term view of what alignment with the new tech hype can do for their company, ignoring the potential downsides. They roll out untested systems, underestimate complexity or even the necessity, and hope that novelty alone will drive the return on investment.

The result is often disappointment—not because the technology lacks potential, but because it's applied too broadly, too soon, and with too little planning and oversight.

Where to from here?

Like blockchain, AI is a legitimate technological innovation with real, transformative potential.

Often, these technologies simply need time to find the right application. While the initial blockchain hype has faded, the technology has found a practical niche in areas like ["asset tokenization"](#) within financial markets. This allows assets like [real estate](#) or company shares to be represented by digital tokens on the blockchain, enabling easier, faster and cheaper trading.

The same pattern can be expected with generative AI. The current AI hype cycle appears to be [tapering off](#), and the consequences of rushed or poorly thought-out implementations will likely become more visible in the coming years.

However, this decline in hype doesn't signal the end of generative AI's relevance. Rather, it marks the beginning of a more grounded phase where the technology can find the most suitable applications.

One of the clearest takeaways so far is that AI should be used to enhance human productivity, not replace it. From people [pushing back against the use of AI to replace them](#), to [AI making frequent and costly mistakes](#), human oversight paired with AI-enhanced productivity is increasingly seen as the [most likely path forward](#).

Recognizing the patterns of tech hype is essential for making smarter decisions. Instead of rushing to adopt every new innovation based on inflated promises, a measured, problem-driven approach leads to more meaningful outcomes.

Long-term success comes from thoughtful experimentation, implementation, and clear purpose, not from chasing trends or short-term gains. Hype should never dictate strategy; real value lies in solving real problems.

JUNE 11, 2025

AI innovation helps shipping ports see what's coming—literally

by Patty Wellborn, [University of British Columbia](#) Edited by [Gaby Clark](#), reviewed by [Andrew Zinin](#)

Credit: Unsplash/CC0 Public Domain

A UBC Okanagan research team has developed an innovative artificial intelligence system that can accurately predict where ships are heading and arriving, potentially helping Canadian ports better prepare for incoming vessels and respond more efficiently to global supply chain disruptions.

Dr. Zheng Lui, a Professor with UBCO's School of Engineering, and doctoral student Chengkai Zhang have created TrajReducer, a framework that increases prediction accuracy and [computational efficiency](#) by analyzing ship trajectories through advanced spatial clustering and cross-dimensional metadata ranking.

The research, published recently in [Ocean Engineering](#), addresses a critical need in maritime logistics where accurate predictions of a vessel's [arrival time](#) are essential for improving [port](#) operations, says Dr. Lui.

"Maritime shipping represents more than 80% of global trade and the ability to accurately predict where cargo vessels are headed and when they will arrive, has never been more important," he says.

"Recent years have shown us how quickly global supply chains can be disrupted—whether by pandemic-related delays, geopolitical tensions or incidents like a [shipping container](#) getting stuck in the Suez Canal for several days. Canadian ports need tools to help them adapt quickly and efficiently."

Traditional methods for predicting where ships are going have been slow and often inaccurate, he says, with about 30% of data omitting the ship's estimated time of departure and arrival.

TrajReducer solves these problems by studying patterns from several thousand ship voyages and using details about each vessel, such as its type, size, speed and direction.

"It is similar to having a smart GPS that not only knows where you've been, but can also predict where you're going based on your driving patterns and what kind of car you drive," explains Dr. Lui. "Instead of checking every single ship journey in our database, TrajReducer quickly identifies the most similar past voyages, including vessel size and weather conditions, to make its prediction."

The system works by looking at a ship's current path and comparing it to carefully chosen similar voyages. This provides a calculation of where a ship is headed with high accuracy, even early in its journey.

The technology has particular relevance for Canada's major ports, including Vancouver, Prince Rupert, Montreal and Halifax which serve as critical gateways for North American trade, says Dr. Lui.

"Canadian ports handle hundreds of millions of tonnes of cargo annually, and even small improvements in operational efficiency can translate to significant economic benefits," he adds. "When a port knows days in advance when a large container ship is arriving, staff can optimize berth assignments, arrange the right equipment and coordinate with rail and trucking networks."

All of these steps can speed up delivery of the products in the shipping containers.

Compared to existing prediction models, TrajReducer's efficiency and reliability offer significant advancements in maritime operations, as it allows for the accurate prediction of vessel destinations without extensive computational requirements.

"The beauty of this system is that it becomes more accurate as it processes more data," adds Zhang. "As global shipping patterns evolve—whether due to new trade agreements, infrastructure changes or climate considerations—TrajReducer adapts and improves its predictions."

This research represents a significant advancement in maritime analytics, with potential applications extending beyond port operations to include maritime safety, environmental monitoring and supply chain optimization.

"This is about more than efficiency—it's about resilience," says Zhang. "When Canadian ports can anticipate and prepare for the arrival of ships accurately, they're better positioned to maintain the flow of goods that Canadians depend on, even when [global trade](#) faces unexpected challenges."

More information: Chengkai Zhang et al, TrajReducer: a cross-dimension indexer-based reducer for vessel destination prediction, *Ocean Engineering* (2025). DOI: [10.1016/j.oceaneng.2025.121404](https://doi.org/10.1016/j.oceaneng.2025.121404)
Provided by [University of British Columbia](#)

JUNE 11, 2025

Nvidia marks Paris tech fair with Europe AI push

by Mona GUICHARD edited by [Andrew Zinin](#)

Nvidia CEO Jensen Huang headlined the first day of the Vivatech tech trade fair and announced a major expansion of AI computing capacity in Europe. Drawing high-powered tech CEOs and a presidential visit, the Vivatech trade fair opened in Paris on Wednesday with a bang as Nvidia boss Jensen Huang announced a major push into Europe.

"In just two years we will increase the amount of AI computing capacity in Europe by a factor of 10," Huang told a packed hall in a southern Paris convention center, striding around the stage wearing his trademark leather jacket.

He also announced a multi-billion-dollar partnership with French AI champion Mistral AI.

President Emmanuel Macron hailed the Nvidia-Mistral tie-up as a "historic" opportunity for France and Europe, urging other local firms to climb aboard.

He had arrived late on Wednesday afternoon for a tour of the show and meetings with European startups about technological sovereignty, a subject dear to his heart.

"We want AI... that's secure, sustainable, humanist," Macron said.

People from around the globe thronged the halls of Vivatech, crammed with stands in blaring colors showing off the latest innovations from startups, tech giants and more traditional firms and patrolled here and there by gesticulating robots.

Around 14,000 startups and more than 3,000 investors were expected in Paris, while organizers forecast total visitor numbers to at least equal last year's 165,000 people.

Nvidia headlining

Nvidia's Huang took top billing with an opening presentation of almost two hours that drew bouts of rapturous applause from attendees.

The US firm's tie-up with Mistral will see the companies build a cloud computing platform powered by 18,000 of Nvidia's "Blackwell" high-end chips worth billions.

Speaking in a panel discussion with Huang and Macron, Mistral chief Arthur Mensch said the offering would be "completely independent" in a nod to the president's sovereignty drive.

"You're no longer relying for your AI workload on certain of the US providers," he promised the audience.

Macron dubbed the Mistral-Nvidia collaboration a "game-changer, because it will increase our sovereignty and it will allow us to do much more" with AI.

Europe "has put its ability to produce things in danger" and "become more and more dependent on the rest of the world," he warned.

French President Emmanuel Macron gave the Cosmo Connected helmet a try during his visit to the Vivatech technology trade fair.

Aside from Mistral, Nvidia will also intensify work with existing partners like Germany's Siemens and France's Schneider Electric, Huang said.

And it will help build multiple data centers in seven European countries.

Europe is well behind competitors like the United States and China in building up the computing power needed to power generative artificial intelligence.

The continent hosts "less than five percent of global computing power, whereas we consume 20%," Macron's office said in a press briefing ahead of the leader's visit to Vivatech.

Trade war

Nvidia has seen export restrictions slapped on its top-performing chips by Washington, with American politicians leery of ceding their country's lead in generative AI.

Remaining high-tech controls on China are at issue in high-stakes trade talks with Beijing.

Huang has warned that the US' superpower rival is nevertheless making swift strides to catch up.

There was little sign of impact from export restrictions on Nvidia's chip sales in its May earnings release.

But the company has warned the braking effect may be larger in the current quarter.

US politics also preoccupies many European tech leaders and policymakers.

Concerns range from Trump's mercurial tariff policy to the continent's ability to stand on its own without US tech giants—and the massive gap in funding for AI development between the two sides of the Atlantic.

"Sovereignty, which wasn't as important in the conversation just a year or two years ago, has become an absolutely strategic priority," Vivatech managing director Francois Bitouzet told AFP.

Macron's hammering on tech sovereignty followed on from his hyping of French and European openness to AI at a Paris global summit in February.

Macron, Mensch and Huang were set to dine together behind closed doors at the president's Elysee Palace residence on Wednesday evening.

© 2025 AFP

JUNE 12, 2025

Training robots without robots: Smart glasses capture first-person task demos

by [Ingrid Fadelli](#), Phys.org edited by [Lisa Lock](#), reviewed by [Robert Egan](#)

Human demonstrations are done with only black ovens (top). The policy transfers zero-shot to the robot with the same oven (middle) and also generalizes to a new oven instance (bottom). The points are color-coded to represent the correspondence. Credit: Liu et al. Over the past few decades, robots have gradually started making their way into various real-world settings, including some malls, airports and hospitals, as well as a few offices and households.

For robots to be deployed on a larger scale, serving as reliable everyday assistants, they should be able to complete a wide range of common manual tasks and chores, such as cleaning, washing the dishes, cooking and doing the laundry.

Training machine learning algorithms that allow robots to successfully complete these tasks can be challenging, as it often requires extensive annotated data and/or demonstration videos showing humans the tasks. Devising more effective methods to collect data to train robotics algorithms could thus be highly advantageous, as it could help to further broaden the capabilities of robots.

Researchers at New York University and UC Berkeley recently introduced EgoZero, a new system to collect ego-centric demonstrations of humans completing specific manual tasks. This system, introduced in a [paper](#) posted to the *arXiv* preprint server, relies on the use of Project Aria glasses, the [smart glasses](#) for augmented reality (AR) developed by Meta.

"We believe that general-purpose robotics is bottlenecked by a lack of internet-scale data, and that the best way to address this problem would be to collect and learn from first-person human data," Lerrel Pinto, senior author of the paper, told Tech Xplore.

"The primary objectives of this project were to develop a way to collect accurate action-labeled data for robot training, optimize for the ergonomics of the data collection wearables needed, and transfer human behaviors into robot policies with zero robot data."

EgoZero, the new system developed by Pinto and his colleagues, relies on Project Aria smart glasses to easily collect video demonstrations of humans completing tasks while performing robot-executable actions, captured from the point of view of the person wearing the glasses.

These demonstrations can in turn be used to train robotics algorithms on new manipulation policies, which could in turn allow robots to successfully complete various manual tasks.

"Unlike prior works that require multiple calibrated cameras, wrist wearables, or motion capture gloves, EgoZero is unique in that it is able to extract these 3D representations with only smart glasses (Project Aria smart glasses)," explained Ademi Adeniji, student and co-lead author of the paper.

"As a result, robots can learn a new task from as little as 20 minutes of human demonstrations, with no teleoperation."

Architecture diagram. EgoZero trains policies in a unified state-action space defined as egocentric 3D points. Unlike previous methods, EgoZero localizes object points via triangulation over the camera trajectory, and computes action points via Aria MPS hand pose and a hand estimation model. These points supervise a closed-loop Transformer policy, which is rolled out on unprojected points from an iPhone during inference. Credit: Liu et al.

To evaluate their proposed system, the researchers used it to collect video demonstrations of simple actions that are commonly completed in a household environment (e.g., opening an oven door) and then used these demonstrations to train a machine learning algorithm.

The machine learning algorithm was then deployed on Franka Panda, a [robotic arm](#) with a gripper attached at its end. Notably, they found that the robotic arm successfully completed most of the tasks they tested it on, even if the [algorithm](#) planning its movements underwent minimal training.

"EgoZero's biggest contribution is that it can transfer human behaviors into robot policies with zero robot data, with just a pair of smart glasses," said Pinto.

"It extends past work (Point Policy) by showing that 3D representations enable efficient robot learning from humans, but completely in-the-wild. We hope this serves as a foundation for future exploration of representations and algorithms to enable human-to-robot learning at scale."

The code for the data collection system introduced by Pinto and his colleagues was [published on GitHub](#) and can be easily accessed by other research teams.

In the future, it could be used to rapidly collect datasets to train robotics algorithms, which could contribute to the further development of robots, ultimately facilitating their deployment in a greater number of households and offices worldwide.

"We now hope to explore the tradeoffs between 2D and 3D representations at a larger scale," added Vincent Liu, student and co-lead author of the paper.

"EgoZero and past work (Point Policy, P3PO) have only explored single-task 3D policies, so it would be interesting to extend this framework of learning from 3D points in the form of a fine-tuned LLM/VLM, similar to how modern VLA models are trained."

*Written for you by our author [Ingrid Fadelli](#), edited by [Lisa Lock](#), and fact-checked and reviewed by [Robert Egan](#)—this article is the result of careful human work. We rely on readers like you to keep independent science journalism alive. If this reporting matters to you, please consider a [donation](#) (especially monthly). You'll get an **ad-free** account as a thank-you.*

More information: Vincent Liu et al, EgoZero: Robot Learning from Smart Glasses, *arXiv* (2025). DOI: [10.48550/arxiv.2505.20290](https://doi.org/10.48550/arxiv.2505.20290)
Journal information: [arXiv](#)

© 2025 Science X Network

JUNE 12, 2025

Innovative detection method makes AI smarter by cleaning up bad data before it learns

by Gisele Galoustian, [Florida Atlantic University](#)

edited by [Lisa Lock](#), reviewed by [Andrew Zinin](#)

In the world of machine learning and artificial intelligence, clean data is everything. Even a small number of mislabeled examples known as label noise can derail the performance of a model, especially those like support vector machines (SVMs) that rely on a few key data points to make decisions.

SVMs are a widely used type of [machine learning](#) algorithm, applied in everything from image and speech recognition to medical diagnostics and text classification. These models operate by finding a boundary that best separates different categories of data. They rely on a small but crucial subset of the training data, known as support vectors, to determine this boundary. If these few examples are incorrectly labeled, the resulting decision boundaries can be flawed, leading to [poor performance](#) on real-world data.

Now, a team of researchers from the Center for Connected Autonomy and Artificial Intelligence (CA-AI) within the College of Engineering and Computer Science at Florida Atlantic University and collaborators have developed an innovative method to automatically detect and remove faulty labels before a model is ever trained—making AI smarter, faster and more reliable.

Before the AI even starts learning, the researchers clean the data using a math technique that looks for odd or unusual examples that don't quite fit. These "outliers" are removed or flagged, making sure the AI gets high-quality information right from the start. The paper is [published](#) in *IEEE Transactions on Neural Networks and Learning Systems*.

"SVMs are among the most powerful and widely used classifiers in machine learning, with applications ranging from cancer detection to spam filtering," said Dimitris Pados, Ph.D., Schmidt Eminent Scholar Professor of Engineering and Computer Science in the FAU Department of Electrical Engineering and Computer Science, director of CA-AI and an FAU Sensing Institute (I-SENSE) faculty fellow.

"What makes them especially effective—but also uniquely vulnerable—is that they rely on just a small number of key data points, called support vectors, to draw the line between different classes. If even one of those points is mislabeled—for example, if a malignant tumor is incorrectly marked as benign—it can distort the model's entire understanding of the problem.

The consequences of that could be serious, whether it's a missed cancer diagnosis or a security system that fails to flag a threat. Our work is about protecting models—any machine learning and AI model including SVMs—from these hidden dangers by identifying and removing those mislabeled cases before they can do harm."

The data-driven method that "cleans" the training dataset uses a mathematical approach called L1-norm principal component analysis. Unlike conventional methods, which often require manual parameter tuning or assumptions about the type of noise present, this technique identifies and removes suspicious data points within each class purely based on how well they fit with the rest of the group.

"Data points that appear to deviate significantly from the rest—often due to label errors—are flagged and removed," said Pados. "Unlike many existing techniques, this process requires no manual tuning or user intervention and can be applied to any AI model, making it both scalable and practical."

The process is robust, efficient and entirely touch-free—even handling the notoriously tricky task of rank selection (which determines how many dimensions to keep during analysis) without user input.

Researchers extensively tested their technique on real and synthetic datasets with various levels of label contamination. Across the board, it produced consistent and notable improvements in classification accuracy, demonstrating its potential as a standard pre-processing step in the development of high-performance machine learning systems.

"What makes our approach particularly compelling is its flexibility," said Pados. "It can be used as a plug-and-play preprocessing step for any AI system, regardless of the task or dataset. And it's not just theoretical—extensive testing on both noisy and clean datasets, including well-known benchmarks like the Wisconsin Breast Cancer dataset, showed consistent improvements in classification accuracy.

"Even in cases where the original training data appeared flawless, our new method still enhanced performance, suggesting that subtle, hidden label noise may be more common than previously thought."

Looking ahead, the research opens the door to even broader applications. The team is interested in exploring how this mathematical framework might be extended to tackle deeper issues in data science such as reducing data bias and improving the completeness of datasets.

"As machine learning becomes deeply integrated into high-stakes domains like health care, finance and the justice system, the integrity of the data driving these models has never been more important," said Stella Batalama, Ph.D., dean of the FAU College of Engineering and Computer Science.

"We're asking algorithms to make decisions that impact real lives—diagnosing diseases, evaluating loan applications, even informing legal judgments. If the training data is flawed, the consequences can be devastating. That's why innovations like this are so critical.

"By improving data quality at the source—before the model is even trained—we're not just making AI more accurate; we're making it more responsible. This work represents a meaningful step toward building AI systems we can trust to perform fairly, reliably and ethically in the real world."

More information: Shruti Shukla et al, Training Dataset Curation by L 1-Norm Principal-Component Analysis for Support Vector Machines, *IEEE Transactions on Neural Networks and Learning Systems* (2025). DOI: [10.1109/TNNLS.2025.3568694](https://doi.org/10.1109/TNNLS.2025.3568694)
Provided by [Florida Atlantic University](#)